

Supplementary Materials for DreamBlend: Advancing Personalized Fine-tuning of Text-to-Image Diffusion Models

Shwetha Ram, Tal Neiman, Qianli Feng, Andrew Stuart, Son Tran, Trishul Chilimbi
Amazon

{shweram, taneiman, fengq, andrxstu, sontran, trishulc}@amazon.com

Abstract

In Sec. 1, we present more qualitative results in addition to Fig. 4 and Fig. 5 and Fig. 8 of the main text. In Sec. 2, we visualize how DreamBlend advances the pareto front for more example subjects from DreamBooth benchmark, in addition to Fig. 7 of the main text. In Sec. 3, we explain details of the human preference studies conducted, present examples of the user interface used and validate statistical significance. In Sec. 4, we present the effects of varying the cross attention guidance and classifier-free guidance. In Sec. 5, we present some implementation details. In Sec. 6, we present comparisons to non-fine-tuning based text-to-image personalization methods.

1. More qualitative results

In Fig. 12, we present more results, in addition to Fig. 4 of the main text. In Fig. 13, we present more results in addition to Fig. 5 of the main text. In Fig. 20, we present more results with SDXL backbone, in addition to Fig. 8 of the main paper.

2. DreamBlend advances the pareto front

In Fig. 14, we visualize how DreamBlend advances the pareto front for more example subjects from the DreamBooth benchmark, in addition to Fig. 7 of the main text.

3. Human preference study

Two user studies were performed, assessing overall preference and diversity, comparing our approach to DreamBooth and Custom Diffusion. An example interface used for these studies is shown in Fig. 15. In the overall preference study shown in Fig. 15a, users chose between an image generated by our method and a baseline method for the same text prompt, considering both subject and prompt fidelity. In the diversity study shown in Fig. 15b, users selected the more diverse collection of four images between

Preference study	Chi-square statistic	P-value
Ours over DB Overall	24.40	7.82e-07
Ours over DB Diversity	17.62	2.70e-05
Ours over CD Overall	42.86	5.88e-11
Ours over CD Diversity	78.27	8.97e-19

Table 3. Results of Chi-square goodness of fit tests on human preference study results. The P-value is very low in all studies.

our method and a baseline. They were asked to consider both subject fidelity and prompt fidelity and select the collection of images which is more diverse in terms of backgrounds, subject poses, etc. For example, in Fig. 15b, the images in the left collection have very similar backgrounds while the images in the right collection are more diverse. The studies comprised of 1000 questions, each question was answered by an average of six people and the order was randomized.

Statistical tests were performed to verify the statistical significance of each study and all results were found to be statistically significant. The results of one-sample binomial test with confidence intervals are summarized in Tab. 4. The results of Chi-square goodness of fit test are summarized in Tab. 3.

4. Effect of varying cross attention guidance and classifier-free guidance

In Fig. 16, we present more examples of the effect of varying cross attention guidance scale, in addition to Fig. 9 of the main text. In Fig. 17, we present the effect of varying both cross attention guidance scale and classifier-free guidance, for the same guidance and edit models.

5. Implementation details

For experiments in Sec. 5 of the main text, we use the pre-trained Stable Diffusion v1.5 model [33] and the SDXL model [?]. We use the HuggingFace Diffusers [41] implementation and the hyperparameters recommended by the

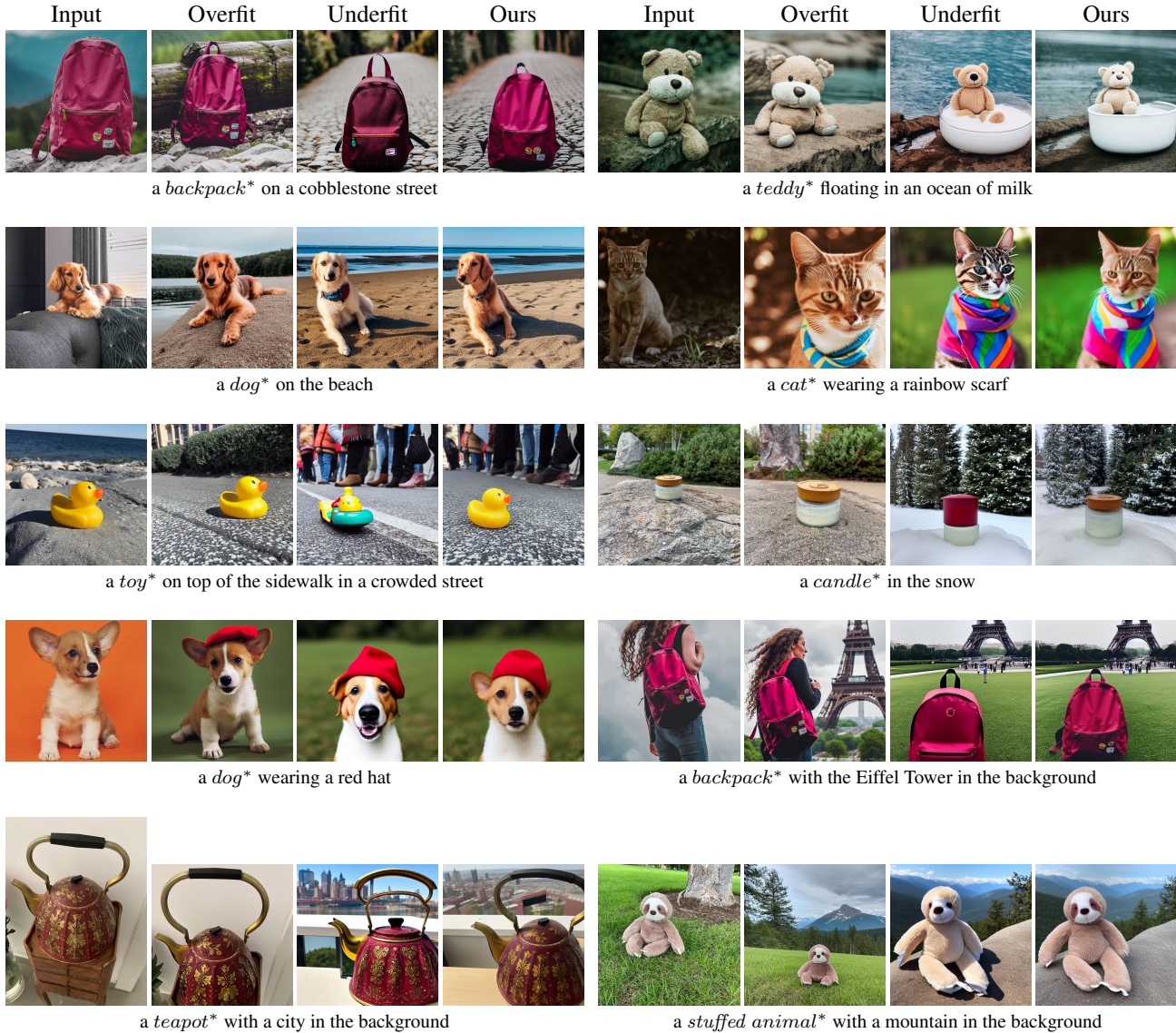


Figure 12. Guided Image Synthesis: Across various subjects and prompts, our approach successfully preserves the layout of the reference underfit image as well as the identity of the input subject. Images generated by the Overfit (Edit) and Underfit (Guidance) models used in our approach are shown for reference.

Preference study	Study result	Binomial p value	CI (ours)	CI (baseline)
Ours over DB Overall	61.11	2.51e-12	[58.08, 64.13]	[35.87, 41.92]
Ours over DB Diversity	61.82	2.51e-09	[58.05, 65.58]	[34.42, 41.95]
Ours over CD Overall	70.16	2.44e-20	[66.21, 74.10]	[25.90, 33.79]
Ours over CD Diversity	72.70	1.46e-35	[69.47, 75.94]	[24.06, 30.53]

Table 4. Results of one-sample binomial tests on human preference study results. CI denotes the 95% Adjusted Wald Confidence Intervals. The lower bound of the CI for our approach is greater than 50% in all studies. Also, the confidence intervals for our approach and those for the baseline approach are well separated in all studies. Further, the exact binomial p value is very low in all studies.

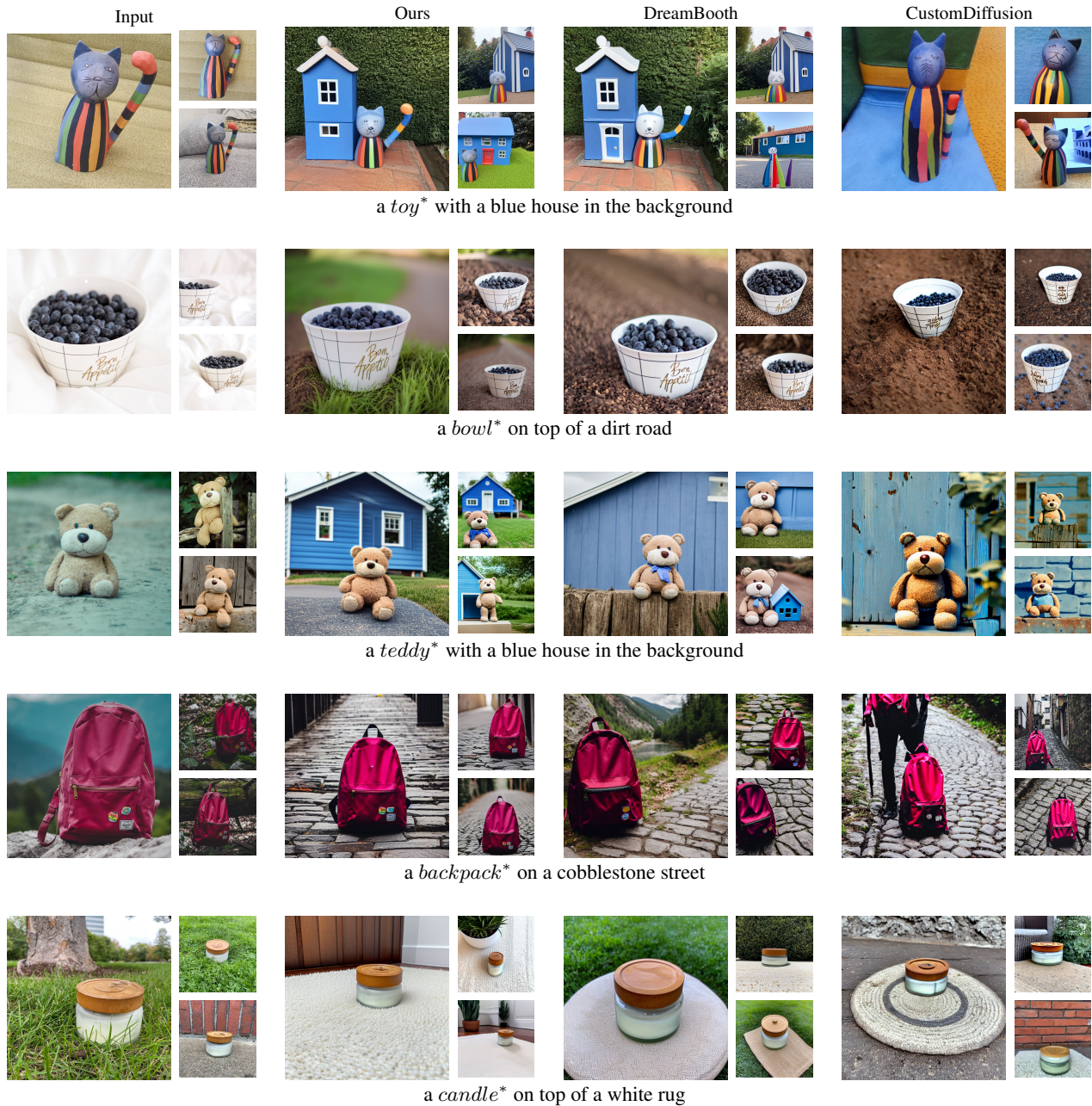


Figure 13. Comparison with prior works: Our approach successfully generates images with better subject fidelity, prompt fidelity and diversity on challenging prompts.

authors. For DreamBooth, we use a learning rate of $5e^{-6}$ and the rare token “sks” to represent the specific subject during fine-tuning. For Custom Diffusion, we use a learning rate of $1e^{-5}$, scaled with effective batch size. For regularization, we use 1000 images of the subject’s category generated by the pre-trained model, with a prior preservation weight of 1.0. We use 50 steps of DDIM forward process for all methods.

We apply our approach, DreamBlend, on results of classical DreamBooth tuning, full fine-tuning for SDv1.5 and LoRA for SDXL. For all subjects, we designate the models at step 100 and step 200 as edit models and all models with lower steps as guidance models. For the step 100 edit model, we use a classifier-free guidance scale of 3.0 and a cross attention guidance scale of 0.1 while for the step 200 edit model, we use 2.0 and 0.07, respectively. As the step

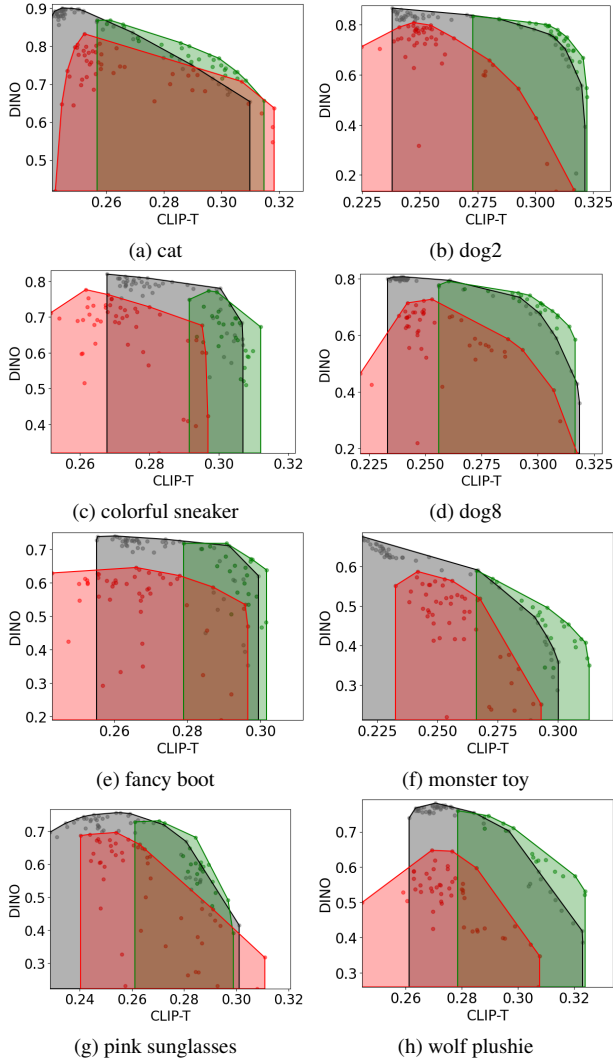


Figure 14. Image alignment (DINO) - text alignment (CLIP-T) space spanned by densely sampled operating points of Dream-Booth (gray), Custom Diffusion (red) and our method (green) for example subjects. Our method advances the pareto front, offering operating points unavailable to existing methods.

200 model has learnt the subject better, it can achieve higher subject fidelity with lower classifier-free guidance.

For calculating metrics, we use the CLIP [30] ViT-B/32 model for CLIP-I and CLIP-T and DINO [2] ViT-S/16 model for DINO metric. Prior to computing text embeddings, we remove any rare token, such as “sks” from the prompt.

6. Comparison with non-fine-tuning based approaches

In this section, we present qualitative comparisons to non-fine-tuning based methods, in addition to Fig. 6 of the

main paper. Comparisons to Textual Inversion and BLIP-Diffusion are in Fig. 18. Comparisons to IP-Adapter and AnyDoor are in Fig. 19. For Textual Inversion, we trained the word embedding for the recommended 3000 steps, logging results every 500 steps and present the best results. For AnyDoor, we generated the background images using the pre-trained StableDiffusion model and used CLIPSeg [23] to generate the segmentation masks.

Guideline: Select the best image after looking at the example images and text prompt carefully. The best image should 1) Depict the subject shown in the example images below 2) Follow the text prompt. Consider both factors and select the image you prefer overall. You have the option to choose "both look good" or "both look bad"

Below are example images of "[V] stuffed animal"



Prompt: a [V] stuffed animal with a mountain in the background

Based on the guideline, make your selection



Select Select

👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍 Both look good 👍


👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎 Both look bad

5/250 images labeled

(a) Overall preference study


Guideline: Select the best set of images after looking at the example images and text prompt carefully. The two sets of images in the questionnaire are generated for the same subject and text prompt. The images should 1) Depict the subject shown in the example images below 2) Follow the text prompt. Select the set of images that is more diverse while adhering to the two factors mentioned above. Here diversity refers to different backgrounds, different poses of the subject, etc. while still being faithful to the subject shown in the example images and the text prompt. You have the option to choose "both look good" or "both look bad"

Below are example images of "[V] stuffed animal"



Prompt: a [V] stuffed animal with a city in the background

Based on the guideline, make your selection



Select Select

👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍👍 Both look good 👍👍

👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎👎 Both look bad 👎👎

(b) Diversity study

Figure 15. Human preference study interface

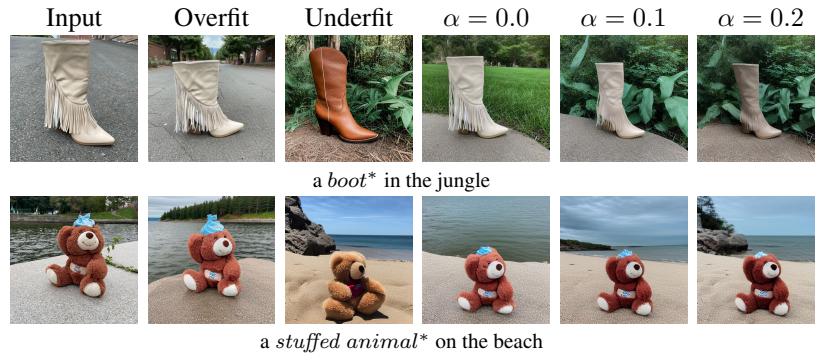
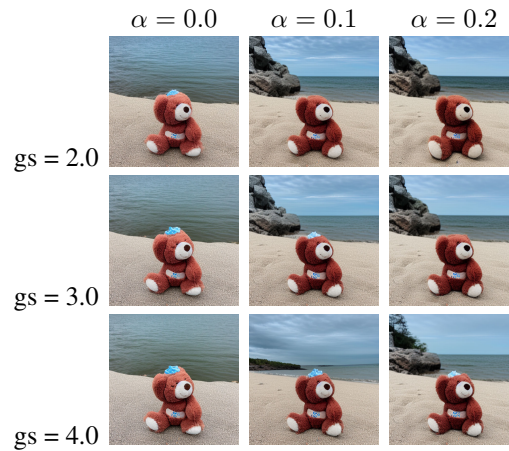


Figure 16. Effect of cross attention guidance scale α , for the same guidance and edit models and classifier-free guidance



(a) Input training image, overfit (edit) image and underfit (guidance) image



(b) Our results varying classifier-free guidance scale (gs) and cross attention guidance scale (α)

Figure 17. Effect of varying classifier-free guidance scale (gs) and cross attention guidance scale (α) for the same guidance and edit models and prompt “a *stuffed animal** on the beach”. Increasing classifier-free guidance improves subject fidelity, while increasing cross attention guidance increases adherence to the layout of the underfit (guidance) image.

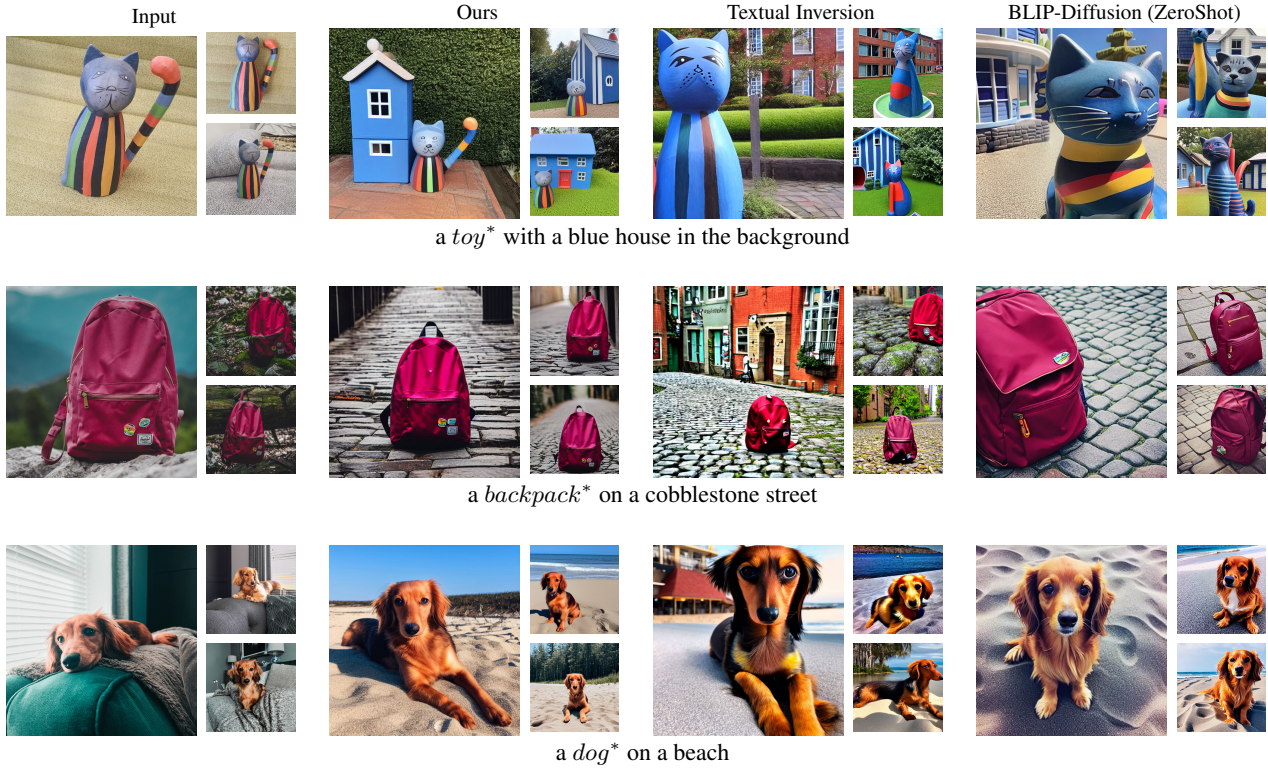


Figure 18. Comparison with non-fine-tuning based methods Textual Inversion and BLIP-Diffusion

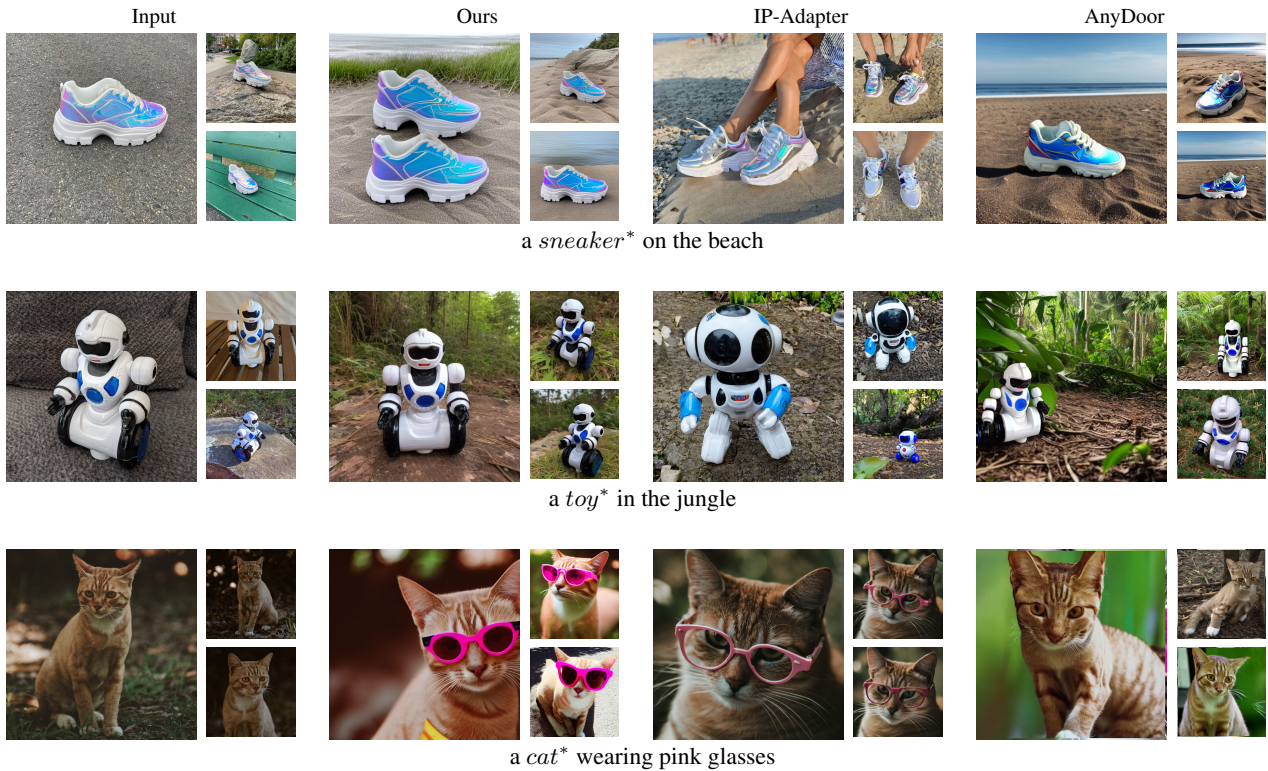


Figure 19. Comparison with non-fine-tuning based methods IP-Adapter and AnyDoor



Figure 20. Qualitative results on SDXL backbone

References

- [1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4
- [2] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 4
- [3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 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. 4
- [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] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 1