Zero-Painter: Training-Free Layout Control for Text-to-Image Synthesis

Supplementary Material

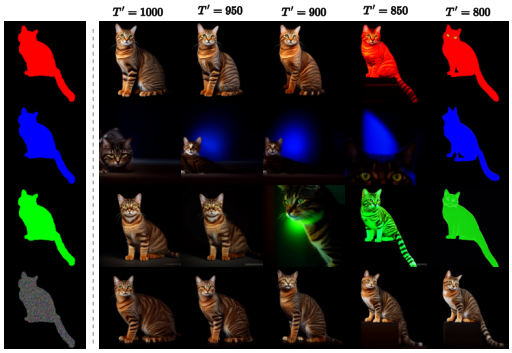

Figure 12. The effect using a starting latent with a specific color in Eq. (8) compared to initializing it with a random ϵ. For each color the images show the corresponding final output when the generation starts at timestep T'

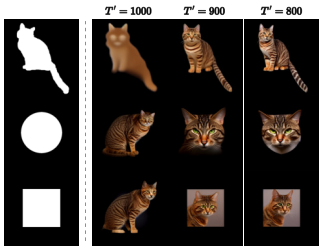

Figure 13. The effect of specifying the starting shape in Eq. (8). The images show the corresponding final output when the starting latent is initialized with a random ϵ as per Eq. (8) and the generation starts at timestep T'. Apart from the starting latent noise and the starting timestep the DDIM process and the Unet have not been modified (no PACA, ReGCA, etc).

Figure 14. Illustration of the generation process of a "cat" image with DDIM. Early timesteps focus on generating rough silhouettes, while later steps generate details. Images show the predicted x_0 at the corresponding timestep t

A. Additional Discussion About the Single Object Generation Stage

As mentioned earlier, [2] demonstrated that early generation stages focus on rough silhouettes and shapes, while later stages concentrate on details. This is illustrated in Fig. 14. In Fig. 14, during the DDIM process, we decode the predicted x_0 at intermediate timesteps t and present the results. Our hypothesis is that skipping early timesteps, while providing some information about the shape, will cause the model to stick to that shape regardless of the prompt. We demonstrate this claim in Fig. 13. We initialize the starting latent using the same algorithm described in the paper, employing a latent black color for the background and a Gaussian for the masked region, but we omit the intermediate blending step from the generation process. Additionally, we eliminate the PACA layer modification, meaning that, apart from the starting noise initialization, there is nothing restricting the model from generating non-zero pixels outside of the masked area. As can be seen, while the model exceeds the masked are for high values of T', for smaller values it sticks to the input shape quite closely.

Finally, in Sec. 3.3, we mention that providing specific color information to the generation will influence the final output of the model, which is undesirable. This effect is illustrated in Fig. 12. Here, we initialize the background using the latent black color, and for the masked area, we use the same Eq. (7) but with a mask of the specified color instead of black. As can be seen, while for high T' the influence on the output color is minimal, for smaller T' the influence grows. For $T' = 800$, which is particularly convenient for shape-awareness, the color influence is particularly high. In contrast, with the random initialization introduced in our paper, the output image exhibits much more organic colors, while the quality loss of the final image is minimal.

B. User Study

In addition to our numerical evaluations we conducted a user study to obtain a more-comprehensive and user-oriented evaluation of our results. The study involved 10 participants, that were tasked with evaluating our chosen competitor methods: eDiff-I using Stable Diffusion 1.4[2, 4], Multidiffusion[3, 41], Gligen [5, 17] and NTLB[12, 22], as well as our own method. We presented the participants with 20 samples consisting of the input layout, the global prompt and the set of per-object local prompts, as well as 5 generated images - one per method, which were shuffled in a random order. Subsequently, the participants were asked the following questions:

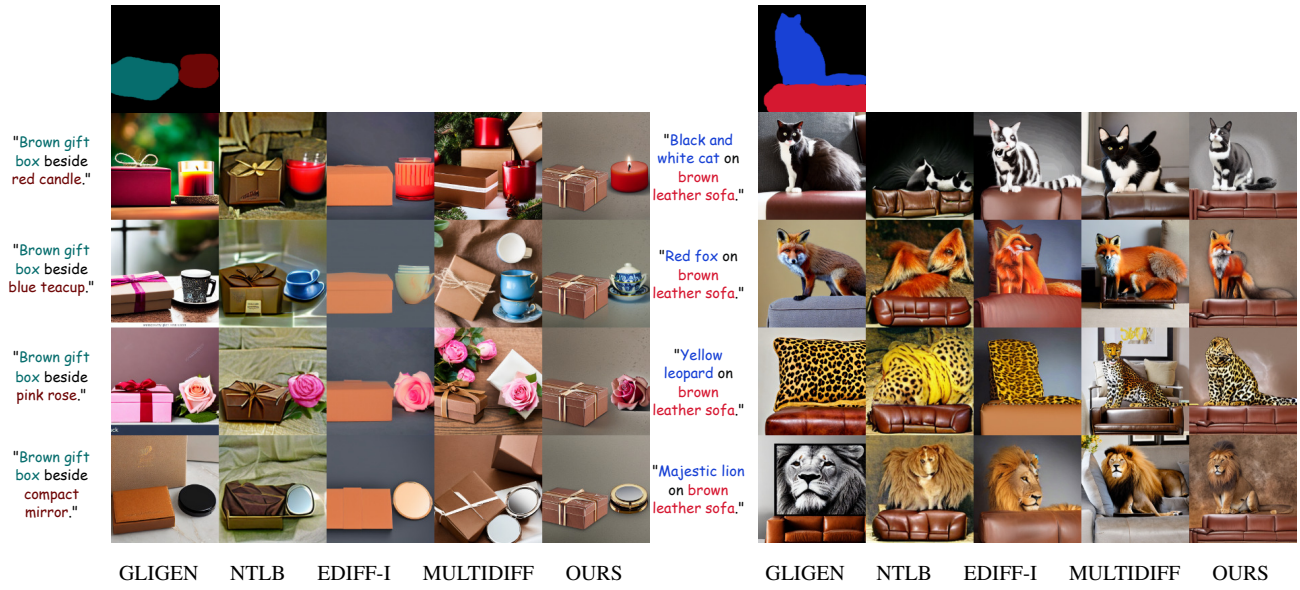

Figure 15. Independent Object Modification: Thanks to the Single-Object Generation stage, which assigns separate object seeds for the objects in the layout, our method enables the modification of some objects while keeping the rest the same.

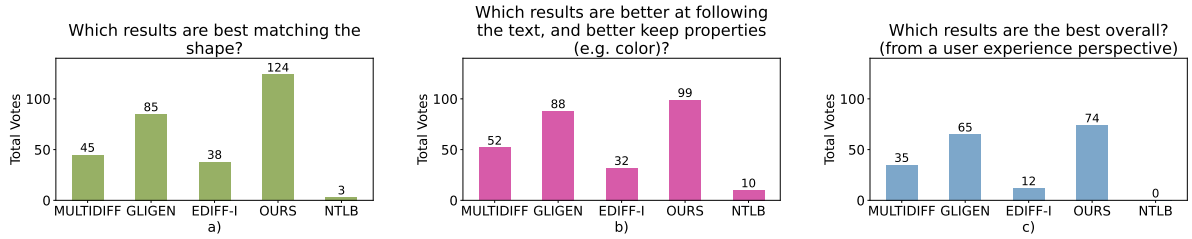

Figure 16. Total votes of each method based on our user study for questions a) *Which results are better at following the text, and better keep properties (e.g. color)?* b) *Which results are best matching the shape?* and c) *Which results are the best overall? (from a user experience perspective)* The user study shows a clear advantage of our method.

(a) *Which results are best at matching the shape?*
(b) *Which results are better at following the text, and better keep properties (e.g. color)?*
(c) *Which results are the best overall? (from a user experience perspective)*

The participants were allowed to choose none or multiple best methods since some approaches can perform equally good or equally bad. After collecting the feedback we calculate the sum of the votes for each method across all 20 samples and 10 participants (200 points max). The results are presented in Fig. 16 demonstrating a clear advantage of our method in both aspects: shape alignment and characteristic preservation of the generated objects.

C. Quantitative Ablation Study

In the main paper, we assessed the significance of two presented modules, PACA and ReGCA, using visual examples. Please see Table 2 for a qualitative comparison, showcasing the importance of PACA and ReGCA.

Model	w/o PACA, w/o ReGCA	PACA+ inpaiting	Ours
CLIP (local)	26.3	26.5	**26.63**
IoU (local)	0.73	0.74	**0.75**

Table 2. Quantitative ablation study.

D. Inference speed

Zero-Painter's inference time is directly influenced by the number of objects generated individually during the SOG stage. This relationship is illustrated in Fig. 17, where we provide a runtime analysis and compare it with other methods. Notably, employing a batch approach for object generation results in a lower computational overhead. Moreover, as existing objects undergo refinement in the later stages of the Comprehensive Composition phase, it becomes feasible to reduce the number of diffusion steps during SOG. This reduction contributes to a decrease in the overall generation

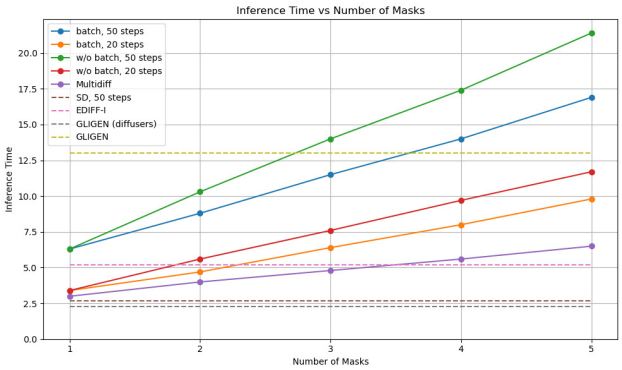

Figure 17. Inference time on A100 GPU for 512x512 resolution

time without compromising the final output quality significantly.

E. Additional Comparisons

Since most of the compared methods, including ours, are stochastic, we conducted additional experiments to demonstrate that the improvements we presented are not accidental but rather consistent. To do this, we sample five random seeds for each image in our visual test set in advance. We generate both our and competitors' images using the same set of seeds, and show the results in Fig. 18. As observed, the results of our method are consistently better aligned with both the input prompt and the shape when compared to all the competitors'.

F. Independent Object Modification

During the Single-Object Generation stage, different objects in the image are generated separately and can use their own independent generation seed. Additionally, the Comprehensive Composition stage does not modify the pre-existing objects too much. This adds increased flexibility to our model. Unlike the competitors, where changing a single word in the prompt can result in dramatically different output images, our model allows changing one or more of the input objects while keeping all the others the same Fig. 15.

Figure 18. A comprehensive comparison of our method to the competitors. For each image we randomly sampled 5 seeds and used all the results.

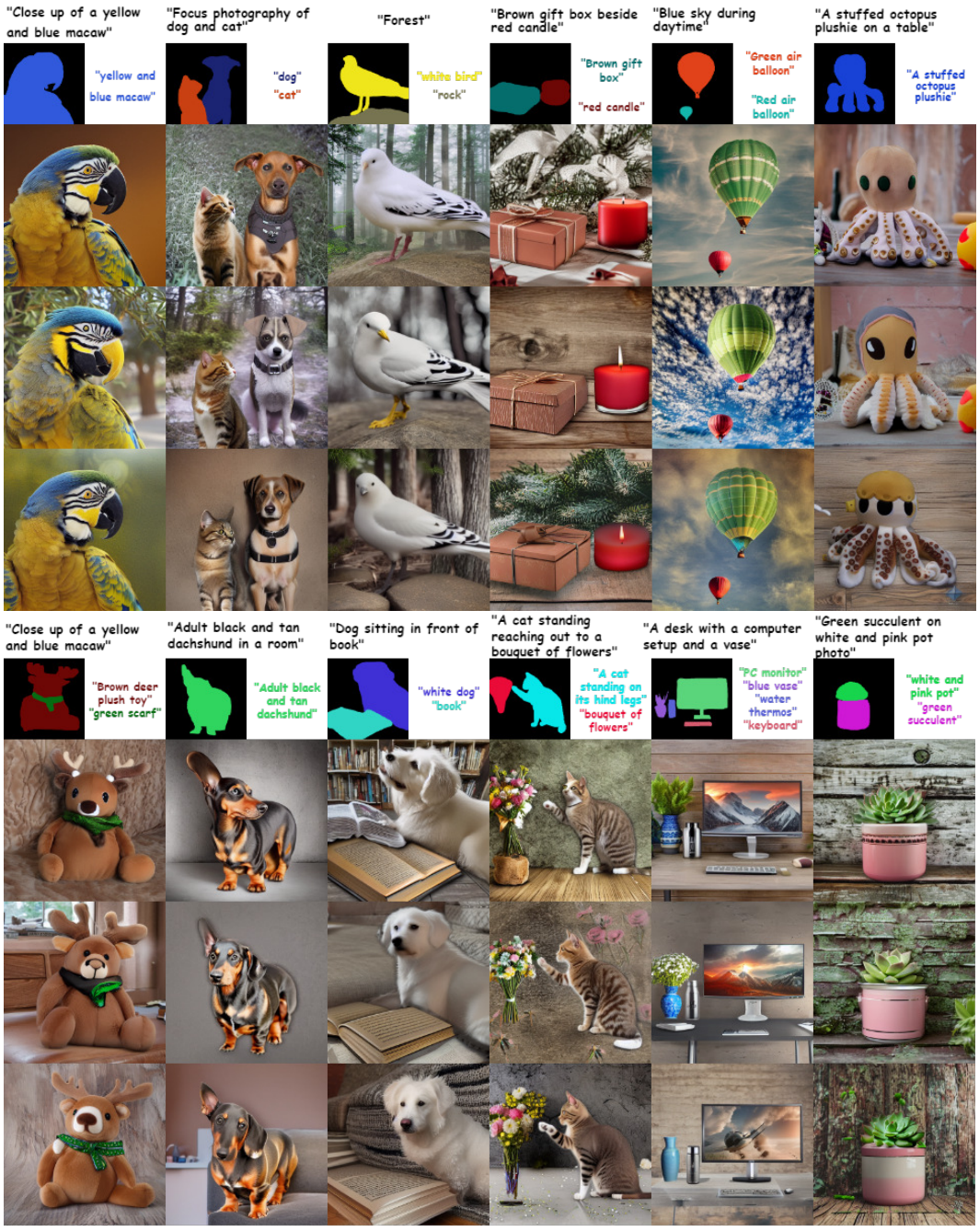

Figure 19. Additional examples of images generated by our model.