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

Weimin Qiu Jieke Wang Meng Tang
University of California Merced
{wqiu5, jwang450, mtang4}@ucmerced.edu

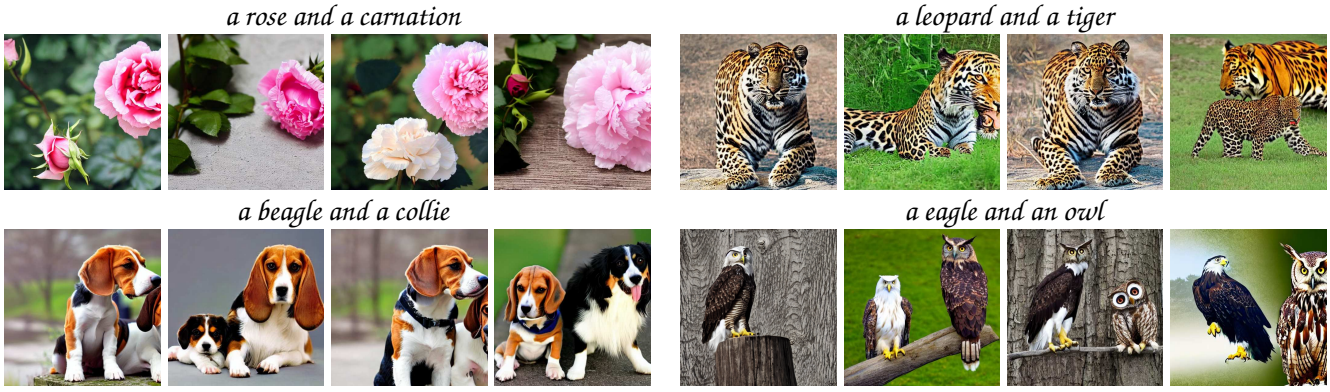


Figure 1. Our Self-Cross guidance addresses subject mixing in particular for similar subjects. Our training-free method can boost the performance of any Unet-based or transformer-based diffusion models such as Stable Diffusion 1, 2, and 3 (shown in Fig. 5 and 6.)

Abstract

Diffusion models achieved unprecedented fidelity and diversity for synthesizing image, video, 3D assets, etc. However, subject mixing is an unresolved issue for diffusion-based image synthesis, particularly for synthesizing multiple similar-looking subjects. We propose Self-Cross Diffusion Guidance to penalize the overlap between cross-attention maps and the aggregated self-attention map. Compared to previous methods based on self-attention or cross-attention alone, our guidance is more effective in eliminating subject mixing. What's more, our guidance addresses subject mixing for all relevant patches beyond the most discriminant one, e.g., the beak of a bird. For each subject, we aggregate self-attention maps of patches with higher cross-attention values. Thus, the aggregated self-attention map forms a region that the whole subject attends to. Our training-free method boosts the performance of both Unet-based and Transformer-based diffusion models such as the Stable Diffusion series. We also release a similar subjects dataset (SSD), a challenging benchmark, and utilize GPT-4o for automatic and reliable evaluation. Extensive qualitative and quantitative results demonstrate the effectiveness of our self-cross diffusion guidance.

1. Introduction

Diffusion-based generative models have made significant progress in recent years in synthesizing high-quality images [41], videos [5], 3D assets [37], etc. With a simple text prompt, diffusion-based generative models such as Stable Diffusion [41] can create highly photorealistic or artistic images of various styles and subjects. Such progress revolutionized many applications including the content creation. However, text-to-image diffusion models still have many issues such as *subject neglect*, *subject mixing*, and *attribute binding*. We are particularly interested in solving *subject mixing* in this work. Many mitigations [8, 17, 31] were proposed to enhance the faithfulness of text-to-image synthesis but the issue of *subject mixing remains*, see failure cases of previous methods [17, 31] in Fig. 1. This is more prominent when synthesizing multiple similar-looking subjects, e.g., a photo of a leopard and a tiger in Fig. 1.

Since self-attention maps reflect the similarity of patches and cross-attention maps reflect the subjects. Intuitively, patches from a subject should not show large similarity to patches from other subjects at least for some timesteps and layers to avoid copying patches from other subjects. Empirically, Fig. 2 shows a failure case for Stable Diffusion [41] with subject mixing and an obvious overlap between cross-

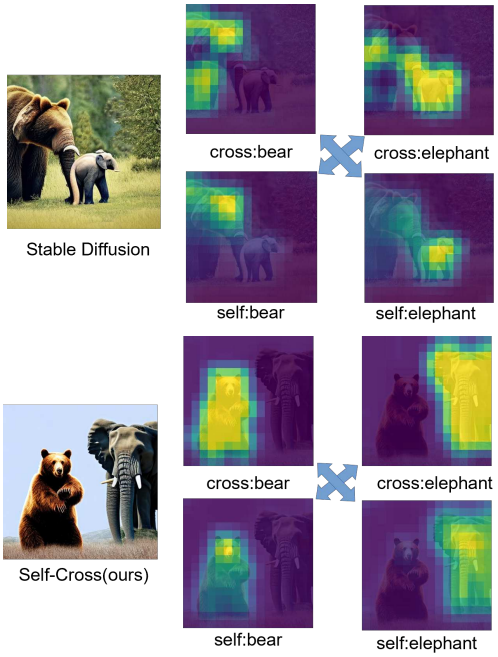


Figure 2. Results of Stable Diffusion [41] and our method with Self-Cross guidance for the same prompt "a bear and an elephant". Images are generated from the same random seed. "cross" means cross-attention map and "self" means the aggregated self-attention map. The overlap between self-attention and cross-attention leads to subject mixing, while Self-Cross guidance reduces overlapping.

attention maps w.r.t. a subject token and self-attention maps w.r.t. patches of another subject. Our insight is that *a subject should not attend to other subjects by its self-attention maps*. As shown in Fig. 2, while overlapping between self-attention and cross-attention maps results in a failure case, our Self-Cross diffusion guidance penalizes such overlaps yielding synthetic images without subject mixing.

Our method is different from previous methods based on cross-attention or self-attention maps alone [3, 17, 31]. We are the first to explore regularization between self-attention and cross-attention maps. What's more, we formulate Self-Cross diffusion guidance for self-attention maps of multiple image patches beyond the most discriminant one. It is well-known that neural networks for perceptron such as image classification tend to focus on the discriminant region of an image, e.g., the beak of a bird. However, apart from the most discriminant patch, other patches with more details are also particularly important for synthesizing multiple similar-looking subjects. We first adaptively identify all image patches corresponding to a subject and then avoid their attendance to other subjects by self-attention maps. As a training-free method, it outperforms other single-patch-based approaches [8, 17] qualitatively and quantitatively.

Previous datasets with prompts for image synthesis are not challenging in terms of subject mixing. What's worse, the commonly used CLIP score doesn't correlate well with human judgment. Therefore, to facilitate image synthesis of similar objects, we release a similar-subject dataset (SSD) consisting of text prompts with similar subjects. To enable benchmarking at scale, we leverage state-of-the-art vision language models such as GPT-4o to evaluate synthetic images of different methods by visual question answering [22].

The main contributions of this paper are as follows.

- We propose Self-Cross diffusion guidance for text-to-image synthesis of similar-looking subjects, which effectively addresses subject mixing as shown in our qualitative and quantitative results.
- Our guidance is training-free and can improve the performance of pre-trained models such as Stable Diffusion.
- We propose similar-subject dataset (SSD), a new benchmark for image synthesis of similar subjects. Our benchmark includes prompts for two or three similar subjects and an effective metric using VLM [22].
- As a side effect evidenced by improved existence and recognizability scores, our method reduced subject neglect.

2. Related Work

Text-to-Image Diffusion Models Given text prompts, text-to-image synthesis aims to generate visually coherent images. Early approaches utilized GANs [16, 25, 43, 44, 48, 54, 59] and autoregressive models [7, 10, 11, 14, 27, 38, 55, 56]. Recently, with ground breaking advancements in diffusion models [9, 20, 21, 30, 33, 46, 47], the focus of text-to-image synthesis has shifted toward diffusion models [4, 13, 36, 39, 41, 43, 58]. Although diffusion models can generate photorealistic images, ensuring faithful adherence to the provided text prompt remains a significant challenge. To tackle this issue, methods like ReCap [45], DALLE3 [4], and SD3 [13] leverage improved image-caption pairs during training or incorporate multiple language encoders to capture more expressive language representations. However, these methods require training models from scratch, which entails substantial computational costs and makes them inapplicable to popular models like Stable Diffusion [41] and Imagen [43]. Furthermore, these models only partially address issues such as subject neglect, subject missing, and attribute binding, leaving room for improvement towards prompt-faithful synthesis.

Guidance for Consistent Text-to-Image Generations Training-free inference-time optimization is an active research area to improve the consistency of the pre-trained text-to-image models. These methods typically extract internal representations of the denoising networks and correct the denoising trajectory to improve the alignment to the given prompt. One way to modulate the denoising trajectory is to replace internal features, such as attention modules in PnP-Diffusion [50], FreeControl [32], Prompt-to-

Prompt [19], DenseDiffusion [26], and MasaCtrl [6]. While effective for image editing and style transfer, these methods fail to address critical issues like subject mixing and attribute misalignment. Another line of work optimizes latents by minimizing guidance loss, in a way similar to classifier guidance [9] and classifier-gree guidance [20]. To address the problems of subject neglect, subject mixing, and attribute binding, many approaches have been proposed. Ge et al. [15], Self-Guidance [12], and DisenDiff [57] design loss functions for image editing and controllable generation. Attention Refocusing [35], BoxDiff [53], TokenCompose [52], and Ge et al. [15] use pre-defined layouts, either from external models or users, as inference-time supervision. However, these methods rely on prior knowledge, which is sometimes unreachable in real world. Therefore, other knowledge-free methods have been developed. Following Attend&Excite [8] and A-STAR [2] steps, CONFORM [31] takes contrastive loss for subject separation and attribute-binding. Linguistic Binding [40] introduces a variant of Kullback-Leibler divergence as the loss for improved consistency. INITNO [17] leverages both cross-attention maps and self-attention maps to refine the initial noise. While they use cross-attention and self-attention separately, we emphasize that the interaction between self-attention and cross-attention is key to eliminating subject mixing and improving faithfulness. Moreover, these methods typically consider the most discriminant patch, which is insufficient for removing subject mixing.

3. Preliminaries

Diffusion Model Latent diffusion model [41] operates in a latent space instead of the pixel space, which largely reduces the computational complexity of image generation. An encoder and a decoder are trained to encode images and decode lower-dimensional latents respectively. Furthermore, cross-attention between prompts and image patches allows controllable image generation with various prompts, such as layouts, semantics, and texts.

In the latent space, the forward process gradually adds Gaussian noises on the latent code z_0 over time until it completely deteriorates to Gaussian noise z_T. While in the reverse denoising process, a denoising network [42] ϵ_θ denoises the latent code z_t iteratively until time step zero z_0. The training objective is formally defined as:

$$\mathcal{L} = \mathbb{E}_{z_t, \epsilon \sim N(0,I), c(y), t} ||\epsilon - \epsilon_\theta(z_t, c(y), t)||^2 \quad (1)$$

where $c(y)$ represents the condition embedding, for example CLIP embedding for text prompt y.

In cross-attention modules, condition embeddings c are projected to query Q and values V. Accordingly, intermediate representations from UNet are projected to keys K. Therefore, the cross-attention maps can be described as:

$$A^c = \text{Softmax}(\frac{QK^T}{\sqrt{d}}) \quad (2)$$

where $\sqrt{d}$ is a scaling factor [51]. Each text token has a cross-attention map with shape $R^{h/P \times w/P}$ for patch size $P \times P$. The cross-attention maps reflect the attendance of text tokens to patches. On the other hand, the self-attention map for each patch indicates the relationship between different patches. We denote the cross attention map of a text token k as $A^c_k \in R^{h/P \times w/P}$ and the self attention map of a patch (x, y) as $A^s_{x,y} \in R^{h/P \times w/P}$.

Attention Based Guidance As discussed in Sec. 2, one methodology for consistent text-to-image generation is attenton-based guidance [3, 8, 17, 31, 35] using self attention or cross attention. Here we briefly introduce the most relevant methods to our Self-Cross diffusion guidance.

To avoid subject neglect, Attend-and-Excite [8] finds the patch with maximum cross attention for each token, and penalizes if the maximum cross attention is small. In other words, it encourages the appearance of specified subjects. Speficailly, the cross-attention response score is defined as,

$$S_{\text{cross-attn}} = \max_{k \in K} S_{\text{cross-attn},k} \quad (3)$$

where K indicates the set of subjects' tokens and

$$S_{\text{cross-attn},k} = 1 - \max(A^c_k). \quad (4)$$

Unlike Attend-and-Excite[8], our Self-Cross diffusion guidance loss is between self-attention and cross-attention. Besides, it's formulated for all relevant subject patches beyond the most discriminant one. The most similar guidance to ours is the self-attention conflict score introduced in INITNO [17]. However, INITNO [17] is still limited to the most discriminant patch, and our method outperformed INITNO [17] by a large margin shown in our experiments. While separation loss [3] penalizes overlap between cross-attention maps, it isn't a training-free method. Our proposed Self-Cross guidance is novel and complementary to existing works on attention-based guidance.

Optimization of Guidance Loss The standard method for optimizing a guidance loss is through gradient descent. To better minimize a guidance loss, Iterative Latent Refinement [8] runs multiple steps of gradient descent until the guidance loss is bellow a threshold. To find a better initial noise, Initial noise optimization [17] (INITNO) optimizes over the mean and variance of initial noise util the initial guidance loss is satisfactory. We adopted the two techniques above for our Self-Cross diffusion guidance.

4. Our Method

We describe Self-Cross diffusion guidance given a pair of similar subjects, namely "a bear and an elephant" with "bear" at index $i = 2$ and "elephant" at index at $j = 5$. Our method can be easily extended to multiple pairs.

Aggregation of Self-Attention Maps Fig. 3 describes the aggregation of self-attention maps. Given the cross-attention map of "bear", we select patches with high responses and visualize their self-attention maps. The diversity of self-attention for different patches means the self-attention map for the most discriminant patch alone can't cover all the regions the subject attends to. Thus, the key to our Self-Cross diffusion guidance is aggregating self-attention maps. If limited to the most discriminant patch (with the highest cross-attention value), other foreground patches may lead to subject mixing, as shown in Fig. 7.

Firstly, we apply simple and efficient Otsu's method [34] on the cross-attention maps, which automatically returns the threshold and masked patches with relatively higher cross-attention values. Secondly, we aggregate all the self-attention maps of masked patches to better represent the whole region that a subject attends to. For subject at i with cross-attention map A_i^c, aggregated self-attention map is:

$$A_i^s = \frac{\sum_{x_m,y_n}(A_i^c[x_m, y_n] \times A^s_{x_m,y_n})}{\sum_{x_m,y_n} A_i^c[x_m, y_n]} \tag{5}$$

Where $[x_m, y_n]$ are the coordinates of the selected patches, $A_i^c[x_m, y_n]$ is the cross-attention value for token at i of the patch at $[x_m, y_n]$, and $A_i^s[x_m, y_n]$ is the self-attention map for the patch at $[x_m, y_n]$.

In summary, we select patches corresponding to a subject and then take a weighted sum of the self-attention maps for these patches, where the weights are the patches' cross attention values. Likewise, we obtain aggregated self-attention maps A_i^s for all subjects at given indexes.

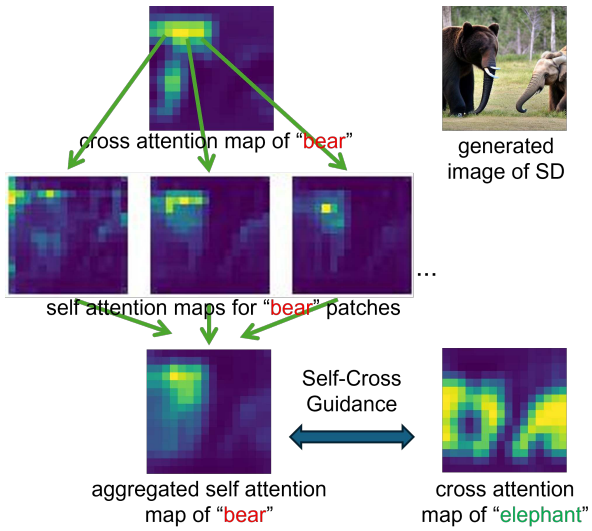


Figure 3. Self-Cross diffusion guidance between the cross-attention of "elephant" and the self-attention of "bear"

Self-Cross Diffusion Guidance The aggregated self-attention maps highlight regions that the subject attends to. Our core assumption is that a subject should not attend to other subjects in the image at some time-steps and layers through its self-attention maps. Hence, we propose to penalize the overlap between the aggregated self-attention map of one subject and cross-attention maps of other similar subjects. This has a different effect compared to a loss between cross-attention maps, as aggregated self-attention can be different from cross-attention, shown in Fig. 3.

The example prompt in Fig. 3 has two pairs of attention maps, aggregated self-attention of "bear" & cross-attention of "elephant" and aggregated self-attention of "elephant" & cross-attention of "bear". The Self-Cross diffusion guidance between the subject at i and the subject at j is defined as their overlap $g(i, j)$,

$$\begin{aligned} g(i,j) = &\sum_{x,y} \min(A_i^s[x,y], A_j^c[x,y]) \\ &+ \sum_{x,y} \min(A_i^c[x,y], A_j^s[x,y]) \end{aligned} \tag{6}$$

where A_i^s is the self-attention map of subject at i and A_j^c is the cross-attention map of subject at j.

Now, let's consider more similar subjects. If there are N similar subjects, then mathematically we would have C_N^2 pairs of subjects. In this case, our Self-Cross diffusion guidance is the average of the C_N^2 ones, as in Eq.7,

$$S_{\text{self-cross}} = \sum_{i,j\in\Omega}^{i\neq j} \frac{g(i,j)}{C_N^2}. \tag{7}$$

where Ω is the set containing all subject indexes. Similar to INITNO [17], we also include the cross-attention response score in our total loss.

$$\mathcal{L}_{total} = S_{\text{self-cross}} + \lambda \cdot S_{\text{cross-attn}}. \tag{8}$$

Spatial Relationship Among Subjects We define Our novel self-cross diffusion guidance with much thought to penalize appearance overlap while not hindering generation capabilities, as shown quantitatively in Tab. 5, Tab. 2 and qualitatively in Fig. 6. To achieve this,

(a) We define the guidance only for early timestamps in the reverse process. Because attention maps from early timestamps are known to be semantically meaningful.

(b) It is known that attention maps of intermediate layers in diffusion models are highly correlated to semantics. So we enforce our guidance for these selected layers.

(c) Other tokens such as verbs and adjectives in a prompt can render subject relationships for image synthesis. These tokens are not involved in our guidance.

See more implementation details in the appendix A.

Overall Pipeline Alg. 1 shows the overall pipeline, which involves initial noise optimization for our total loss 8. We apply Self-Cross guidance to the first half of the reverse process, which is more relevant for the semantic structure of synthesized images. Similar to previous work [8, 17], iterative refinement is conducted to ensure losses are below specified thresholds.

Algorithm 1: T2I with Self-Cross Guidance

Input: A text prompt P and indices Ω of subjects
A pre-trained T2I diffusion model such as $SD(\cdot)$
Max iterations: $\tau_{MaxAlterStep}$, $\tau_{MaxIter}$
Thresholds: $\tau_{\text{cross-attn}}$, $\tau_{\text{self-cross}}$
A set of iterations for refinement $\{t_1, t_2, ..., t_k\}$.
Output: Generated image z^0.

Noise Initialization Step:
Noise pool $\mathcal{P} \leftarrow \{\}$, t $\leftarrow T$;
do INITNO (loss $\mathcal{L}_{total} = S_{\text{self-cross}} + \lambda S_{\text{cross-attn}}$)
return noise pool $\mathcal{P}$
$z^t \leftarrow \arg\min_{z^t \in \mathcal{P}} \mathcal{L}_{total}$

Reverse Process:
while t $\geq \tau_{MaxAlterStep}$ **do**
  ▷ Compute attention and losses
  $S_{\text{cross-attn}}, S_{\text{self-cross}}, \_ \leftarrow SD(z^t, P, t)$
  if t $\in \{t_1, t_2, ..., t_k\}$ *and*
  ($S_{\text{cross-attn}} > \tau_{\text{cross-attn}}$ *or*
  $S_{\text{self-cross}} > \tau_{\text{self-cross}}$) **then**
    ▷ Iterative Refinement Steps
    $i = 0$
    while ($S_{\text{cross-attn}} > \tau_{\text{cross-attn}}$ *or*
    $S_{\text{self-cross}} > \tau_{\text{self-cross}}$) *and* $i < \tau_{MaxIter}$
    do
      $z^t \leftarrow SGD(\mathcal{L}_{total})$
      $i = i + 1$
    end
  else
    $z^t \leftarrow SGD(\mathcal{L}_{total})$
  end
  $z^t \leftarrow SD(z^t, P, t)$
  t = t −1
end
▷ Continue the remaining steps without guidance
while t > 0 **do**
  $\_, \_, z^t \leftarrow SD(z^t, P, t)$
  t = t −1
end
return z^0

5. Experiments

Datasets and Baselines Similar to previous work [8, 17, 31] on consistent text-to-image generation, we first report results on three datasets including animal-animal, animal-object, and object-object prompts based on Stable Diffusion models.[1] After visualization of experimental results, We found some prompts are visually distinct and not challenging enough. Hence, we introduce our new dataset **Similar Subjects Dataset (SSD)** containing 31 prompts with two subjects(SSD-2) and 21 prompts with three subjects(SSD-3). The former ones are designed for primitive baselines, e.g., Stable Diffusion 1, and the latter ones are intended for stronger baselines, e.g., Stable Diffusion 3. In the new dataset, subjects usually share similar structures with distinguishable details, e.g., different textures for leopards and tigers shown in Fig. 1. Qualitative and quantitative results show more subject mixing for each method with the new dataset than the original ones. Additionally, We also verify the spatial reasoning capacity of our method on 2D-spatial and 3D-spatial from T2ICompBench [23] [24]. These subsets emphasize spatial relationships between two subjects.

We compare our method to the original Stable Diffusion [41] as well as other training-free methods including Initial Noise Optimization (INITNO) [17] and CONFORM [31]. Separate-and-Enhance [3] is relevant to our work regarding subject mixing, but it necessitates fine-tuning, making it incomparable to training-free approaches.

5.1. Qualitative results

Self-Cross *v.s.* baselines We provide the qualitative comparisons in Figure 4, Figure 5, and Figure 6 using SD1.4, SD2.1, and SD3-medium respectively. For each prompt, we used the same list of random seeds for all methods. Self-Cross diffusion guidance successfully addressed the issue of subject mixing in most cases. For example, in Figure 4, given the prompt "a bird and a rabbit", Stable diffusion [41], INITNO [17], and CONFORM [31] generated birds with rabbits'ears, while our method generated faithful images without subject mixing. Our method synthesized better images for extremely similar subjects too, e.g., "a hummingbird and a kingfisher". Hummingbirds are recognized for their iridescent plumage and long, slender beaks, whereas kingfishers typically feature bold, vibrant colors like blue and have shorter, stout beaks. Appendices E and G show more qualitative results, failure cases, and discussion.

Self-attention and Cross-attention maps Fig. 3 presents a cross-attention map and multiple self-attention maps for top-responsive image patches, along with an aggregated

[1] We used the original dataset except for replacing the word "mouse" with "rat" for any Animal-Animal prompt including "mouse". This avoids ambiguity, as "mouse" can refer to a rodent or a computer mouse.

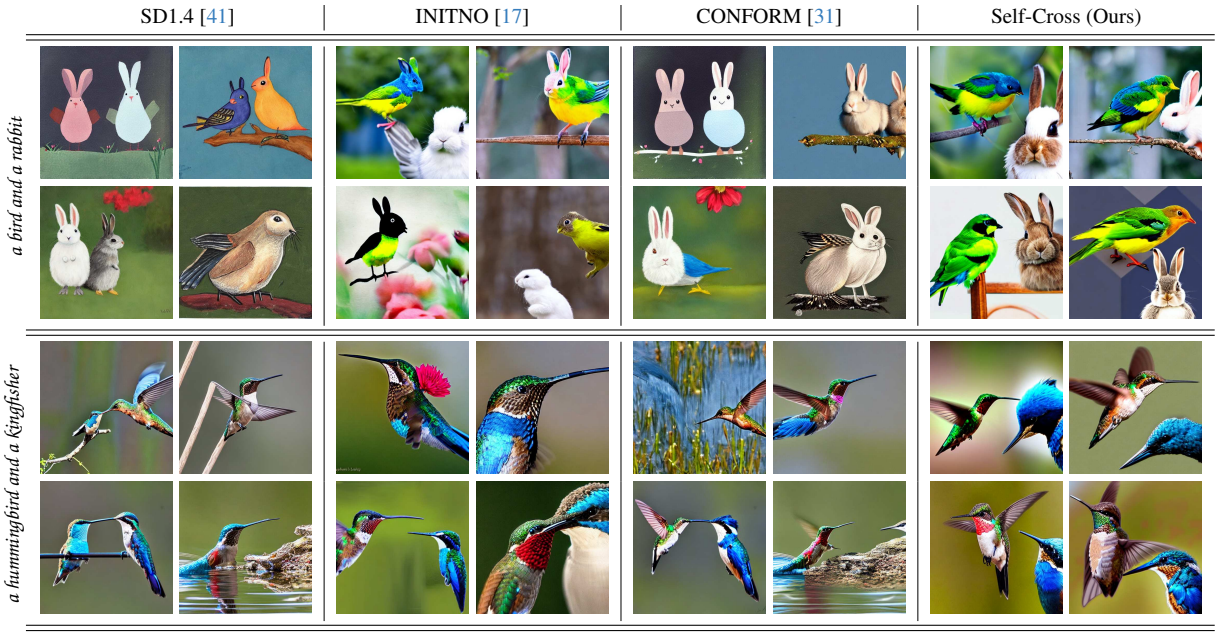


Figure 4. Qualitative comparisons of Self-Cross (ours) to SD1.4 [41], INITNO [17], CONFORM [31]. For each prompt in the left column, we sample four seeds and show the results of different methods.

self-attention map highlighting the region the entire subject("bear") attend to. Note that the aggregated self-attention map is different from the original cross-attention map, which motivated Self-Cross diffusion guidance rather than using cross-attention maps alone. The next section explores different self-attention aggregation schemes and their impact on image synthesis.

Self-Cross guidance with the increasing number of patches To see the effect of self-attention aggregation, we compare our Self-Cross diffusion guidance using different patches. Specifically, given a cross-attention map, we select

(a) the patch with the maximum cross-attention value.
(b) the top 16 patches w.r.t. cross-attention value.
(c) masked patches which typically have more than 16 patches. The details of masking are in Sec. 4.

Fig. 7 shows results with three example prompts, which clearly demonstrate less subject missing with more patches. If only the patch with the maximum cross-attention is used in Self-Cross guidance, the subjects are still mixing. The mixing is less severe but not eliminated with 16 patches, for example mixing in terms of hair, teeth, and feet. The best is when we use all masked patches in the otsu mask and define an aggregated self-attention map for our guidance, in which case the sutle mixing is eliminated.

5.2. Quantitative results

Quantitative results include TIFA-GPT4o scores and text-text similarities. Our TIFA-GPT4o is a variant of the recently proposed TIFA metric [22], which is much more correlated with human judgment compared to the widely used CLIP score. Our main quantitative results are TIFA-GPT4o scores, while CLIP scores should be interpreted with caution (see appendix D for details).

Methods	SSD-3			
Question Types	Ext	Rec	w/o M	t-t sim
SD3 Medium	33.54	30.31	70.08	73.82
Self-Cross(Ours)	**57.92**	**53.08**	**77.15**	**74.96**

Table 1. Quantitative results using SD3 for prompts with three subjects.

%	Animal-Animal	Animal-Obj	Obj-Obj	SSD-2
SD1.4	76.5	79.2	76.4	70.0
INITNO	82.2	84.0	82.3	72.0
CONFORM	81.9	84.6	82.0	70.7
Self-Cross(Ours)	**84.3**	**84.7**	**82.5**	**73.6**

Table 2. Average text-text similarities (↑) on different methods

%	Animal-Animal	SSD-2	2D-Spatial	3D-Spatial
SD2.1	81.68	73.01	78.71	77.73
INITNO	83.15	73.20	-	-
Self-Cross(Ours)	84.02	73.53	**80.4**	**80.0**
CONFORM	85.21	73.87	-	-
CONFORM+Ours	**85.88**	**74.91**	-	-

Table 3. Average text-text similarities (↑) on different methods

TIFA-GPT4o Scores TIFA [22] aims to assess the faithfulness of the generated image to the input prompt. It uses a vision-language model to perform Visual Question Answering (VQA) on the generated image, guided by specific questions about the image's contents. In our implementation, we leverage a more advanced vision-language model,

Methods	Animal-Animal			Animal-Object			Object-Object			SSD-2		
Question Types	Ext	Rec	w/o M	Ext	Rec	w/o M	Ext	Rec	w/o M	Ext	Rec	w/o M
SD1.4 [41]	39.51	29.70	72.24	67.84	53.72	90.36	34.15	31.89	94.22	30.77	28.09	77.47
INITNO [17]	89.39	77.09	82.26	98.37	**78.00**	95.97	96.20	90.42	95.17	61.34	55.53	79.70
CONFORM [31]	89.63	78.00	84.22	98.37	75.09	97.16	78.58	73.12	97.48	67.54	59.90	79.50
Self-Cross(Ours)	**94.55**	**87.79**	**92.94**	**99.60**	75.17	**98.30**	**98.95**	**93.19**	**98.65**	**77.67**	**70.92**	**86.45**

Table 4. TIFA-GPT4o Scores (↑) on four benchmarks: Animal-Animal, Animal-Object, Object-Object, and our proposed Similar Subjects Dataset. Inspired by TIFA [22], we employ GPT4o [1] as the VQA model to evaluate three aspects: Existence of both subjects (Ext), Recognizability of both subjects (Rec), and Absence of Mixing of subjects (w/o M). GPT4o is prompted with several True/False questions, and we report the percentage of True responses as the scores. The list of question prompts is given in the supplementary materials. Best results are highlighted in **bold** and second best results are shown with underline.

Methods	Animal-Animal			SSD-2			2D-Spatial				3D-Spatial			
Question Types	Ext	Rec	w/o M	Ext	Rec	w/o M	Ext	Rec	Rel	w/o M	Ext	Rec	Rel	w/o M
SD2.1 [41]	61.63	38.23	67.51	52.84	36.98	84.93[2]	76.16	73.37	32.46	90.21	69.1	66.17	47.72	90.33
INITNO [17]	80.96	47.18	63.80	68.70	46.10	73.85	-	-	-	-	-	-	-	-
Self-Cross(Ours)	89.53	55.03	78.79	77.35	45.36	77.08	**87.37**	**83.90**	**35.48**	**91.24**	**87.07**	**82.60**	**57.96**	**91.27**
CONFORM [31]	96.88	**70.19**	92.49	**83.71**	53.90	89.76	-	-	-	-	-	-	-	-
CONFORM+Ours	**97.83**	69.00	**94.66**	81.39	**55.98**	**93.20**	-	-	-	-	-	-	-	-

Table 5. Quantitative benchmarks with SD2.1 TIFA-GPT4o Scores (↑) on two challenging benchmarks: Animal-Animal and our proposed Similar Subjects Dataset. Best results are highlighted in **bold** and second best results are shown with underline. vanilla SD2.1 usually generates only one subject of the prompt so its 'w/o M' score is high.

Figure 5. Quantitative comparisons between original SD2.1 [41] and our method.

Figure 6. Quantitative comparisons between SD3-medium [41] and our method.

(a) 1 patch (b) 16 patches (c) masked patches

Figure 7. Our Self-Cross guidance works the best with masked patches, which verifies our assumption that all patches of a subject not just the most discriminant one need to be considered for eliminating subject mixing.

GPT4o [1], to provide more precise feedback. Given a text prompt in the form of "a class A and a class B", we developed questions to (1) verify the existence of both subjects; (2) confirm the recognizability of both subjects, ensuring no artifacts or distortions are present; and (3) check that the image does not exhibit a mixture of the two subjects, i.e., subject mixing. These three aspects capture both local and global information about the image, serving as comprehensive and reliable metrics including **Ext**, **Rec**, **w/o M**, and **Rel** scores respectively. W/o score is most relevant to our study of subject mixing, while Ext and Rec scores are effective in assessing the issue of subject neglect. Rel measures the spatial relationship of subjects. The full set of question prompts we designed are in the appendix C.

As shown in Table 4, when built on Stable Diffusion 1.4, our method is overwhelmingly better than baselines including SD [41], INITNO [17], and CONFORM [31] on Animal-Animal and Similar Subjects datasets(SSD-1), especially for reducing subject mixing (w/o M score). For animal-animal prompts, we achieved a w/o M score of **92.94%**, which is **8.7%** better than the second-best score (84.22% CONFORM). For similar subjects dataset(SSD-1), our w/o M score of **86.45%** is **6.7%** better than the second-best score (70.70% INITNO). Besides, for less challenging datasets with distinct subjects including Animal-Object and Object-Object, we still achieved better results.

Interestingly, although our method focuses on addressing $subjectmixing$, it can also reduce $subjectneglect$ and improve the recognizability/fidelity of generated subjects to some extent. However, visualization of the attention maps shows the inefficiency of $attend\&excite$ in Stable Diffusion 2.1. As we rely on $attend\&excite$ to encourage the existence of different subjects, our performance is influenced in Stable Diffusion 2.1. So, for Stable Diffusion 2.1, we apply CONFORM for the first few steps to ensure the existence of subjects and apply our method in later steps(indicated as CONFORM+ours), as shown in Tab. 5. To further prove the effectiveness of our method on DiT, we also implement our method on Stable Diffusion 3 medium as shown quantitatively in Tab. 1 and qualitatively in Fig. 6.

Text-text similarities Text-text similarities, also known as BLIP scores, are the similarity between captions generated by a vision-language foundation model [28] and the original prompts used to synthesize the images. This metric captures subjects and attributes from the original prompt, then measures the coherency and consistency of the generated content with the textual descriptions. As shown in Table 2, Table 3, and Table 1, compared with other methods, our approach, including the combined one, reaches SOTA performances on all benchmarks.

6. Conclusion and Future Work

Subject mixing remains a persistent issue for diffusion-based image synthesis, particularly for similar-looking subjects. We propose Self-Cross diffusion guidance to boost the performance of any diffusion-based image synthesis of similar subjects. Our method is motivated by the overlap between self-attention maps and cross-attention maps for mixed subjects, which is penalized by the proposed self-cross diffusion guidance loss during inference. Further more, we aggregate self-attention maps for multiple patches to a single attention map. In other words, our formulation involves all relevant patches of a subject beyond the most discriminant one. We are the first to reduce overlap between cross-attention and aggregated elf-attention maps, while previous methods are limited to the self-attention map from one patch or rely on cross-attention maps alone for guidance. We utilize standard gradient-based optimization and initial noise optmization [17] for minimizing our guidance loss during inference. We also released SSD, a new dataset of similar subjects for image synthesis, and leveraged the latest vision large language model (GPT-4o) for automatic and reliable evaluation of different methods. Qualitative and quantitative results show significant improvement over previous approaches. Our Self-Cross guidance greatly reduced subject mixing while also reducing the issue of object neglect as a side effect.

In the future, we will extend our approach to video generation of similar subjects that face challenges of subject mixing. We also anticipate the issue of subject mixing to be less prominent with newer backbone models but not disappear. We will keep exploring variants of Sef-Cross diffusion guidance for finer-grained subjects synthesis and addressing other issues such as attribute binding.

Acknowledgement This research was supported by the Department of Defense under funding award W911NF-24-1-0295 and by Google Cloud research credits program.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7, 8

[2] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 3

[3] Zhipeng Bao, Yijun Li, Krishna Kumar Singh, Yu-Xiong Wang, and Martial Hebert. Separate-and-enhance: Compositional finetuning for text2image diffusion models. In *SIGGRAPH*, 2024. 2, 3, 5

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, 2023. 3

[7] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023. 2

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 2, 3, 5, 4

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2, 3

[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 2

[11] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 2

[12] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 1

[14] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[15] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7545–7556, 2023. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[17] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 1

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[20] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 2, 3, 1

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[22] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 2, 6, 7, 4

[23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv: 2307.06350*, 2023. 5

[24] Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–17, 5555. 5

[25] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10124–10134, 2023. 2

[26] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7701–7711, 2023. 3

[27] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11523–11532, 2022. 2

[28] 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. 8

[29] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. 1

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 2

[31] Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9005–9014, 2024. 1, 2, 3, 5, 6, 7, 8, 4

[32] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2

[34] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1979. 4

[35] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7942, 2024. 3, 4

[36] 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

[37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 1

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. In *arxiv*, 2022. 2

[40] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8, 4

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[44] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pages 30105–30118. PMLR, 2023. 2

[45] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation, 2023. 2

[46] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[48] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16515–16525, 2022. 2

[49] Yifan Zhou Xingang Pan Tianyi Wei, Dongdong Chen. Enhancing mmdit-based text-to-image models for similar subject generation. *https://arxiv.org/abs/2411.18301*, 2024. 1

[50] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 2

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[52] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8553–8564, 2024. 3

[53] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 3

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 2

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. Featured Certification. 2

[56] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multimodal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023. 2

[57] Yanbing Zhang, Mengping Yang, Qin Zhou, and Zhe Wang. Attention calibration for disentangled text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4764–4774, 2024. 3

[58] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. *arXiv preprint arXiv:2403.05121*, 2024. 2

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019. 2

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

Supplementary Material

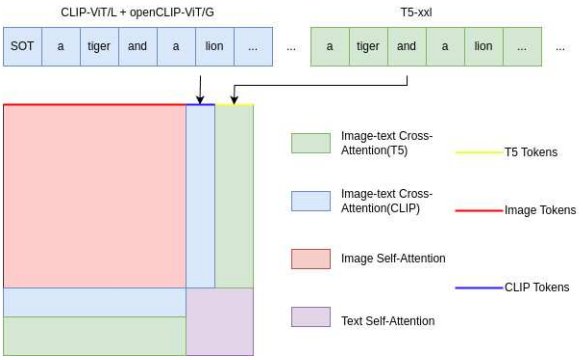


Figure 8. attention maps in multimodal diffusion transformer.

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.

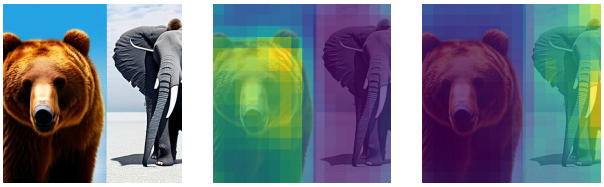


image self-attn "bear" self-attn "elephant"

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

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

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

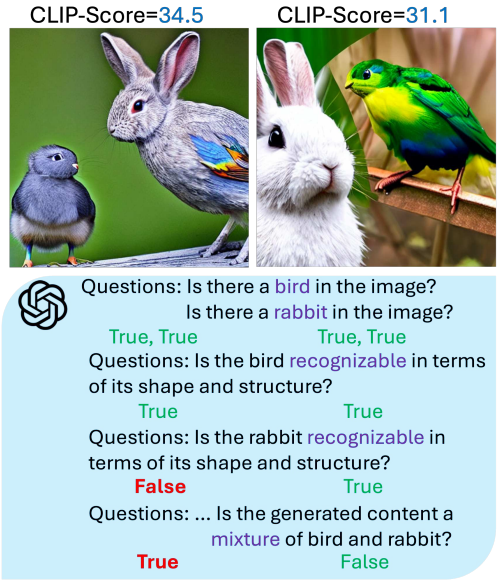


Figure 11. (Left): image generated by CONFORM [31]; (right): image generated by our approach under the same seed. Left image shows a higher CLIP score. However, there are obvious content mixing issues in the left image, which GPT4o is able to capture with VQA. This is an example that CLIP score is not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

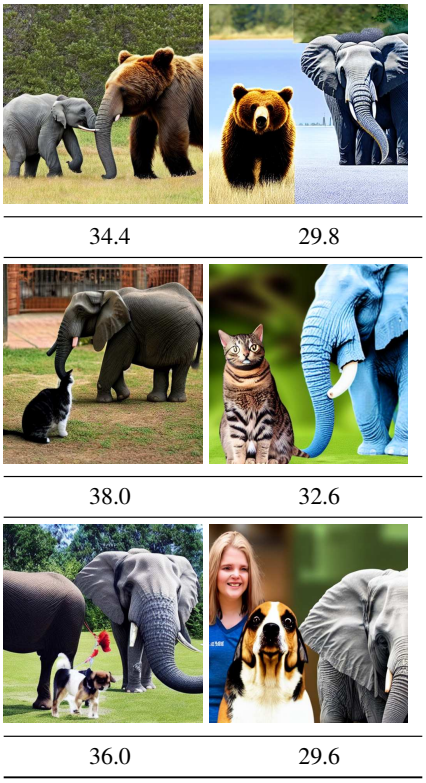


Figure 12. CLIP Scores (↑) for synthetic images generated by CONFORM [31] (left) and our self-cross guidance (right). CLIP scores are unreliable for measuring image quality w.r.t. subject mixing.

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.

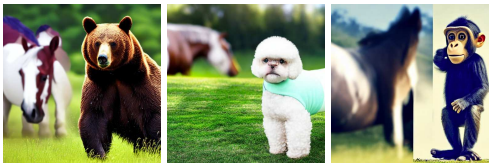
Unlike Attention Refocusing, which relies on a language model to pre-define the layouts, our method operates independently of external knowledge, making it more versatile and applicable to a wider range of scenarios. The superior results in existence and recognizability highlight our approach's ability 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-

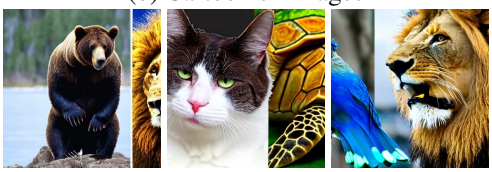
Metric (↑)	SD1.4 [41]	Attn-Refocus [35]	Self-Cross (Ours)
Ext	39.51	86.99	**94.55**
Rec	29.70	64.45	**87.79**
w/o M	72.24	**93.80**	92.94
CLIP score	31.0	**33.9**	33.2
Text sim	76.5	79.8	**84.3**

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.

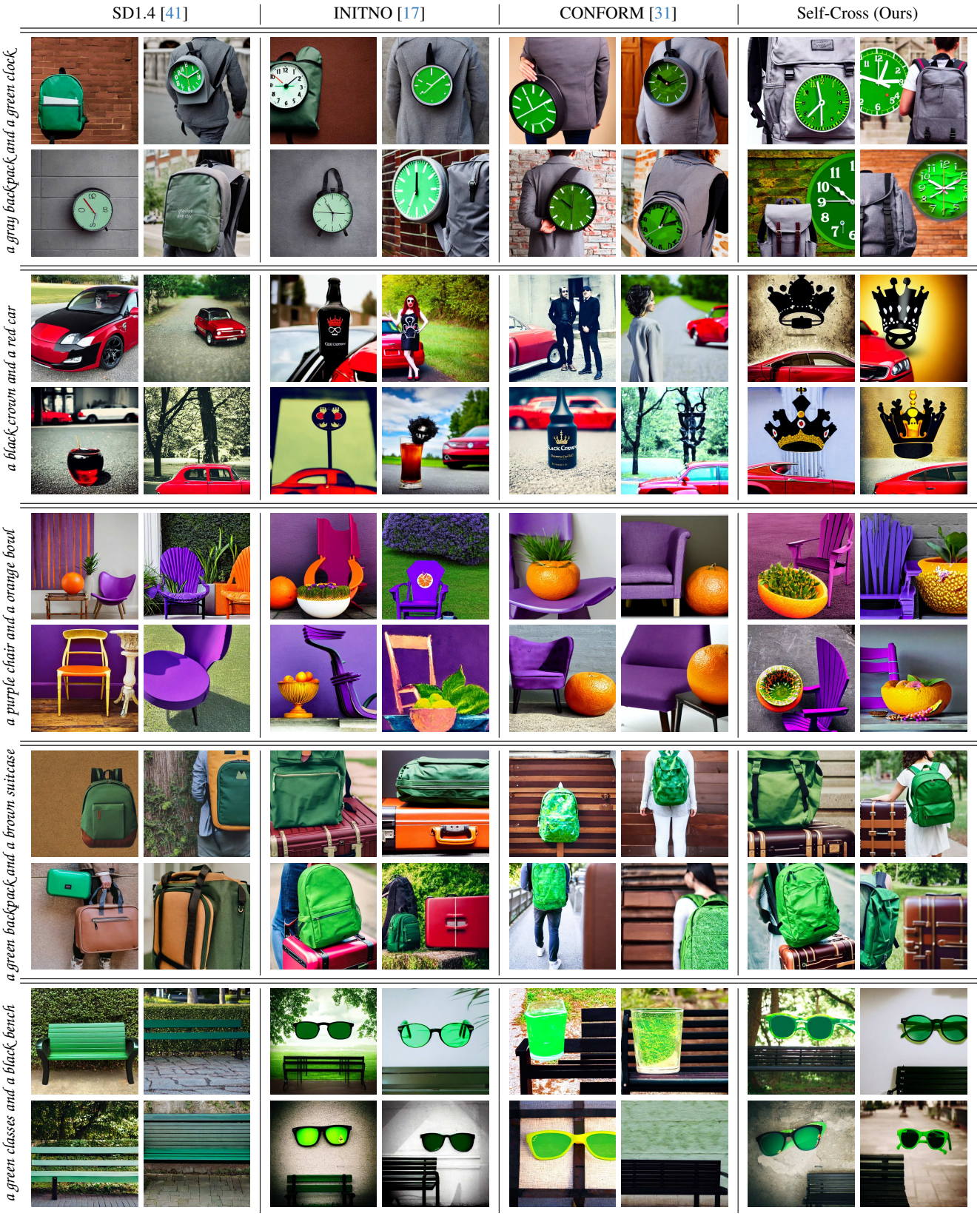


Figure 14. More qualitative comparisons of Self-Cross (ours) to SD1.4 [41], INITNO [17], CONFORM [31]. For each prompt in the left column, we sample four seeds and show the results of different methods.

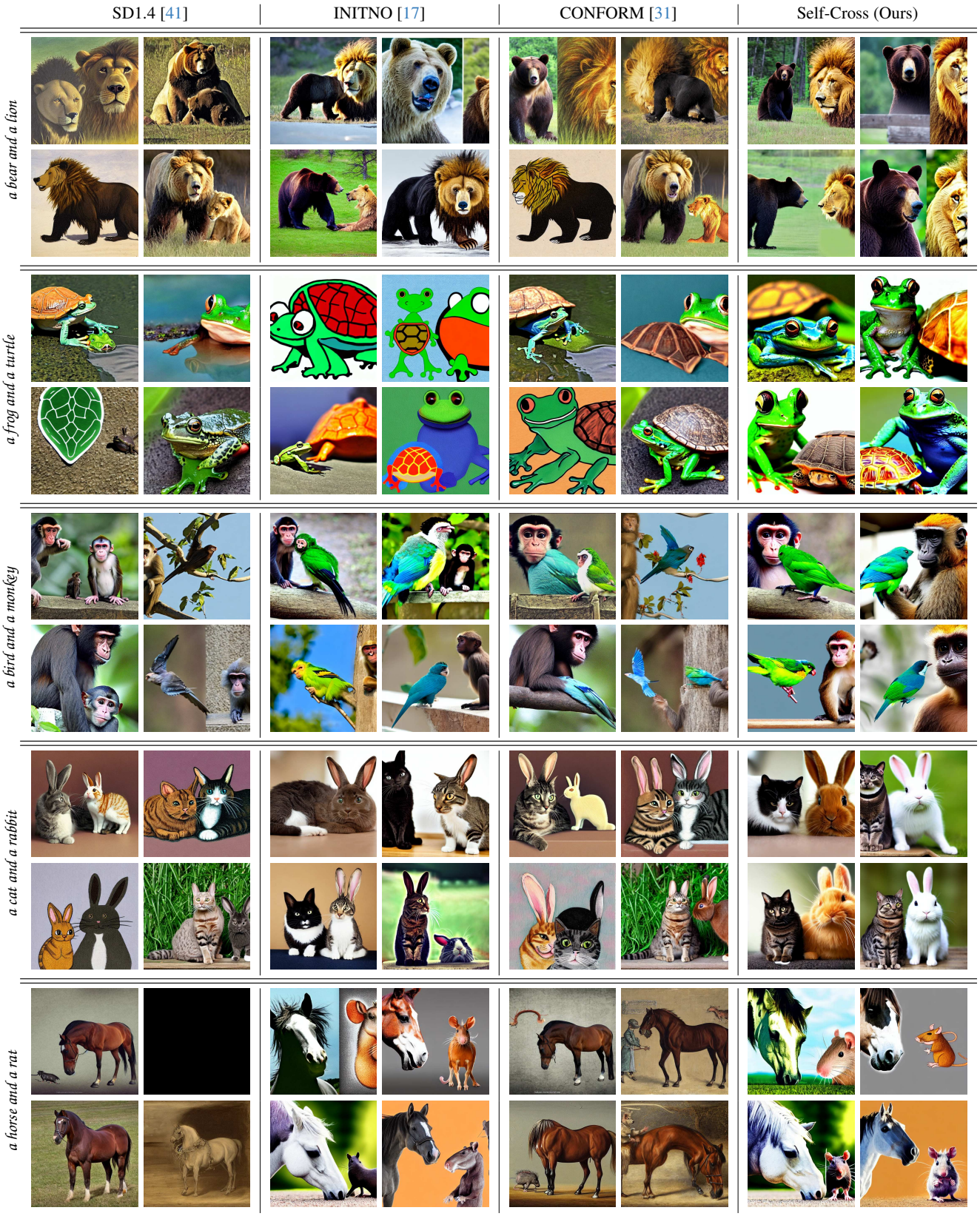


Figure 15. More qualitative comparisons of Self-Cross (ours) to SD1.4 [41], INITNO [17], CONFORM [31]. For each prompt in the left column, we sample four seeds and show the results of different methods.