ICCV #1621

Penalizing Boundary Activation for Object Completeness in Diffusion Models

Supplementary Material

Summary

This supplementary material encompasses five components:

A. Exploration of object incompleteness across different T2I models.
B. Ablation study on another Stable Diffusion version.
C. Comprehensive visualization comparisons demonstrating the changes in attention maps.
D. Extended results comparison between Stable Diffusion and our method.
E. Demonstration of failure cases.

A. Object Incompleteness across Different T2I Models

In Sec. 3, we demonstrated through experiments that the issue of object incompleteness exists in Stable Diffusion 2.1. Building upon these findings, we further investigate whether this issue also affects other T2I models and different versions of Stable Diffusion. To this end, we conducted additional experiments using a variety of models, including DALL-E mini [1], GLIDE [2], and Stable Diffusion 1.5.

As shown in Tab. 6, for all models, the object incompleteness rate (OIR) of the results, is greater than 50%. Specifically, DALL-E mini [1] achieved 57.3%, GLIDE [2] demonstrated 59.7%, and both Stable Diffusion versions, 2.1 and 1.5, exhibited performances of 54.5% and 52.3%, respectively. These findings suggest that the incompleteness problem is not specific to a single model or version but is a common issue across different T2I architectures. This reinforces the need for further exploration into methods to mitigate object incompleteness in image generation tasks.

T2I Model	Complete	Incomplete	OIR
DALL-E mini [1]	406	544	57.3%
GLIDE [2]	383	567	59.7%
Stable Diffusion 1.5	432	518	54.5%
Stable Diffusion 2.1	453	497	52.3%

Table 6. Statistics on the incompleteness of objects of different T2I models. OIR is the abbreviation for object incompleteness rate.

Method	OIR ↓	Time Cost (s) ↓	CLIP-IQA ↑
Stable Diffusion 1.5	52.3 %	5.16	0.670
Ours w/o Cross Constraint	45.6 %	5.49	0.641
Ours w/o Self Constraint	40.7 %	5.51	0.659
Ours	**19.1 %**	5.55	**0.706**

Table 7. Ablation study on whether to use the cross-attention constraint and self-attention constraint of our method across three different metrics, where OIR is the abbreviation for object incompleteness rate.

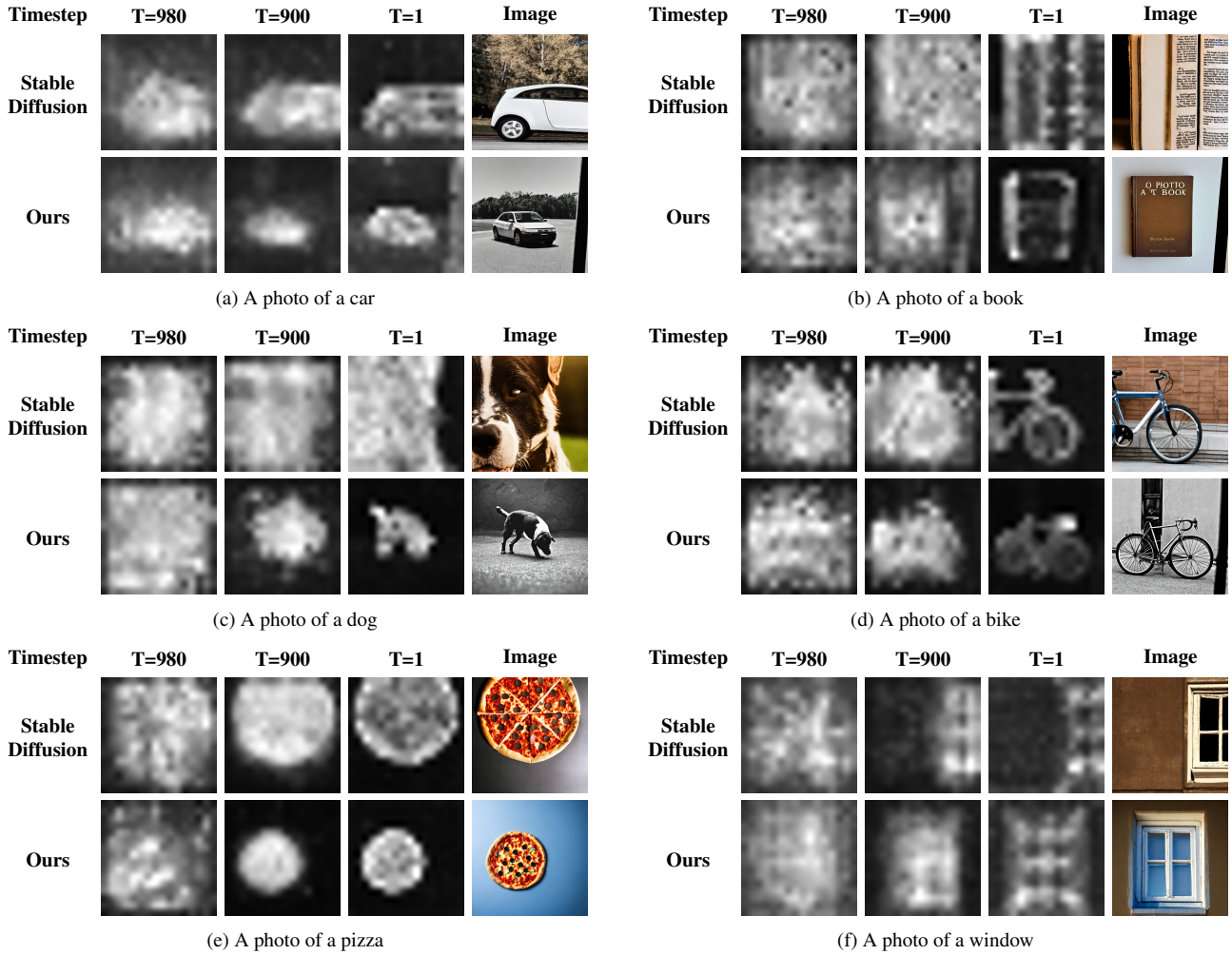
B. Ablation Study on Stable Diffusion 1.5

We conducted an ablation study on Stable Diffusion 2.1 in the main manuscript to demonstrate the effectiveness of the two constraints in our method. To further validate that our method can be applied to different Stable Diffusion models, we deployed the same ablation study on Stable Diffusion 1.5.

As shown in Tab. 7, our method performs best in OIR, the incompleteness rate is 19.1%, which is significantly lower than the original model of 52.3%. This result is consistent with our test results on Stable Diffusion 2.1, indicating that our method has significant effectiveness across multiple models.

Impact of the cross-attention constraint. In the ablation studies of Stable Diffusion 1.5, we removed the cross-attention guidance and maintained all other settings constant. As shown in the second row in Tab. 7, excluding cross-attention constraints caused the object incompleteness rate to rise significantly by 26.5%. Without cross constraint, it struggles to effectively associate their spatial placement with semantic meanings and the alignment between an object's description and its appearance deteriorates, leading to suboptimal image generation. This is further supported by a decline in the CLIP-IQA score by 0.065 when the cross-attention constraint is removed. In terms of efficiency, removing cross-attention reduced image generation time by just 0.06 seconds per sample, which accounts for a small fraction (just over 1%) of the overall time cost. These findings demonstrate again that cross-attention guidance is indispensable for generating complete and high-quality images.

Impact of the self-attention constraint. To investigate the impact of the self-attention constraint, we conducted experiments only using cross-attention guidance. The results, detailed in the third row of Tab. 7, show that removing self-attention constraints increased OIR by 21.6%. Self-attention reveals the layout assigned by the model to objects in the image. Without it, an object struggles to break away from the original layout, resulting in its continued incompleteness. Although the absence of self-attention caused a smaller reduction in the CLIP-IQA score (0.047) com-



(a) A photo of a car

(b) A photo of a book

(c) A photo of a dog

(d) A photo of a bike

(e) A photo of a pizza

(f) A photo of a window

Figure 8. Visualization comparison on the variation of attention maps across time steps.

pared to cross-attention, it still impacted object positioning and structure negatively. From a computational perspective, eliminating self-attention reduced image generation time by 0.04 seconds per sample. This reduction is minor.

In conclusion, cross-attention and self-attention constraints contribute distinct yet complementary benefits. The cross-attention constraint ensures semantic completeness and object detail, while the self-attention constraint refines spatial accuracy and structural coherence. The integration of both constraints yields superior image quality, as demonstrated by the lowest OIR and the highest CLIP-IQA score in the last row of Tab. 7. This result highlights the importance of combining the two mechanisms to achieve complete objects in images while maintaining high quality.

C. Visualization of Attention Maps

Fig. 8 shows the changes in the attention maps after applying our method. Each set of images displays the results from Stable Diffusion and those after applying our method, along with the attention maps during the denoising process. We can observe that at the initial $T = 980$, the noise is nearly identical. However, by $T = 900$, our method rapidly enhances the complete representation of the object, causing a significant difference in the attention maps and the final image's completeness during the subsequent denoising process.

D. Extended Results

Fig. 9 shows more generated complete samples from our method. Each prompt in the figure corresponds to four images: on the left column are the incomplete results from Stable Diffusion, and on the right are the results after applying our method. The images on the left contain only partial structures of the objects, while our method effectively generates the complete objects, including all their parts.

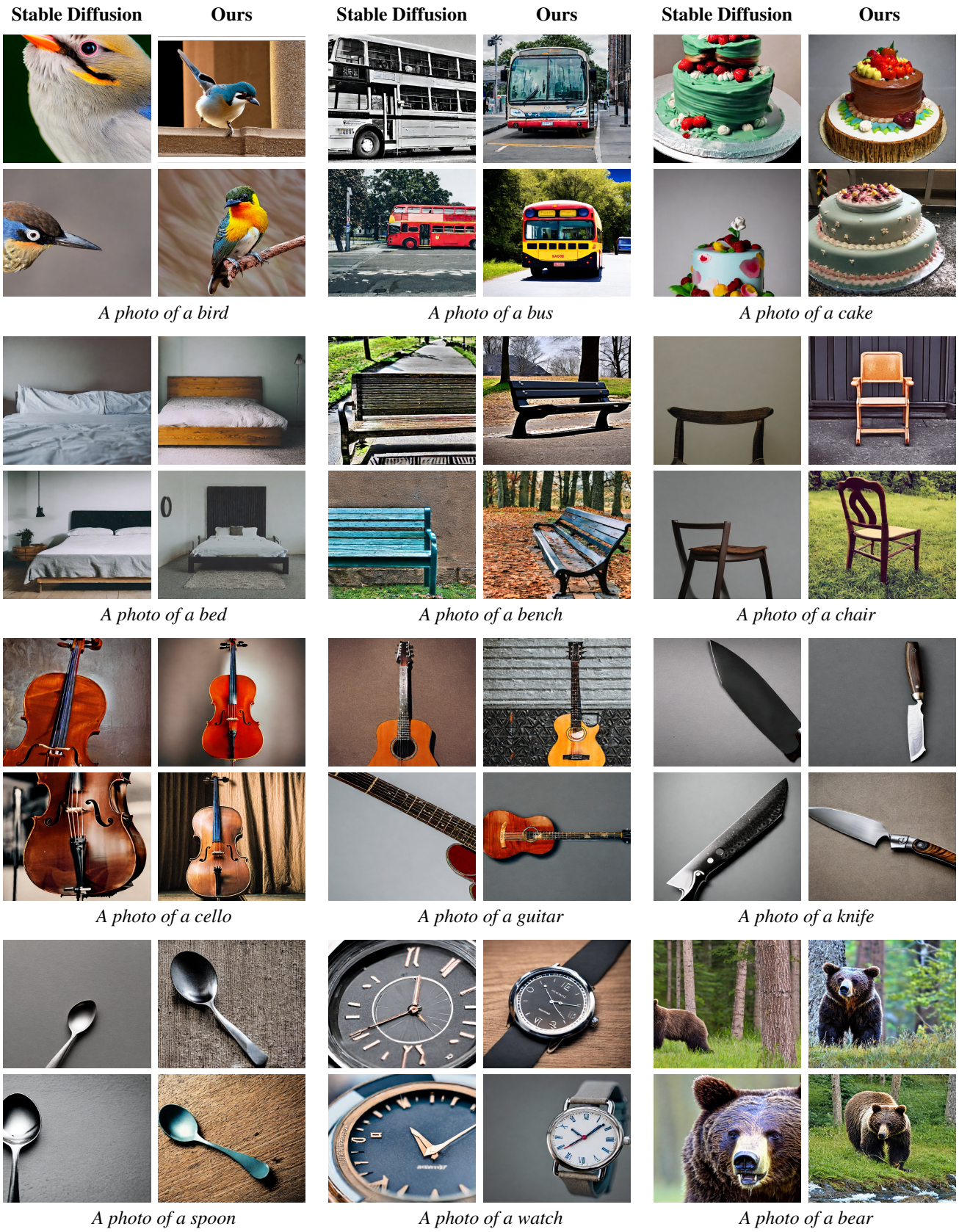


Figure 9. Visualization comparison between Stable Diffusion and our method.

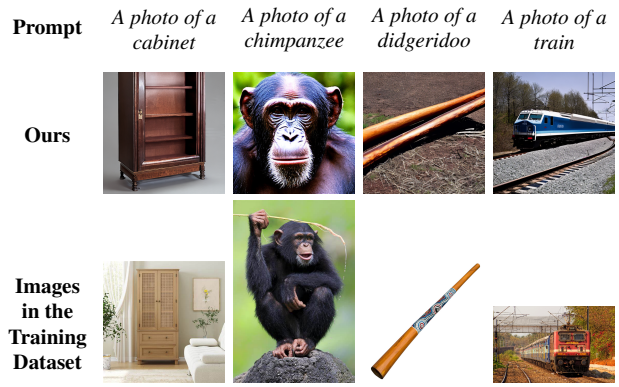


Figure 10. Visualization of failure cases in our method and their comparison with images in the training dataset.

E. Failure Cases

In the statistical results, our method still struggles to address the incompleteness issue in a few cases. As shown in Fig. 10, we listed several typical cases and compared them with original images in the training dataset. The results in the first column show objects that emphasize details, causing parts to remain outside the image boundaries, such as the cabinet. The second column's results show objects with a strong tendency to generate incomplete results, such as the chimpanzee, which often results in only the face or other representative parts being generated. The third column presents elongated objects that are difficult to fully display within one image, leading to incomplete results, as seen with the didgeridoo. The fourth column presents large, continuous objects, such as the train, which, due to its size, sometimes results in incomplete generations. Despite these challenges, our method demonstrates significant improvement in most cases, and with further refinement, it has the potential to better handle these specific failure scenarios.

References

[1] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 2021. 2, 1

[2] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*. PMLR, 2022. 2, 1