INITNO: Boosting Text-to-Image Diffusion Models via Initial Noise Optimization

Supplementary Material

A. Scalability of INITNO

The proposed method, *i.e.*, INITNO, demonstrates adaptability to various text-to-image diffusion models. Visual results generated by integrating INITNO with SD v1.4, SD v1.5, and SD v2.1 are presented in Fig. 1, 2, and 3, respectively. Our approach consistently led to improvements across these versions. It is also noticeable that enhancements to the foundational model can improve the alignment of the generated images with the text prompts to some extent, SD v2.1 performs better than SD v1.4 and SD v1.5. However, these models still grapple with issues such as subject neglect, subject mixing, and incorrect attribute binding, which our method effectively handles.

Furthermore, INITNO is complementary to existing methods and can mutually enhance performance. As shown in Fig. 4, INITNO can be integrated into Attend-and-Excite [2] to boost performance. Specifically, we remove the self-attention conflict loss and only add the initial noise optimization process before the Attend-and-Excite denoising process. Intuitively, INITNO directs noise towards the valid region at the initial denoising timestep, substantially reducing the burden on Attend-and-Excite.

B. Additional qualitative comparison

Fig. 5 and 6 present additional qualitative comparisons on complex text prompts. It can be observed that our method generates semantically more plausible and photorealistic results than its counterparts, successfully capturing all input concepts.

C. Diversity evaluation

To assess the diversity of images generated by the proposed method, we randomly sample 30 noises for the given text prompt and synthesize the corresponding images. As depicted in Fig. 7, INITNO maintains a high level of diversity in generated images.

D. Visualization of the attention maps

Additional visualizations of attention maps are provided in Fig. 8. Our method facilitates a reasonable allocation of attention, thereby yielding semantically consistent results.

E. Grounded text-to-image synthesis

Fig. 9 presents more visual results on grounded text-to-image synthesis. Owing to sufficient and appropriate optimization, our method achieves superior alignment between generated images and the given text and layout conditions.

F. Limitations

While our approach enhances the fidelity of given text prompts, several limitations merit consideration. First, our approach is constrained by the expressive capacity of the foundational model, as corroborated in Fig. 1, 2, and 3. If the text prompts fall outside the distribution of text descriptions learned by the foundational model, the resulting images may not correspond to the text prompts.

Second, our method relies solely on the first timestep of the denoising process to provide cross-attention maps and self-attention maps for initial latent space partitioning. Despite the first step offering the strictest constraints, dominating subsequent steps, it is partly due to computational overhead, as extracting information from all timesteps is computationally prohibitive. Fortunately, several studies [6, 7, 9] are exploring fewer denoising timesteps (≤ 5) for synthesizing high-quality images, which hold promise for addressing these challenges.

References

[1] 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 *ICCV*, 2023. 4, 5

[2] 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. In *SIGGRAPH*, 2023. 1, 4, 5

[3] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 4, 5

[4] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. In *BMVC*, 2023. 4, 5

[5] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 4, 5

[6] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 1

[7] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 1

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz,

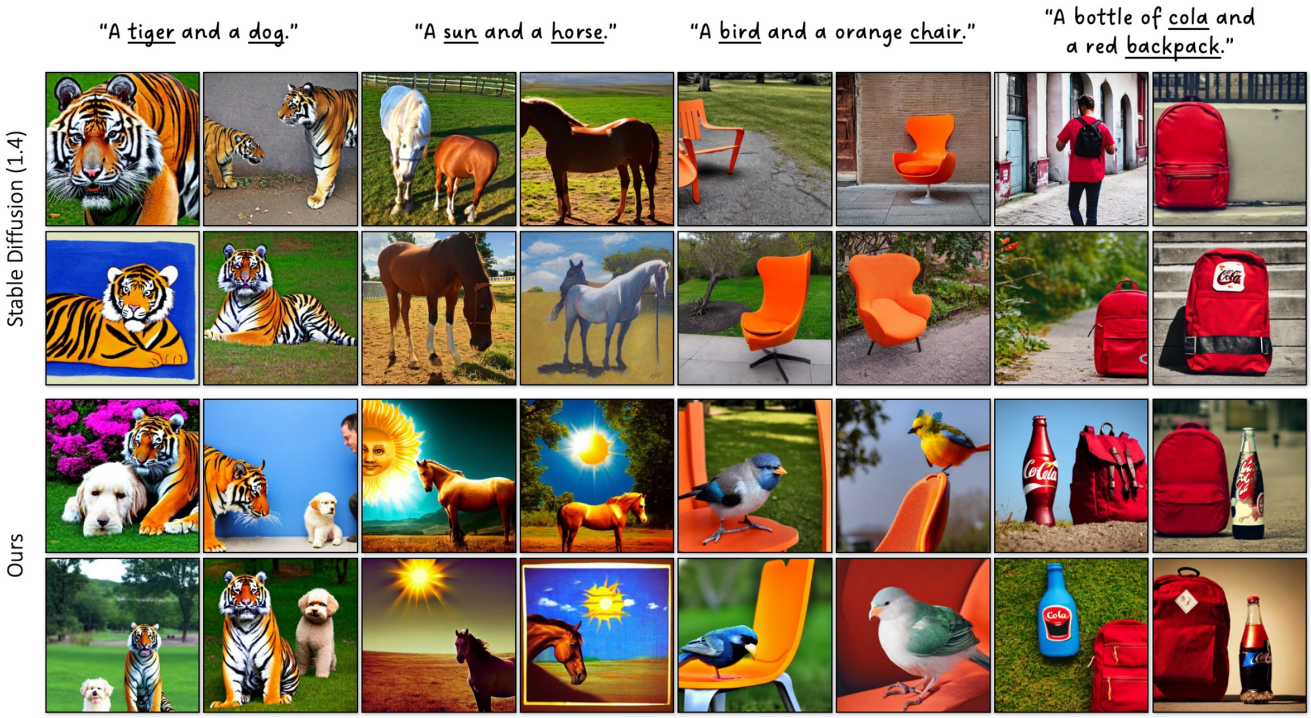


Figure 1. **Example results synthesized by SD v1.4 with INITNO.**

Figure 2. **Example results synthesized by SD v1.5 with INITNO.**

Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4, 5

[9] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 1

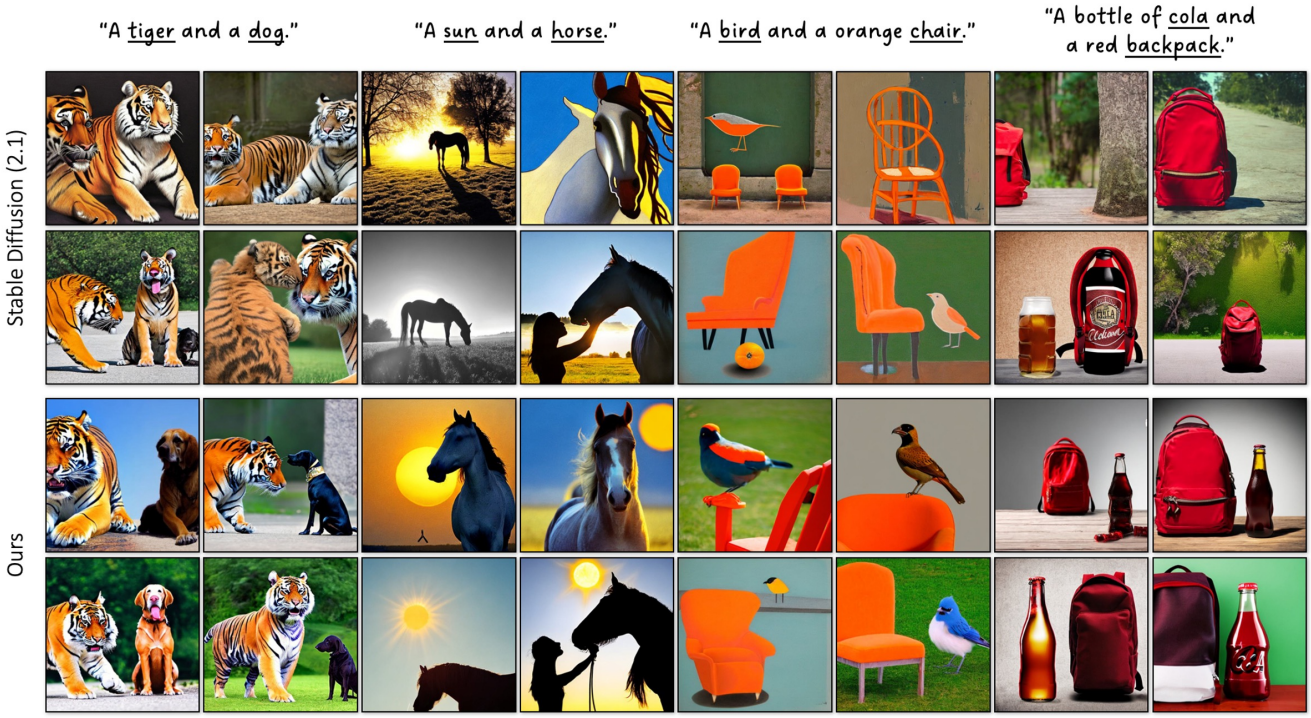


Figure 3. **Example results synthesized by SD v2.1 with INITNO.**

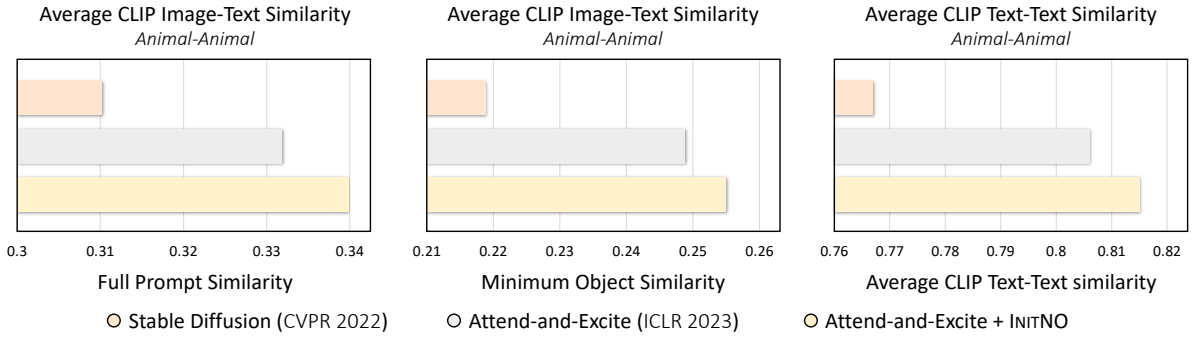


Figure 4. **Objective evaluation.**

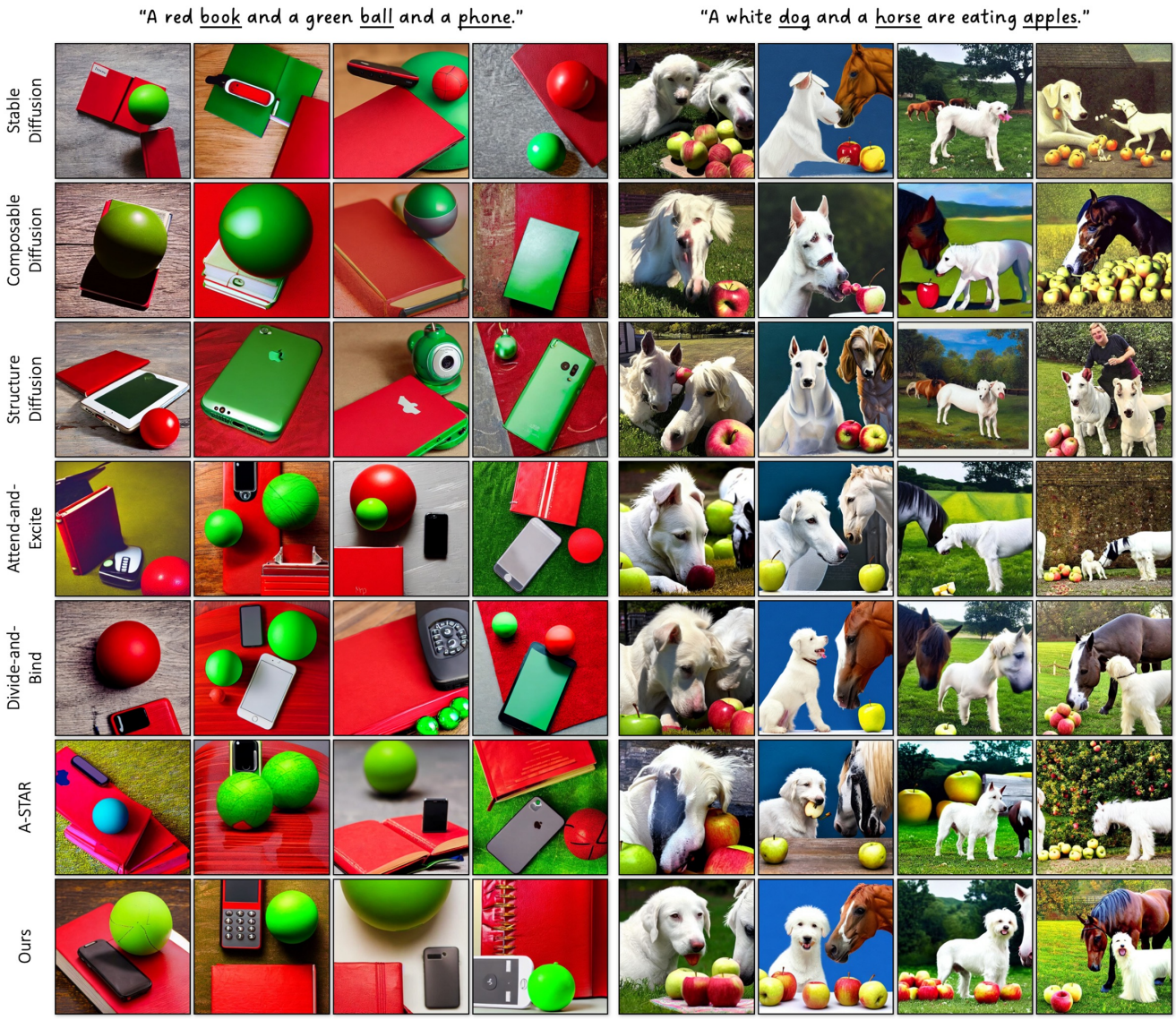


Figure 5. **Qualitative comparison with complex text prompts.** Each image is generated with the same text prompt and random seed for all methods. The subject tokens are highlighted in underline. From top to bottom: Stable Diffusion [8], Composable Diffusion [5], Structure Diffusion [3], Attend-and-Excite [2], Divide-and-Bind [4], A-STAR [1], and Ours.

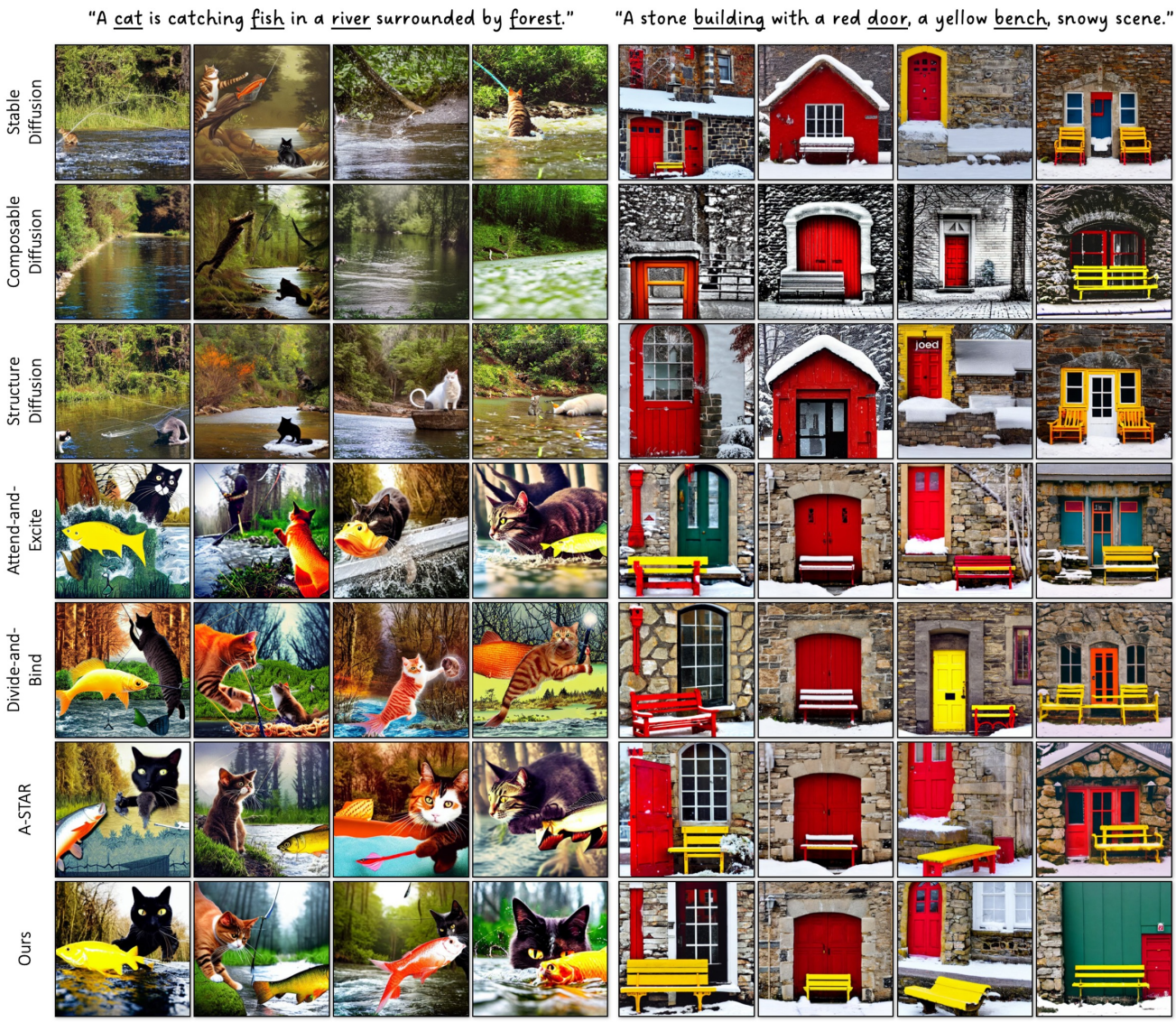
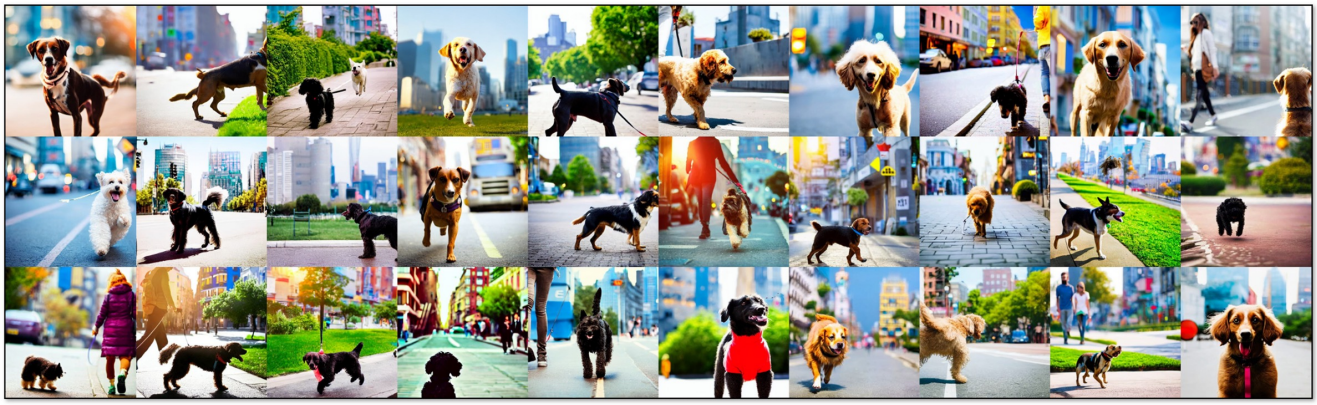


Figure 6. **Qualitative comparison with complex text prompts.** Each image is generated with the same text prompt and random seed for all methods. The subject tokens are highlighted in underline. From top to bottom: Stable Diffusion [8], Composable Diffusion [5], Structure Diffusion [3], Attend-and-Excite [2], Divide-and-Bind [4], A-STAR [1], and Ours.

"A campfire and a tree in the desert, starry sky."

"A dog walking on the city street."



"A cup of coffee and an orange."

Figure 7. **Example results synthesized by InitNO (ours).**

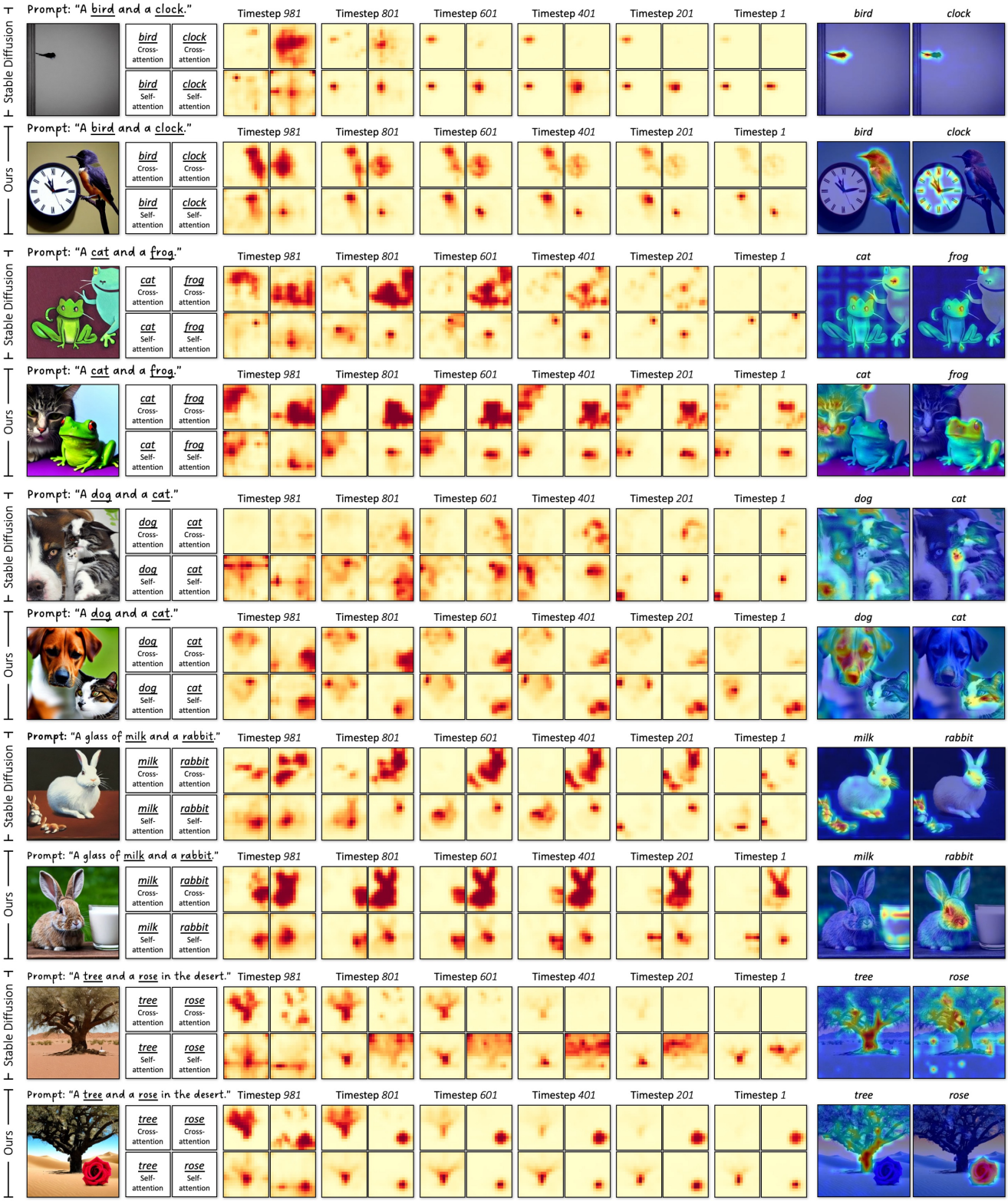


Figure 8. **Visualization of the attention maps.**

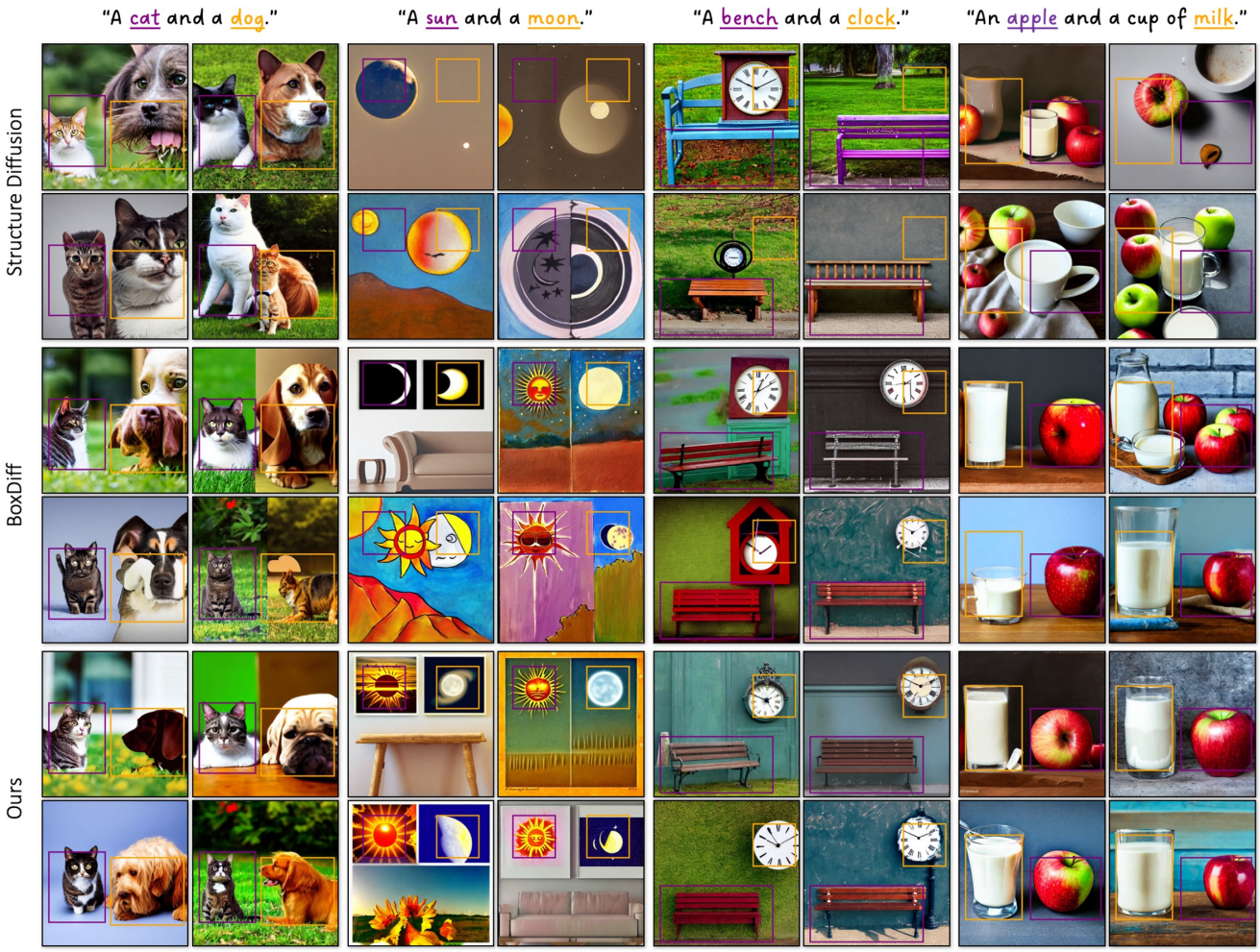


Figure 9. **Grounded Text-to-Image.**