AnyScene: Customized Image Synthesis with Composited Foreground

Supplementary Material

A. Evaluation Benchmark

In constructing a fair benchmark for quantitatively evaluating our proposed task, we utilized the publicly available COCO [3] dataset to construct composite foregrounds. Specifically, we employed the segmentation annotations of 80 object categories in COCO to extract foreground elements and recombined them onto a canvas to generate "composite foregrounds" as task inputs. Subsequently, we synthesized complete scenes guided by the overall descriptions in the dataset and compared different methods' generative abilities by assessing the quality of the synthesized scenes. We employed three metrics to evaluate the model's generative capabilities from the perspectives of image quality, text-image consistency, and aesthetic appeal. The definitions and calculation formulations are as follows:

- FID(Fréchet inception distance) [1]. It assess the quality of generated images by measuring the distribution difference between real and generated images, which is formulated as:

$$\mathrm{FID}(x, g) = \|\mu_x - \mu_g\| + \mathrm{Tr}\left(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x \Sigma_g}\right),$$

  where x represents the overall feature distribution of real images extracted by the inception network [9], and g represents that of the generated imgaes. We calculated the FID values of the generated images using real images from the COCO as the reference ferature distribution.
- CLIP-Score [4]. This metric evaluates the model's ability to generate images matching text descriptions based on the similarity between CLIP embeddings of generated images and input text.
- LAION Aesthetic [8]. Proposed by LAION, it is a set of lightweight models predicting ratings given by people when asked "How much do you like this image on a scale from 1 to 10?". We calculate the aesthetic quality of generated images using LAION's provided predictor model [7].

B. User Studies

To compare the generative performance of AnyScene and alternative methods on the proposed task from the perspective of subjective quality assessment by users, we organized a series of user studies. For customized scene generation performance on the proposed task, we chose Inpaint-Anything [5, 10] as the primary baseline method for comparison. As for alternative methods for single-object scene customization, we choose DreamBooth [6] and CustomDiffusion [2]. Based on several objects from the DreamBooth dataset and some collected from the internet, we constructed 30 groups of foreground elements and compose them into the composited foreground. To construct the customized text prompt, we manually wrote ten different scene context descriptions, such as: "with sunset", "on the street, background is Eiffel Tower", "surrounding with flowers", etc. We generated images using these prompts combined with the composited foregrounds, producing four images per prompt for each method. The generated images, anonymized and randomly grouped into 30 sets, were rated by ten annotators from diverse age groups and professions based on the following criteria:

- Image Quality: Scoring the overall quality of the generated images based on personal aesthetic preferences.
- Scene Harmony: Whether the foreground elements correctly and reasonably blend into the generated scene, presenting visual harmony.
- Object Fidelity: Whether the foreground elements maintain good recognizability in the generated images, without noticeable layout deformation or detail loss.

Based on these criteria, the annotators scored the images on a scale of 1 to 4 based on these criteria. The final evaluation results are shown in the Table. 1. It is evident that compared to Inpaint Anything [5, 10], our method showed better generative quality in all three metrics according to user preferences. As for the subject-driven methods [2, 6], although these tend to generate overall harmonious images by learning specific visual subjects as visual concepts, they did not show superior results in scene harmony compared to AnyScene in user evaluations. This indicates that AnyScene effectively considers the semantic information of the foreground during generation, synthesizing the overall scene based on the foreground elements, thereby achieving harmonious scene generation. In summary, the results of the user studies affirm the effectiveness of AnyScene in the proposed task and its applicability in generation tasks like poster creation and scene customization that require balancing image quality, scene harmony, and object fidelity.

Table 1. Comparative user study results for customized image generation performance across different methods.

	Image Quality	Scene Harmony	Object Fidelity
DreamBooth [6]	2.23	2.59	1.61
CustomDiffusion [2]	2.31	2.48	1.39
Inpaint Anything [10]	2.80	2.41	3.11
Ours	**3.26**	**2.92**	**3.45**

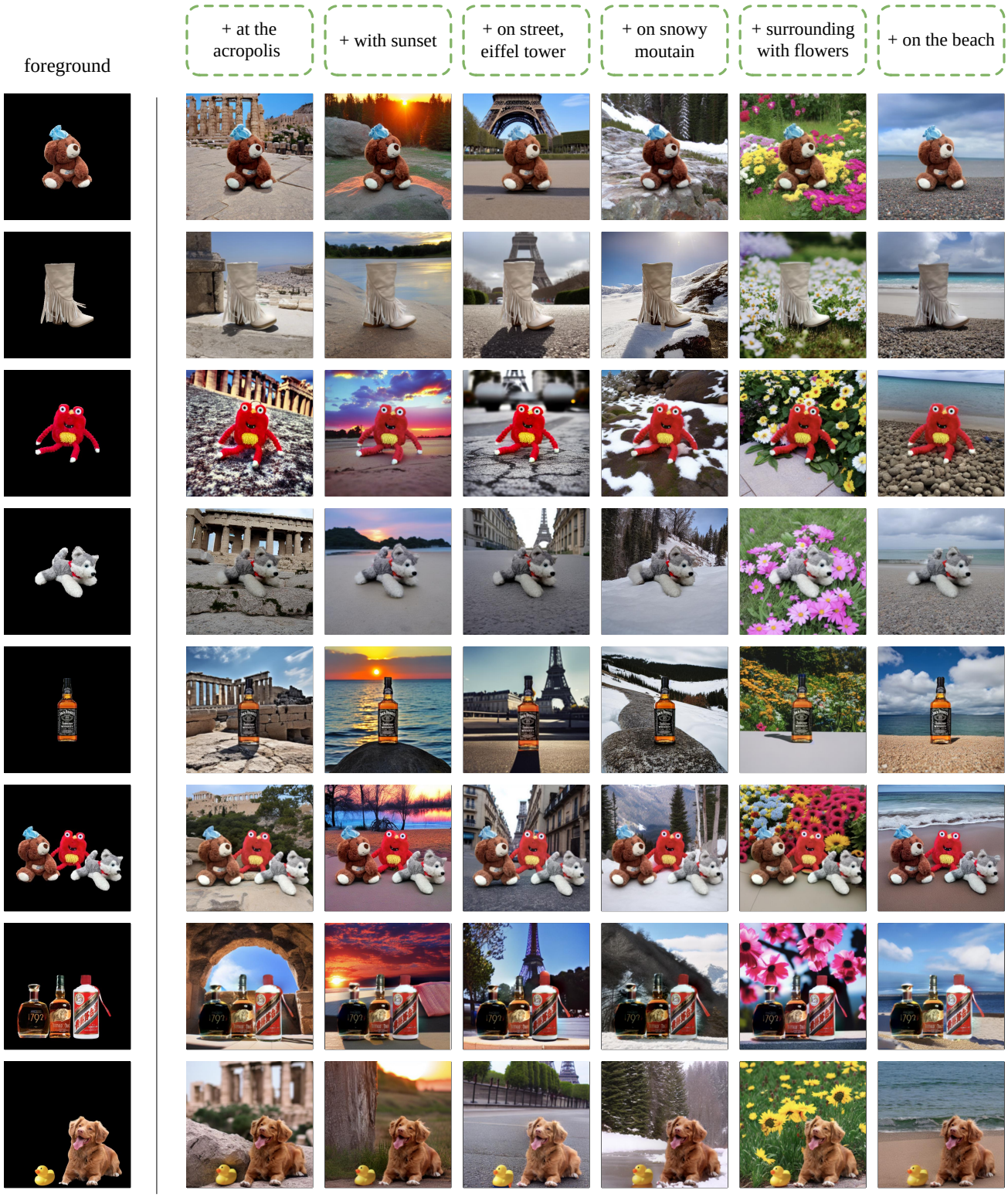

Figure 1. Customized scene generation results synthesized via proposed AnyScene conditioned with specific foregrounds and prompts of various scene contexts.

C. Cutomized Image Generation Results

In order to vividly demonstrate the generative capabilities of AnyScene, Fig. 1 showcases more results produced by AnyScene. As illustrated, AnyScene can generate high-quality customized images based on a given foreground, combined with various scene prompts.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 1

[2] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[5] 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

[6] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 1

[7] Christoph Schuhmann. LAION-Aesthetics Predictor V1. https://github.com/LAION-AI/aesthetic-predictor, 2022. 1

[8] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[10] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 1