Customize your NeRF: Adaptive Source Driven 3D Scene Editing via Local-Global Iterative Training

Supplementary Material

A. Model Architecture

Our foreground-aware NeRF is constructed on Instant-NGP [5] with an additional head to predict the editing probability to conduct decomposed rendering. For simplicity in implementation, this editing probability head is designed similarly to the RGB prediction head. It is composed of a hidden layer with 64 hidden dimensions followed by the sigmoid activation to keep the values between 0 and 1. When rendering the foreground or background regions separately, we avoid truncating the editing probability gradient with the threshold operation. Instead, we encourage the editing probability to approach 1 or 0 through a parameterized sigmoid activation and utilize the processed probabilities for decomposed rendering via a simple linear combination. We find that this soft mask operation better facilitates complete shape optimization during the editing process compared to a threshold operation, as demonstrated by the ablation study presented in Section C.

B. Implementation Details

Dataset preparation. In this paper, we conduct different editing operations using datasets such as BlendedMVS [10] and LLFF [4]. For 360-degree scenes from these datasets, we employ COLMAP [7] to extract camera poses. This method can also be applied to real-world scenes captured by users.

Rendering. Due to the memory limitation of the GPU and the high overhead of SDS loss, we need to render down-sampled images. For the BlendedMVS dataset, we render images with 3x downsampling for NeRF training and 5x downsampling for NeRF editing. For the high-resolution LLFF and IBRNet [9] datasets, the two downsampling factors are 15 and 28. For the bear statue dataset [2], the two downsampling factors are 4 and 7. Since we render the foreground for a separate local editing operation, we combine the rendered foreground image with a random solid color background to avoid confusion between the foreground and the single background of the same color.

C. More Ablation Studies

Visualization of intermediate results. We present the rendered editing probabilities and foreground/background images in Figure 1 after editing. It can be observed that the rendered editing probabilities can adapt well to the shape of the modified foreground regions, enabling the precise rendering of both foreground- and background-only images. For instance, after the training process of the edited NeRF, the editing probability aligns well with the cartoon dinosaur, which is larger than the original dinosaur statue. In the background image in the second row, we observe dark shadows in the foreground region, which indicates that NeRF does not picture the covered background regions in non-360 scenes. However, in the full rendering image, this shadow will be filled with foreground contents when the new objects being added are similar in shape to the original objects.

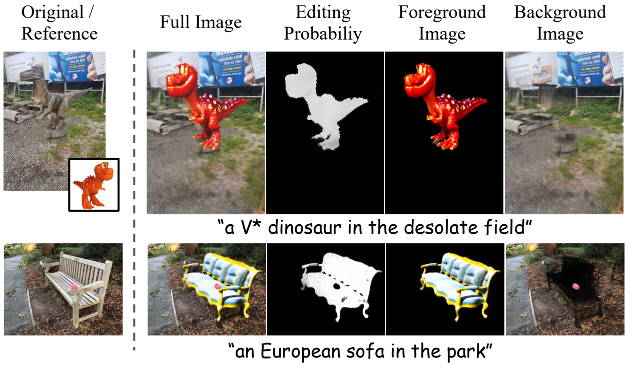


Figure 1. Visualization of intermediate results.

Ablation studies of other training strategies. We conduct more ablation studies on training strategies in Figure 2. Upon removal of the background preservation loss, significant changes can be observed in the background regions after editing compared with the original scene, indicating the necessity of the background preservation loss. We also ablate different post-processing methods for editing probabilities. When binarizing these probabilities using a threshold (i.e., w/o Soft Mask), the obtained editing probabilities are incomplete, leading to incomplete foreground in the edited results, such as the missing left front leg of the dog in the 3-rd row of Figure 2. By scaling the values between 0 and 1 using a sigmoid function to obtain a soft mask, *i.e.*, the strategy adopted in our paper, we can achieve complete editing probabilities and consequently, complete editing results.

Effect of different class word. We use the class word in the local editing stage to promote reasonable geometry in the case of a single reference image. In order to explore the effect of the class word, we modify the class prior to be inconsistent with the reference image. As shown in Figure 3, we compare different class words such as "dog", "corgi", and "cat". The final results reflect the texture of the reference image and the geometric characteristics associated with the class word we use.

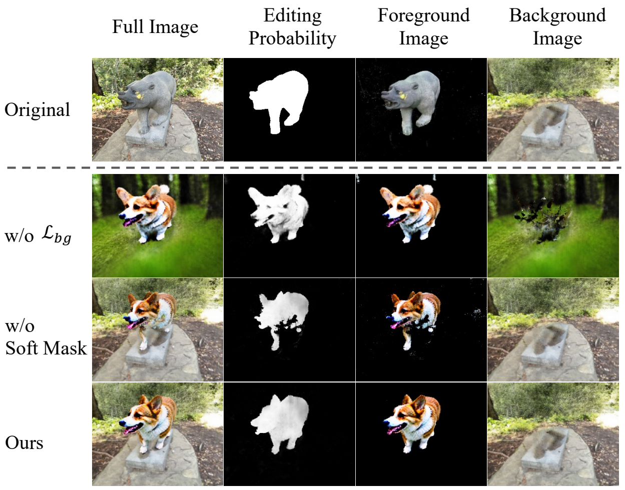
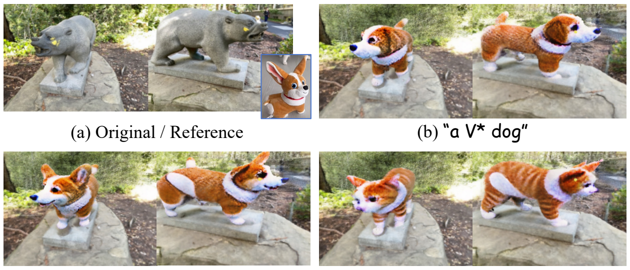


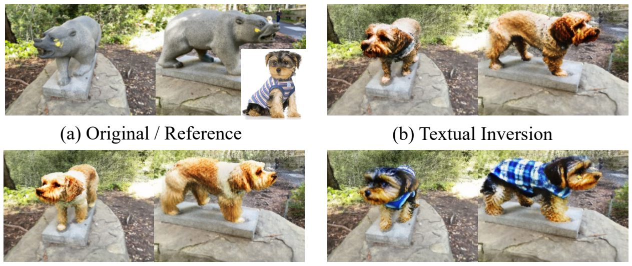
Figure 2. Ablation studies of other training strategies. w/o Soft Mask denotes utilizing a threshold to binarize the editing probabilities.



(a) Original / Reference (b) "a V* dog"

(c) "a V* corgi" (d) "a V* cat"

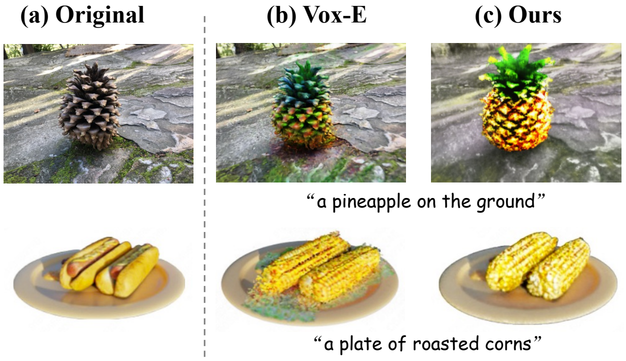
Figure 3. Visualization of different class words used in local editing prompt.



(a) Original / Reference (b) Textual Inversion

(c) DreamBooth (d) Custom Diffusion

Figure 4. Comparison between different subject-aware T2I generation methods for reference subject learning.



(a) Original (b) Vox-E (c) Ours

"a pineapple on the ground"

"a plate of roasted corns"

Figure 5. Qualitative comparison with Vox-E.

Different subject-aware T2I generation methods. We compare the image-driven editing results leveraging Custom Diffusion [3] for reference subject learning with two other subject-aware T2I methods, *i.e.*, Textual Inversion [1], and DreamBooth [6] and the visualization results are presented in Figure 4. Given only one reference image, editing results generated with Custom Diffusion most closely resemble the reference image. Though the other two methods can generate similar categories and shapes, they do not match the reference image in texture. Therefore, in our paper, we ultimately select Custom Diffusion for reference subject learning.

D. More Visualization Results

Comparison with Vox-E. We present qualitative comparison with Vox-E [8] in Figure 5. Our method generates more realistic editing results than Vox-E. In the 1st row, the editing results by VoX-E maintain the shape of the pinecone and have an unnatural color. Besides, some odd color artifacts can be observed in both rows of Vox-E's edits, due to Vox-E's attention-based post-blending strategy, and our editing results strike a balance between text prompt alignment and background preservation.

More qualitative results. We provide more image-driven and text-driven editing results in Figure 7 and Figure 8 respectively. We have also included video editing results at `https://customnerf.github.io/`, which provides dynamic visualization of editing results.

Failure cases. In addition to the experimental results above, we also present the failure cases of CustomNeRF in Figure 6. For instance, in the first row, as the Custom Diffusion cannot always generate images with the subject identical to those in the reference image, the edited scene shows the dog's clothing different from that in the reference image. In the second row, due to the considerable shape difference between a hat and a hamburger, the replaced hamburger takes on an inverted shape, larger at the bottom and smaller at the top, similar to the hat's structure.

References

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-

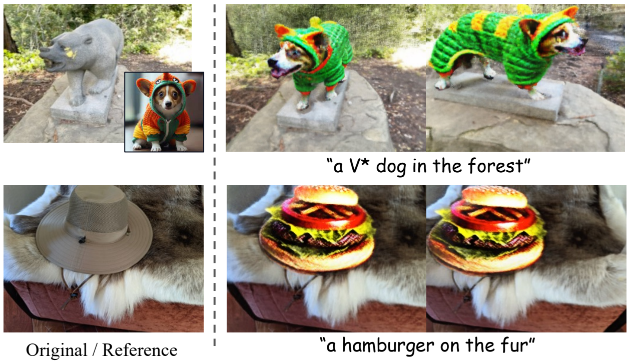


Figure 6. Failure cases.

image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[2] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. *arXiv preprint arXiv:2303.12789*, 2023. 1

[3] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2

[4] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *TOG*, 2019. 1

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 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 *CVPR*, 2023. 2

[7] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[8] Etai Sella, Gal Fiebelman, Peter Hedman, and Hadar Averbuch-Elor. Vox-e: Text-guided voxel editing of 3d objects. In *CVPR*, 2023. 2

[9] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 1

[10] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 1

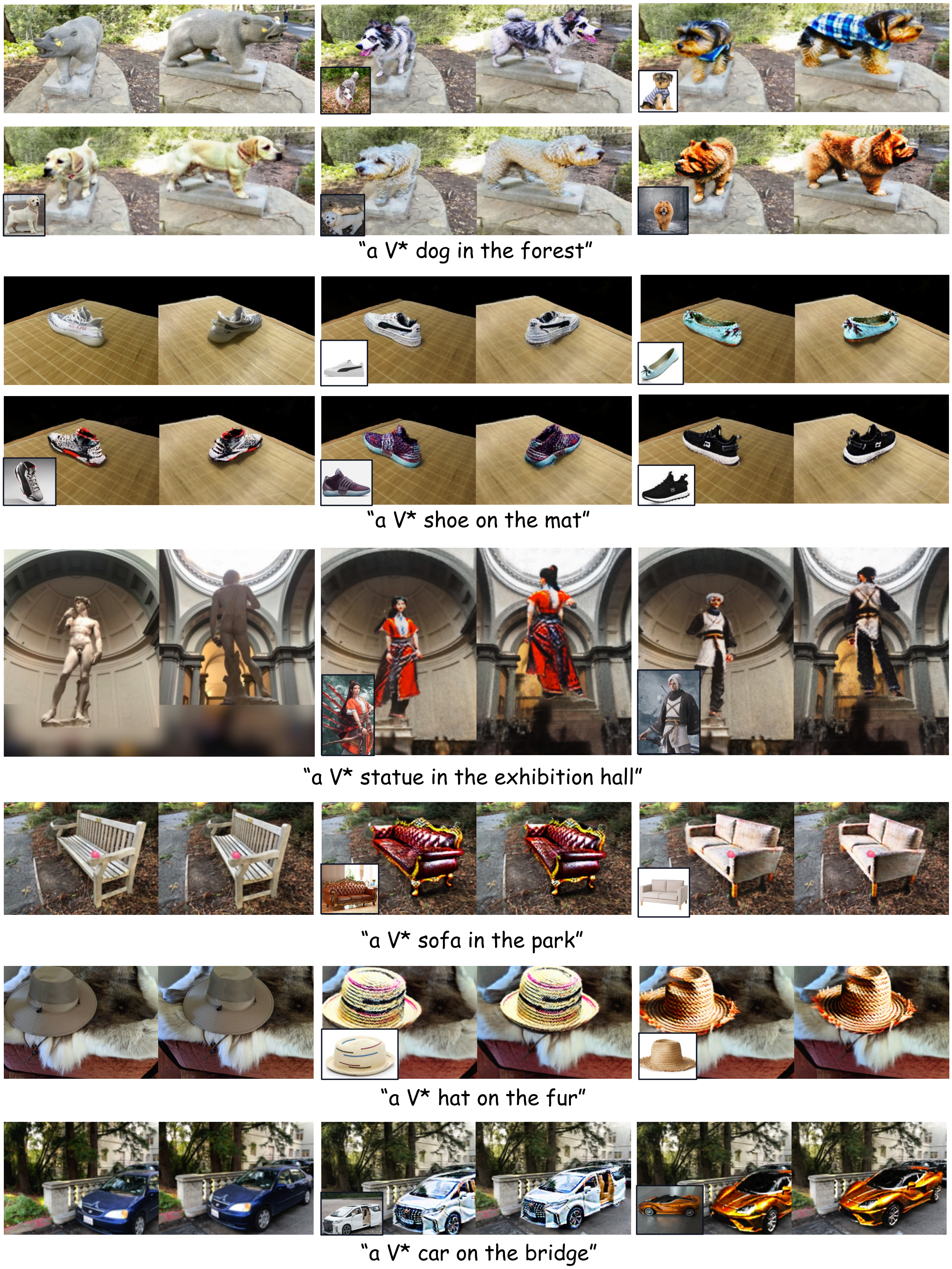


Figure 7. More visualization results of image-driven editing.

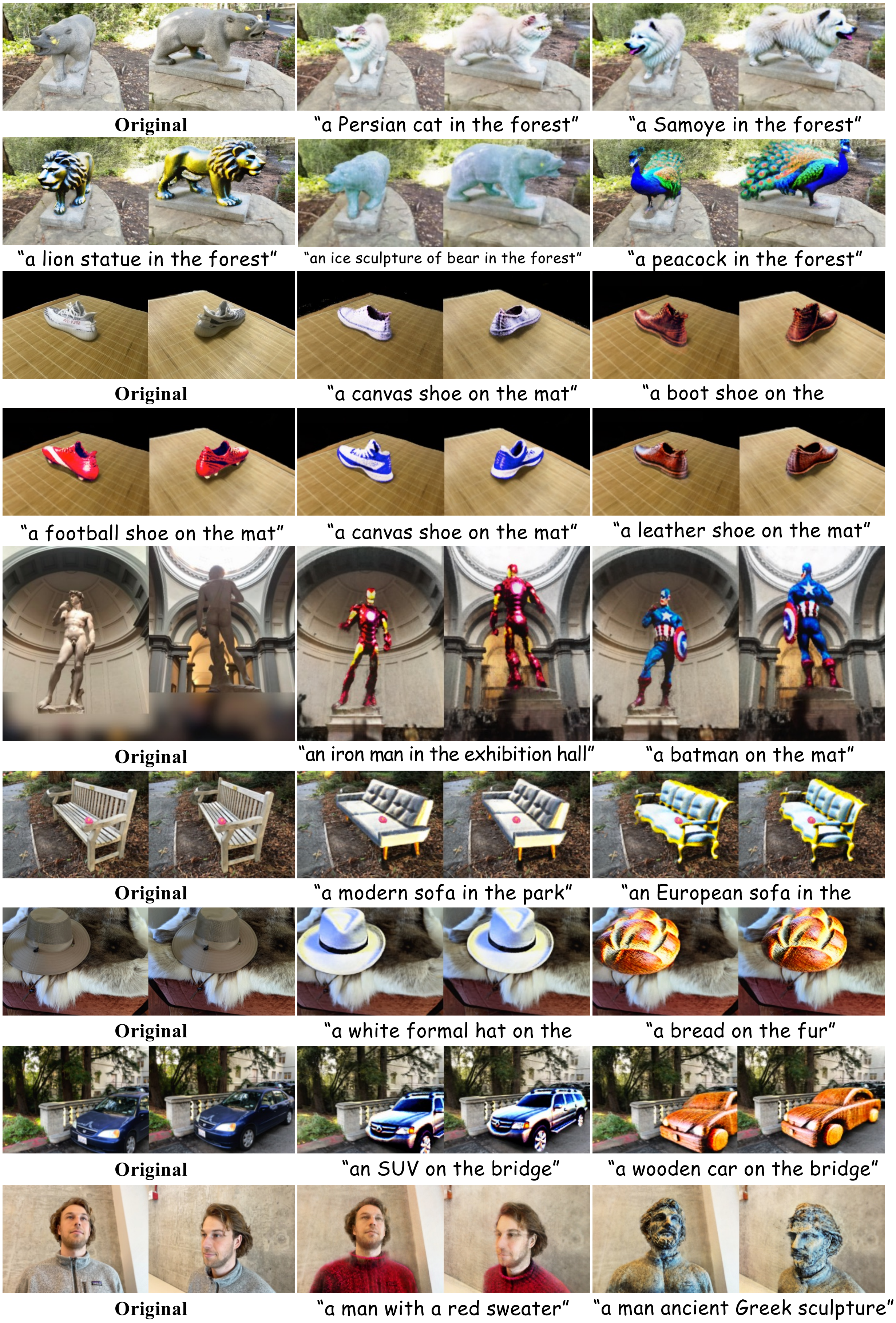

Original
“a Persian cat in the forest”
“a Samoye in the forest”
“a lion statue in the forest”
“an ice sculpture of bear in the forest”
“a peacock in the forest”
Original
“a canvas shoe on the mat”
“a boot shoe on the
“a football shoe on the mat”
“a canvas shoe on the mat”
“a leather shoe on the mat”
Original
“an iron man in the exhibition hall”
“a batman on the mat”
Original
“a modern sofa in the park”
“an European sofa in the
Original
“a white formal hat on the
“a bread on the fur”
Original
“an SUV on the bridge”
“a wooden car on the bridge”
Original
“a man with a red sweater”
“a man ancient Greek sculpture”

Figure 8. More visualization results of text-driven editing.