DreamDissector: Learning Disentangled Text-to-3D Generation from 2D Diffusion Priors Supplementary File

Zizheng Yan[123], Jiapeng Zhou[123], Fanpeng Meng[123], Yushuang Wu[123], Lingteng Qiu[123], Zisheng Ye[123], Shuguang Cui[213], Guanying Chen[13], and Xiaoguang Han[213⋆]

[1] Shenzhen Future Network of Intelligence Institute [2] SSE, CUHKSZ
[3] Guangdong Provincial Key Laboratory of Future Networks of Intelligence, CUHKSZ

1 More Results

1.1 Text Splitting.

The Category Score Distillation Sampling (CSDS) requires splitting the input text prompts into individual objects. We employ GPT-4 for this purpose, a method commonly used and effective for information extraction [5]. We empirically found that GPT-4 has the ability to split very complex text prompts, as shown in Figure 1.

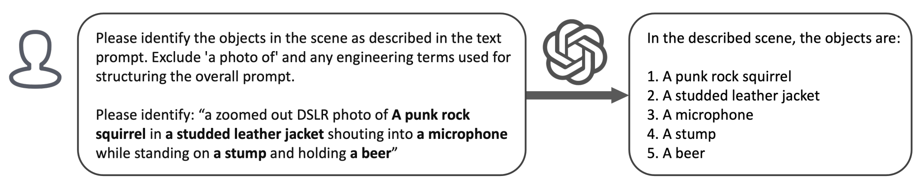

Fig. 1: Text splitting.

1.2 Applications on texture editing

We provide more results on text-guided texture editing, as shown in Figure 2. It can be observed that our method offers greater controllability compared to TEXTure [3].

1.3 Limitations

DreamDissector is likely to fail when objects are in very close contact, such as the body and clothing. We present two examples of this failure in the figure below. The primary reason is the challenge of obtaining clean NeCFs for such complex interactions.

⋆ Corresponding Author.

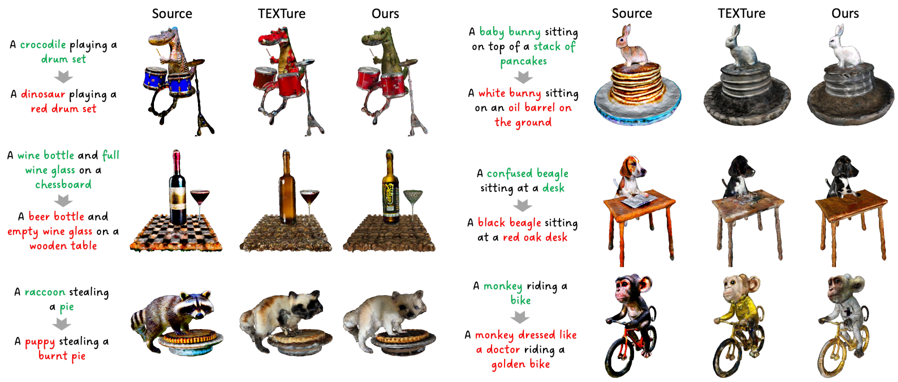

Fig. 2: Text-guided texture editing.

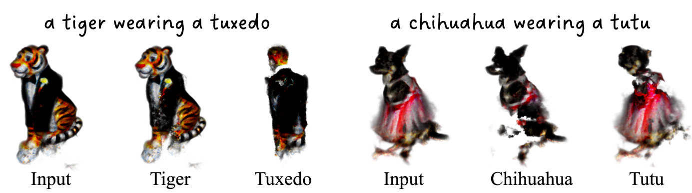

Fig. 3: Failure cases.

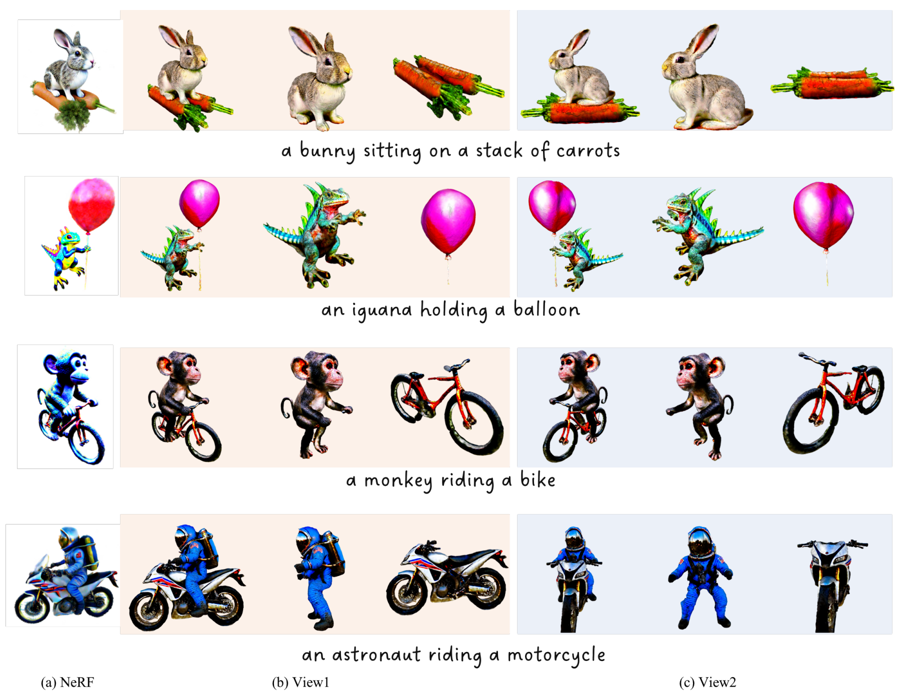

Fig. 4: Qualitative results based on MVDream [4].

1.4 Results on MVDream

We adopt Dreamfusion [2] as the backbone method for generating the initial text-to-3D NeRF for our main results. To verify the versatility of DreamDissector against different backbone methods, we employ MVDream [4], a recently proposed text-to-3D method, as the backbone. Results are shown in Figure 4. It can be observed that DreamDissector successfully dissects MVDream and produces independent textured meshes with improved geometries and textures.

1.5 Results on disentangled text-to-3D generation

We present additional results on disentangled text-to-3D generation, including those featured in the main paper. These results and text prompts are depicted in Figure 5, 6 and 7.

1.6 Comparisons with the baselines

Additional comparisons are shown in Figure 8. It should be noted that negative prompting baseline, being intended to generate independent objects, does not associate with composed objects. Therefore, we regard the entire NeRF as the composed object. We also show the results of a text-guided scene generation method, Set-the-Scene [1], shown in Figure 9. These results illustrate the superior performance of our method.

References

1. Cohen-Bar, D., Richardson, E., Metzer, G., Giryes, R., Cohen-Or, D.: Set-the-scene: Global-local training for generating controllable nerf scenes. In: ICCVW (2023)
2. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
3. Richardson, E., Metzer, G., Alaluf, Y., Giryes, R., Cohen-Or, D.: Texture: Text-guided texturing of 3d shapes. arXiv preprint arXiv:2302.01721 (2023)
4. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation (2023)
5. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Chen, E.: Large language models for generative information extraction: A survey. arXiv preprint arXiv:2312.17617 (2023)

Fig. 5: Qualitative results based on Dreamfusion [2].

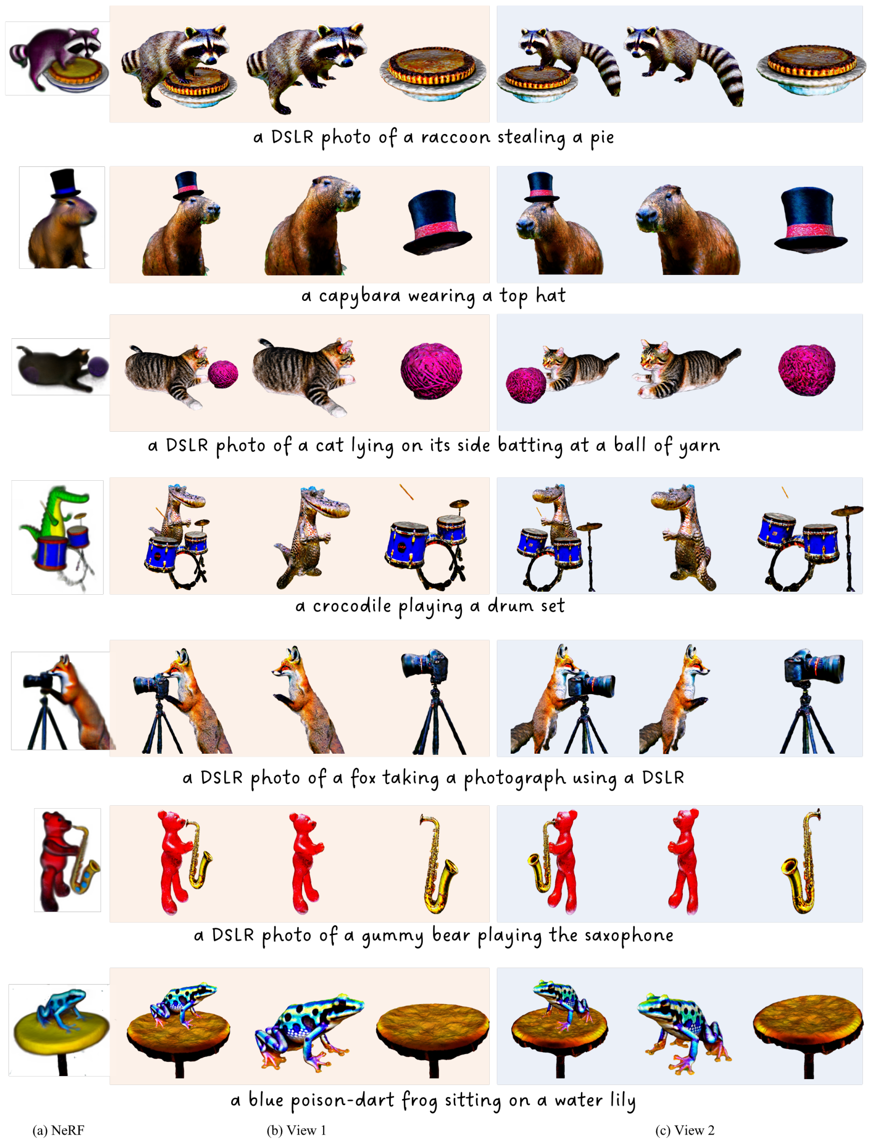

Fig. 6: Qualitative results based on Dreamfusion [2].

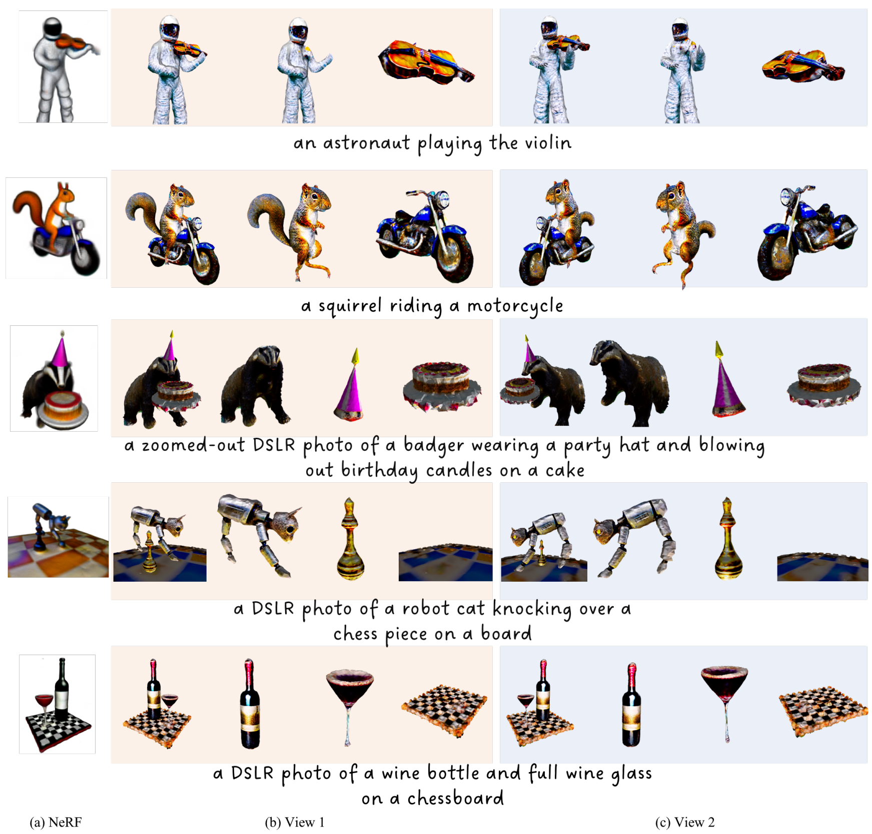

Fig. 7: Qualitative results based on Dreamfusion [2].

Fig. 8: Comparison with baseline methods.

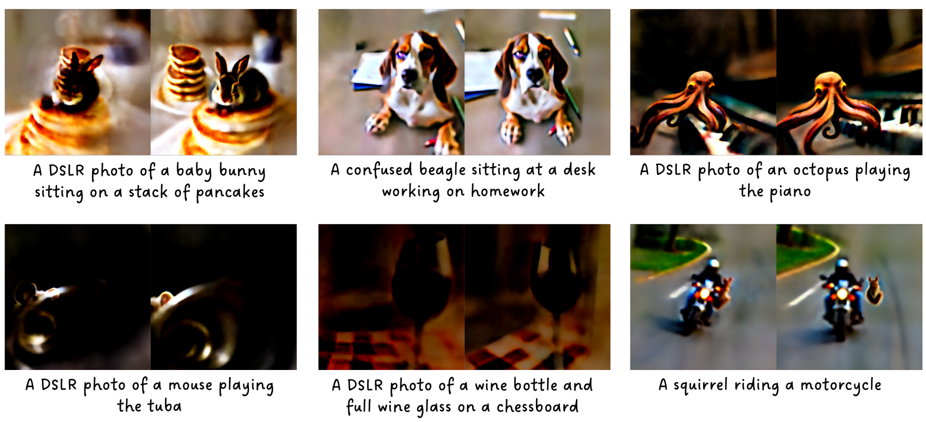

Fig. 9: Results on Set-the-Scene. We show the results on set-the-scene. It can be observed that set-the-scene struggles to model the object-interected scenes.