DiGA3D: Coarse-to-Fine Diffusional Propagation of Geometry and Appearance for Versatile 3D Inpainting -Supplementary Material-

Jingyi Pan[1] Dan Xu[2*] Qiong Luo[1,2*]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
jpan305@connect.hkust-gz.edu.cn, {danxu, luo}@cse.ust.hk

1. Additional Implementation Details

To select reference views, we utilize K-means clustering with $K = 3$ to identify the views that are closest to the cluster centers as our reference views. During the coarse stage, we set $\lambda = 0.6$ for our AFP mechanism to propagate reference attention features into other attention features effectively. In the fine stage, we set the guidance scale to 7.5, the condition scale for depth to 1.0, and the condition scale for texture to 0.8. Some parameters will be adjusted based on the specific scenario.

Discussion on K-means for selecting reference views. We compared the method of selecting reference views using K-means clustering with the method of randomly selecting reference views on the object removal task using the ground truth SPIn-NeRF dataset [6]. We found that in the coarse stage, there was not much difference between the two methods in propagating attention features from reference views to other views. However, in the fine stage, the reference views selected by K-means produced more stable clusters, resulting in more consistent and accurate outcomes when warping reference views to other views.

2. Ablations on Using Different 2D Inpainters

We conduct qualitative ablation studies using different text-guided 2D inpainters, specifically SD-Inpainter [7] and PowerPaint [10], within our methods applied to the SPIn-NeRF [6] datasets. As shown in Fig. 1, our method achieves consistent inpainting results across different 2D inpainters. We observe in (b) that the SD-Inpainter sometimes struggles to deliver successful removal results with complex prompts. In contrast, PowerPaint effectively uses negative prompts to describe the objects to be removed, yielding more accurate results.

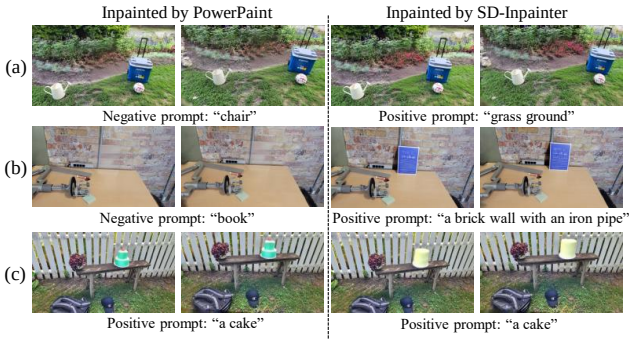

Figure 1. Ablations on using different 2D inpainters, *i.e.*, PowerPaint [10] and SD-Inpainter [7]. (a) and (b) display comparisons for object removal tasks, whereas (c) presents comparisons for object replacement tasks.

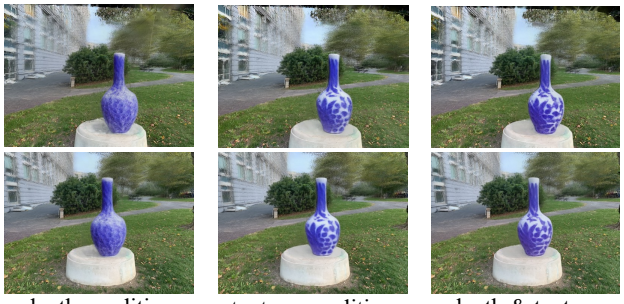

Figure 2. Additional ablation study on the TG-SDS loss.

3. Ablations on TG-SDS loss

As illustrated in Fig. 2, we conduct additional ablation studies on our TG-SDS loss with positive text prompts, such as 'a vase textured with some flowers', which includes intricate texture details and specific geometry. By integrating both texture and depth conditions into the TG-SDS loss, we can achieve improved texture and detailed geometry not only for foreground objects but also for the background.

*Corresponding authors.

Figure 3. A failure case of the object replacement task.

4. Additional Quantitative Results

We present additional no-reference measurements on two key metrics, specifically MUSIQ [4] and Corrs (number of high-quality correspondences between random pairs of frames). These metrics are commonly utilized to evaluate the aesthetic and geometric quality of images. We provide a comparison with NeRFiller across various scenes, demonstrating the capability of both object removal and replacement tasks. As indicated in Tab. 1, our method achieves significantly superior results on both MUSIQ and Corrs metrics, underscoring the enhanced aesthetic and geometric quality facilitated by our approach.

Methods	Removal		Replacement	
	MUSIQ ↑	Corrs ↑	MUSIQ ↑	Corrs ↑
NeRFiller [8]	65.55	7343	65.25	7223
DiGA3D (Ours)	**68.89**	**7421**	**68.70**	**7512**

Table 1. Results of the two tasks with MUSIQ and Corrs.

5. Additional Qualitative Results

We provide supplementary qualitative results for a range of inpainting tasks utilizing the SPIn-NeRF dataset [6], LLFF dataset [5], MipNeRF360 dataset [1], and Instruct-NeRF2NeRF dataset [3].

6. Additional Results for Object Removal

As presented in Fig. 4, we present three additional object removal examples across different scenes from the SPIn-NeRF [6] dataset. In the first two scenes, we successfully remove objects that lack corresponding ground truth data in the original dataset. This removal is achieved using text prompts.

7. Additional Results for Object Re-Texturing

In Fig. 5, we present additional object re-texturing results across various scenes and prompts. These further demonstrate the effectiveness of our method.

8. Additional Results for Object Replacement

Furthermore, as illustrated in Fig. 6, we present additional object replacement results to further evaluate the diversity and generalizability of our methods. By employing different text prompts within a single scene, we produce various object replacement outcomes.

9. Details of User Study

Similar to GaussianEditor [2], we created six questions with the videos of novel view rendering results for the object re-texturing task questionnaire (including the scenes presented in our main paper), each featuring the original scene, text instructions, and re-texturing results from IN2N [3], GaussianEditor [2], GaussCtrl [9], and our method, all labeled randomly. Participants selected their preferred outcome, and after 18 participants completed the questionnaires, we collected a total of 108 votes.

10. Limitations and Future Work

The object replacement task in 360-degree scenes may encounter multi-face Janus problems when the replaced object significantly differs in shape and appearance from different views. We aim to address this issue by designing view priors in the future.

Analysis of Failure Cases: As shown in Fig. 3, we show a failure case where we attempt to 'replace the tractor with a cup of coffee'. The handle of the coffee cup is visible in multiple views, causing a multi-face issue. This common challenge may arise from the substantial geometric changes from a tractor to a coffee cup, and the limitation of the diffusion model and SDS optimization for fine-grained geometric inpainting, particularly noticeable in view 3.

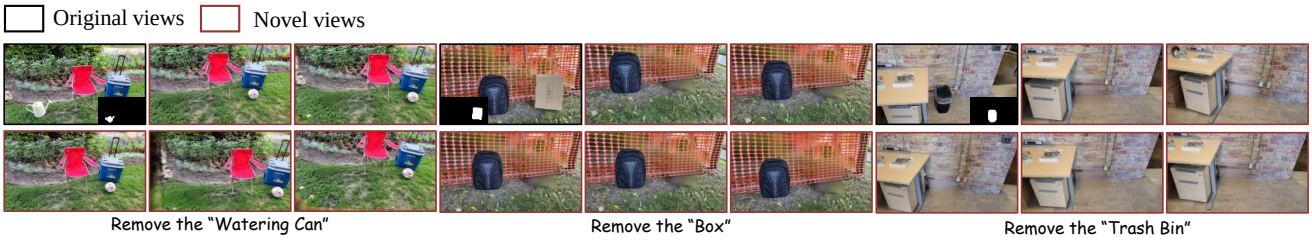

Figure 4. Additional object removal results.

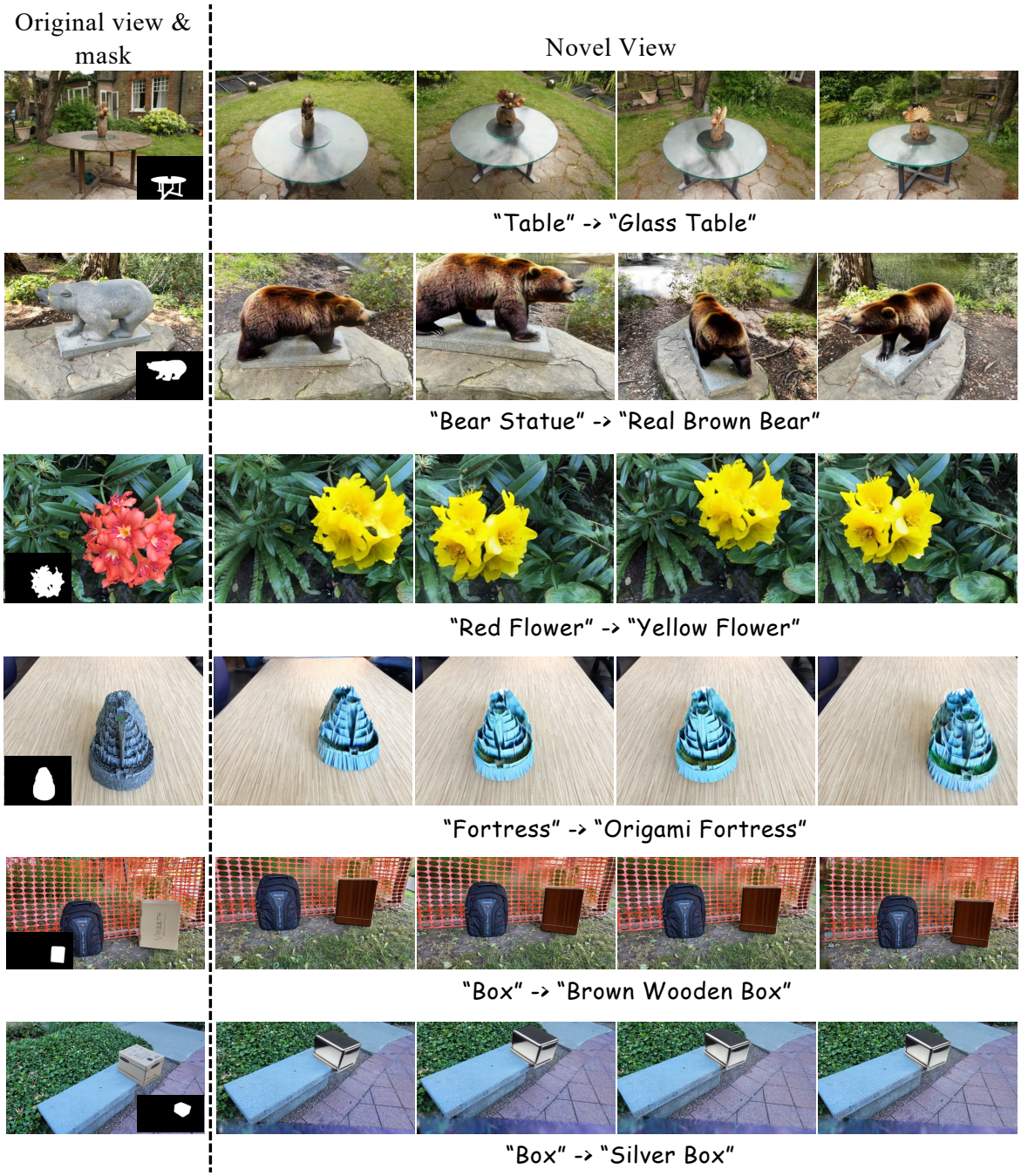

Figure 5. Additional object re-texturing results.

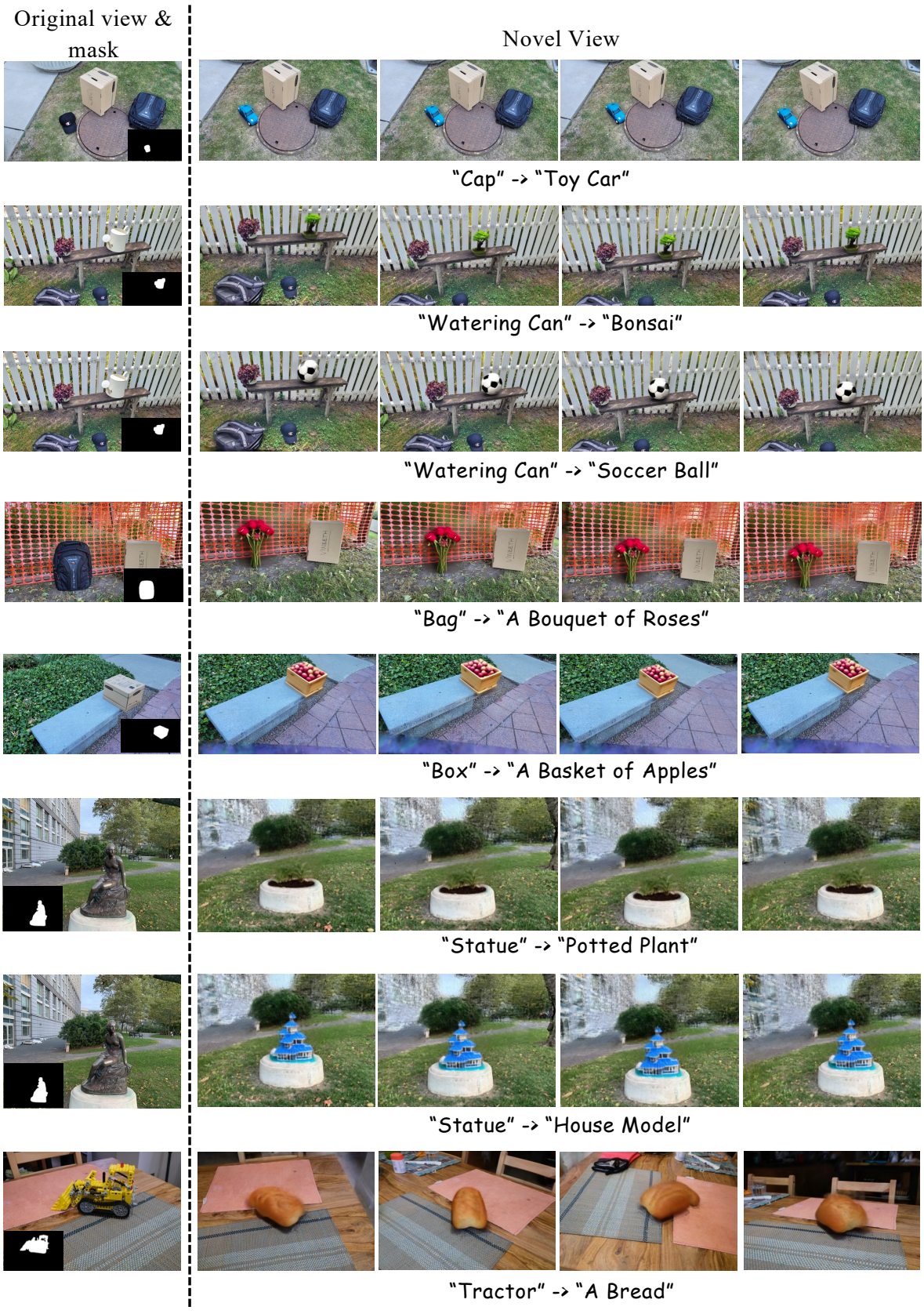

Figure 6. Additional object replacement results.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2

[2] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21476–21485, 2024. 2

[3] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 2

[4] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 2

[5] 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. *ACM Transactions on Graphics (TOG)*, 2019. 2

[6] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 1, 2

[7] 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

[8] Ethan Weber, Aleksander Holynski, Varun Jampani, Saurabh Saxena, Noah Snavely, Abhishek Kar, and Angjoo Kanazawa. Nerfiller: Completing scenes via generative 3d inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20731–20741, 2024. 2

[9] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. Gaussctrl: Multi-view consistent text-driven 3d gaussian splatting editing. In *European Conference on Computer Vision*, pages 55–71. Springer, 2024. 2

[10] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*, 2023. 1