A-STAR: Test-time Attention Segregation and Retention for Text-to-image Synthesis (Supplementary Material)

Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena,
Koustava Goswami and Balaji Vasan Srinivasan
Adobe Research, Bengaluru India
{aishagar,skaranam,josephkj,apoorvs,koustavag,balsrini}@adobe.com

1. Appendix

In Section 1.1, we show more results for both attention overlap and attention decay as well as provide evidence to show baseline Stable Diffusion [3] does not degrade with our losses in cases where it is already capturing all concepts in the prompt. In Section 1.2, we give more implementation details. In Section 1.3, we provide additional quantitative results where we report results of an ablation experiment that calculates CLIP image-text similarities like Figure 9 in the main paper. We also report SentenceTransformer based text-text similarity scores in this section. In Section 1.4, we provide additional qualitative results comparing our method with Attend-Excite [1] on top of baseline Stable Diffusion. Finally, we conclude with some discussion on limitations of our method in Section 1.5.

1.1. Attention Overlap and Attention Decay

As discussed in the main paper, we identified two key issues with existing diffusion models: attention overlap and attention decay. Here, we show more examples.

In Figure 1, we demonstrate the issue of attention overlap with four examples. We notice that overlapping high-response regions in the attention maps lead to the *elephant* getting missed in the generated output image in the first example, the *dog* in the second example, and the *man* in the third example. For instance, in the first example, there is significant overlap in the regions that correspond to high activations for both *elephant* and *giraffe* attention maps. Since they are highly activated in the same pixel regions, the final generated image is unable to distinguish between the two subjects and is able to pick only one of the two. Similar reasoning follows for examples in columns 2 and 3. In column 4, we demonstrate the issue of incorrect attributes getting binded to the subjects due to attention overlap. Here, the attention map of *bowl* has high responses for the same regions where the *turtle* and other objects, leading to a mixup in the properties of the turtle and bowl (see final image where even the *turtle* is yellow).

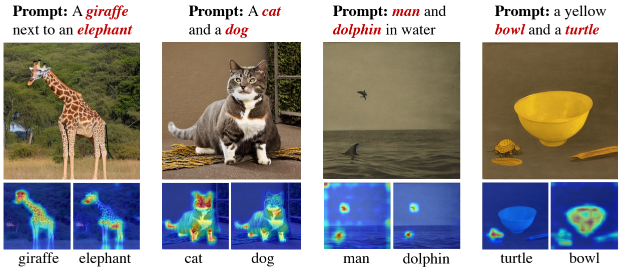
Prompt: A *giraffe* next to an *elephant*
Prompt: A *cat* and a *dog*
Prompt: *man* and *dolphin* in water
Prompt: a yellow *bowl* and a *turtle*
giraffe
elephant
cat
dog
man
dolphin
turtle
bowl

Figure 1: Examples demonstrating the issue of attention overlap in baseline Stable Diffusion

In Figure 2, we show more examples to demonstrate the issue of attention decay. In the first example, one can note the *ship* is missing in the baseline image output. Looking at the cross-attention maps for *ship* across the denoising timesteps with baseline Stable Diffusion, it is clear that the information for this concept is present at the beginning but is not retained towards the end. Concretely, the pixel regions that were initially highly activated in the *ship* attention map is very sparsely activated at the end. This results in the *ship* not showing up in the final generation. A similar phenomenon can be observed with the *forest* concept in the other example. Note that in both cases, with our proposed method (see A-STAR attention maps), we are able to correct this issue.

In Figure 3, we provide results for baseline Stable Diffusion and A-STAR for a set of prompts where the baseline model already captures the input semantics well. The motivation of this experiment is to show that in these cases, with our proposed losses, we are not degrading baseline performance. Let us consider the first example (top left) where the baseline model already has well-separated attention maps for *bird* and *garden*, resulting in both concepts being captured in the generated image. In this case, even after applying our losses with A-STAR, there is no degradation in the generated image and both concepts show up. Similarly, in

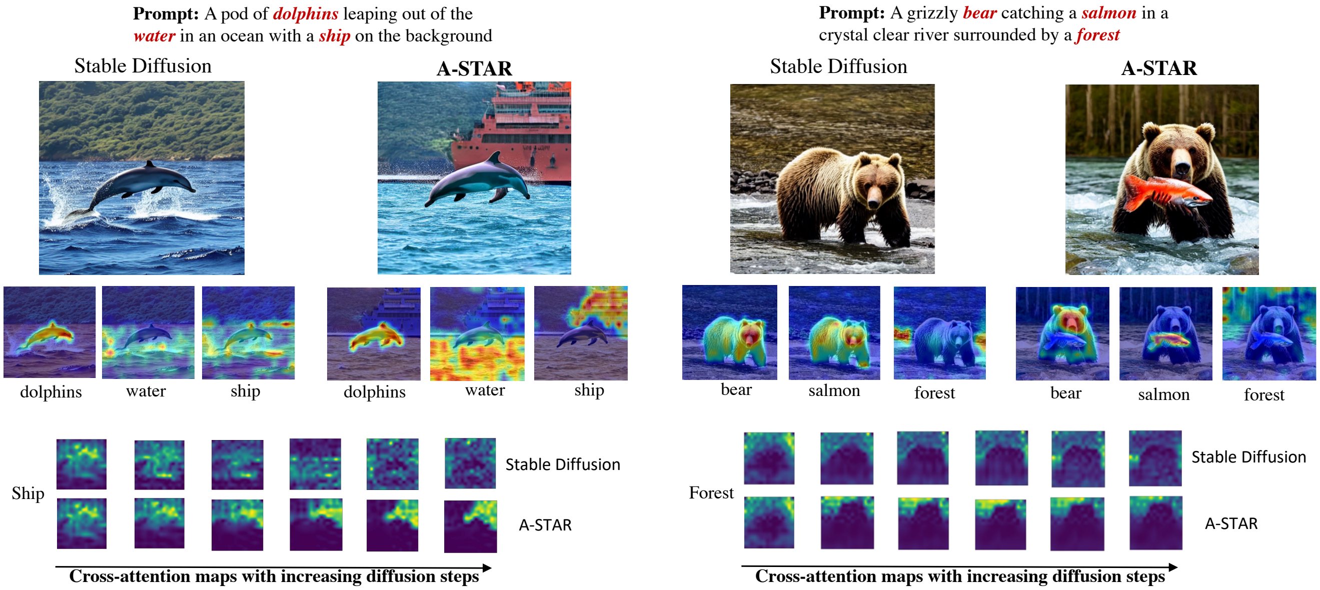
Prompt: A pod of dolphins leaping out of the water in an ocean with a ship on the background
Stable Diffusion
A-STAR
dolphins
water
ship
dolphins
water
ship
Ship
Stable Diffusion
A-STAR
Cross-attention maps with increasing diffusion steps
Prompt: A grizzly bear catching a salmon in a crystal clear river surrounded by a forest
Stable Diffusion
A-STAR
bear
salmon
forest
bear
salmon
forest
Forest
Stable Diffusion
A-STAR
Cross-attention maps with increasing diffusion steps

Figure 2: Examples demonstrating the issue of attention decay in baseline Stable Diffusion

the second example (top right), the baseline model has well-separated attention maps for *cat* and *table*, and this remains the case after applying the A-STAR losses, leading to both models giving the desired output. Similar observations can be made from the other two examples as well.

1.2. Implementation Details

Given an input text prompt, we consider all the possible subjects (e.g., nouns) while computing the two proposed losses. Let $\mathcal{C}$ denote the set of subjects identified given the prompt. To compute the losses $\mathcal{L}_{seg}$ and $\mathcal{L}_{ret}$, we first normalize the outputs of the cross-attention layers from the DDPM model to a range between 0 and 1 to obtain the attention maps $\mathbf{A}_t^m \; \forall m \in C$. Note that $\mathcal{L}_{seg}$ considers all possible pairs of subjects present in the input text prompt. We next discuss how we compute the ground truth binary mask $\mathbf{B}_{t-1}$ used in $\mathcal{L}_{ret}$. Given the attention maps $\mathbf{A}_t^m$ for a subject m at timestep t, we first determine a bounding box for the pixel regions with high activations and set all pixels within the bounding box to be 1 (and rest to 0), giving us the binary mask. Note that the mask computed at timestep t gets utilised in the $\mathcal{L}_{ret}$ at timestep $t-1$.

1.3. Additional Quantitative Results

In the main paper in Fig 9, we showed CLIP image-text similarity comparisons with several existing diffusion models. In Fig 4 in this supplementary document, we show this graph for an ablation experiment to demonstrate the individual impact of our proposed losses. As can be seen from Fig 4 here, across all the three scenarios, while each of attention segregation and attention retention losses improve the performance of baseline Stable Diffusion, we obtain the best performance when both of them are used in conjunction.

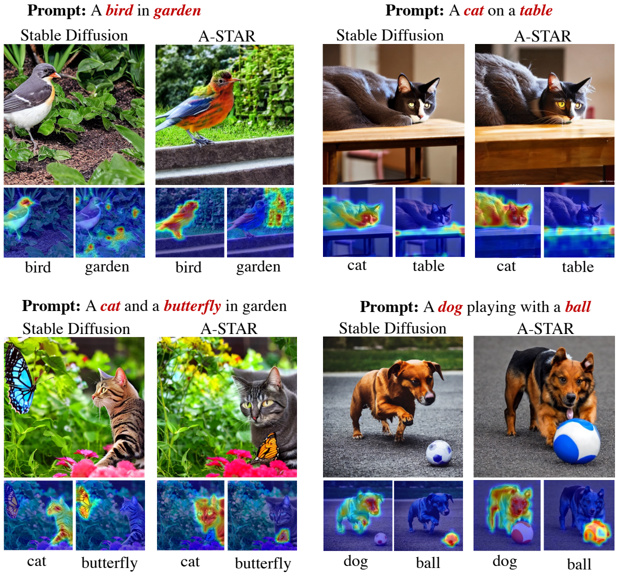

Figure 3: The first column shows the generated images and the corresponding attention maps for a set of prompts where baseline Stable Diffusion captures the text semantics well in the generation. In the second column, we show the generations for the same prompt and seed using A-STAR in order to demonstrate that A-STAR does not degrade the quality of generated images in terms of capturing semantics where baseline Stable Diffusion already works well.

Method	Animal - Animal	Animal - Object	Object - Object
Stable [3]	0.59	0.68	0.63
Attend-Excite [1]	0.66	0.74	0.72
A-STAR	**0.68**	**0.75**	**0.73**

Table 1: Text-text similarities between the text prompts and BLIP-generated captions over the generated images.

In Table 1, we show results corresponding to Table 1 in the main paper with cosine similarities computed using SentenceTransformer [2] embeddings. Specifically, we take the input prompt and the BLIP-generated caption, compute their respective SentenceTransformerembeddings, and compute their cosine similarities. As can be seen from Table 1 here, A-STAR outperforms the baselines across all the three categories.

1.4. Additional Qualitative Results

In Fig 5 here, we show more qualitative results comparing the performance of our proposed method with Attend-Excite on top of baseline Stable Diffusion. In each case, A-STAR gives more photorealistic imagery that captures all the input concepts. For instance, in the second column, A-STAR has both bear and turtle clearly captured in the final generation whereas both baseline Stable Diffusion and Attend-Excite fail. Similarly, in the fourth example, A-STAR generates both the horse and the bird whereas the other models either miss out one or both of these concepts.

1.5. Limitations

In this section, we discuss a few limitations of our proposed method. In Figure 6(a), both the baseline model as well as A-STAR generate the *red carpet* and the *table* but lack an understanding of the relationship between the two concepts. In such cases, A-STAR is limited by the capabilities of the base model and as we discussed in both our proposed losses, we are currently not accounting for explicit relationship modeling between the concepts. However, given a computational model that captures these relationships, it can conceivably be added to our losses to reflect these relationships in the final output.

In Figure 6(b), while A-STAR ensures both concepts (*man* and *dolphin* in first and *giraffe* and *elephant* in second) are captured in the final image, it may perhaps be more desirable to have these images generated at specific camera poses/viewpoints so as to capture these concepts more holistically. With advances in the ability to control diffusion model outputs [4], we can integrate our losses with such controlled generation techniques to improve these aspects.

References

[1] 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. *arXiv preprint arXiv:2301.13826*, 2023. 1, 3

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 3

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3

[4] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3

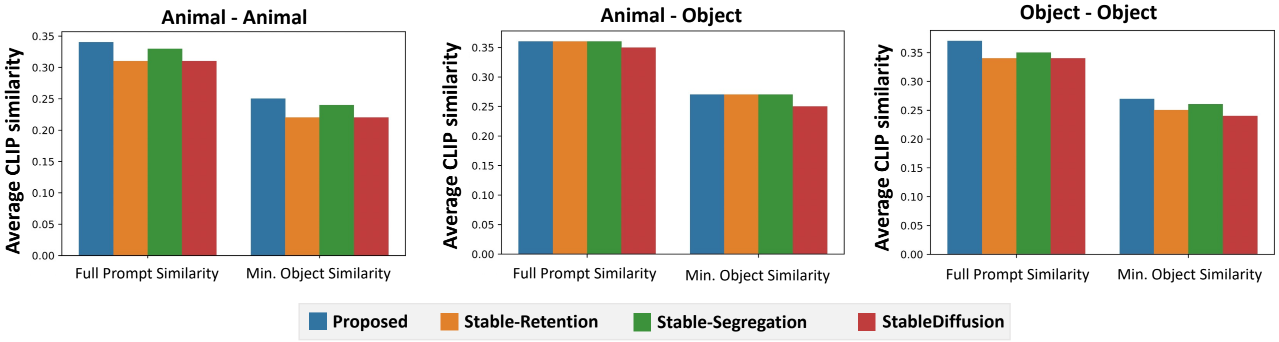

Figure 4: Ablation Study: Comparing Average CLIP image-text similarities between the text prompts and generated images

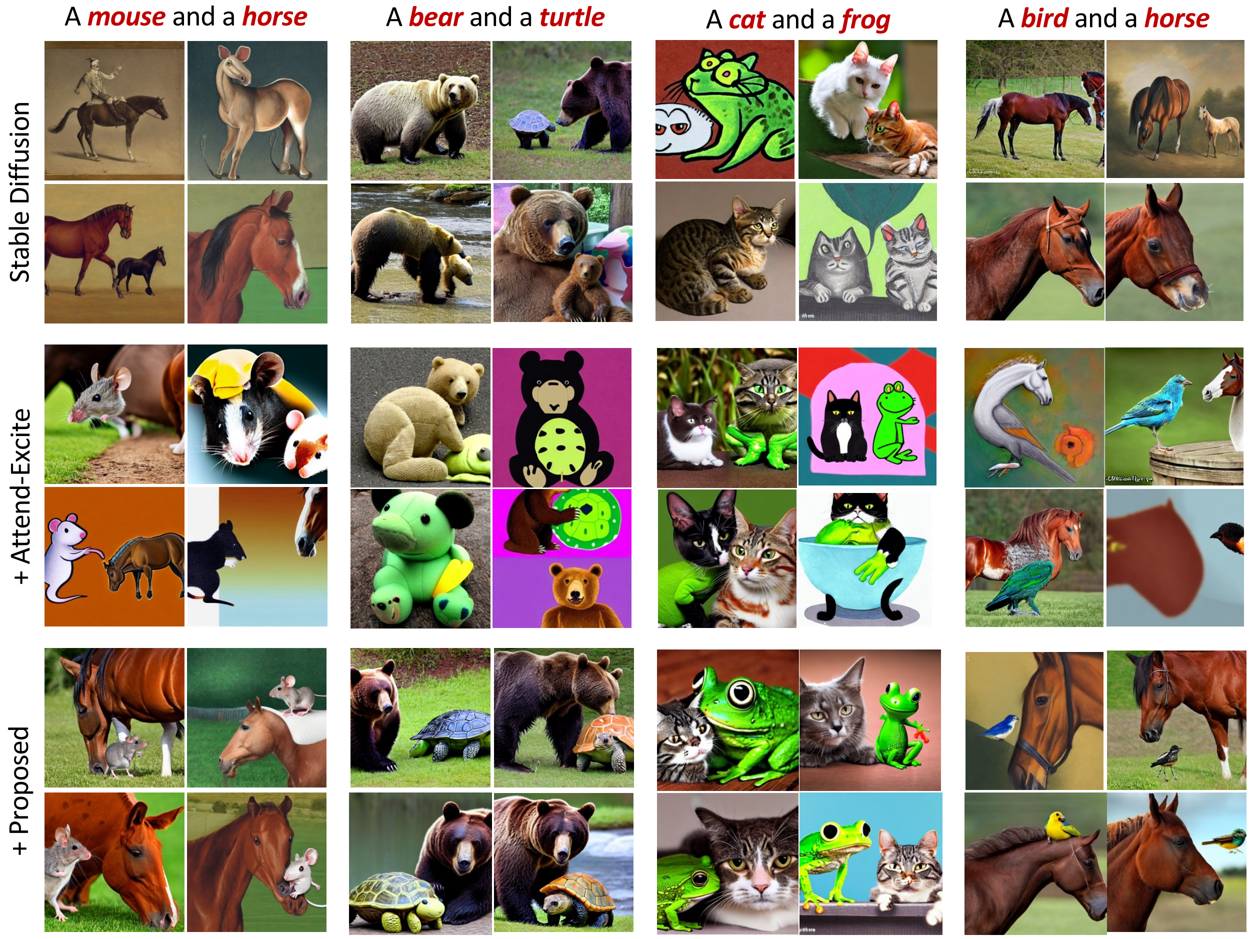

Figure 5: More comparison results of the proposed method vs Attend-Excite applied on top of base Stable Diffusion.

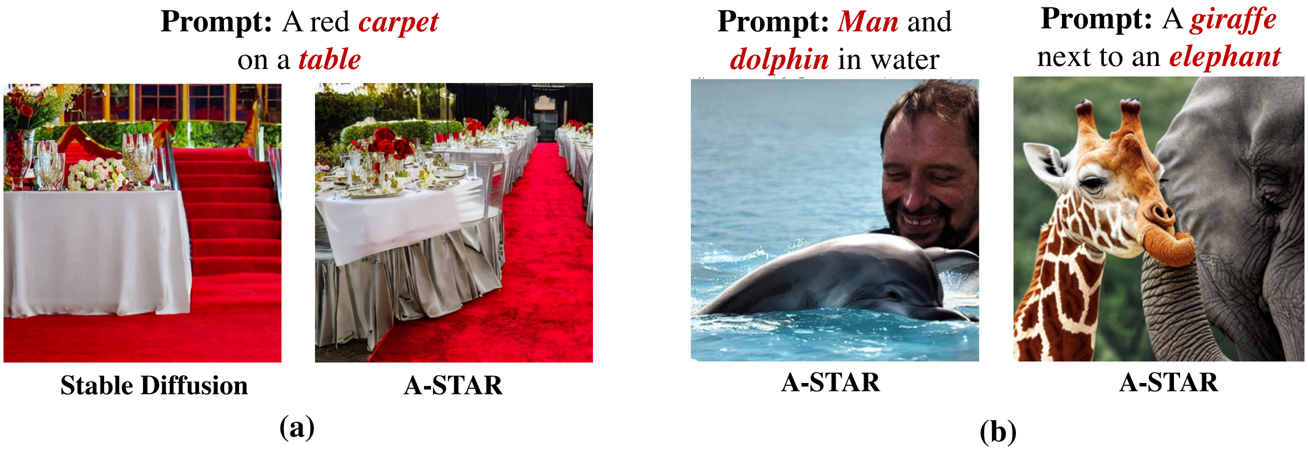
Prompt: A red carpet on a table
Stable Diffusion
A-STAR
(a)
Prompt: Man and dolphin in water
A-STAR
Prompt: A giraffe next to an elephant
A-STAR
(b)

Figure 6: A-STAR limitations.