-Supplementary Material-
GaussianDreamer: Fast Generation from Text to 3D Gaussians by Bridging 2D and 3D Diffusion Models

Taoran Yi[1], Jiemin Fang[2]†, Junjie Wang[2], Guanjun Wu[3], Lingxi Xie[2], Xiaopeng Zhang[2], Wenyu Liu[1], Qi Tian[2], Xinggang Wang[1]†‡
[1]School of EIC, Huazhong University of Science and Technology [2]Huawei Inc.
[3]School of CS, Huazhong University of Science and Technology
{taoranyi, guajuwu, liuwy, xgwang}@hust.edu.cn
{jaminfong, is.wangjunjie, 198808xc, zxphistory}@gmail.com tian.qi1@huawei.com

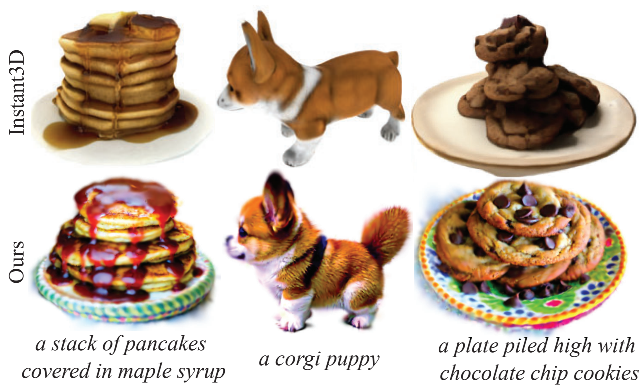


Figure 1. Visual comparisons with Instant3D [7].

A. Appendix

A.1. More Results

Quantitative Comparisons. In Tab. 1, we use CLIP [11] similarity to quantitatively evaluate our method. The results of other methods in the table come from the concurrent Instant3D [7] paper. The results of Shap-E [4] come from the official source, while DreamFusion [9] and ProlificDreamer [15] results come from implementation by threestudio [2]. The implementation version of DreamFusion is shorter in time than the official report we mention in the main text. During the evaluation, we use a camera radius of 4, an elevation of 15 degrees, and select 120 evenly spaced azimuth angles from -180 to 180 degrees, resulting in 120 rendered images from different viewpoints. We follow the Instant3D settings, randomly selecting 10 from the 120 rendered images. We calculate the similarity between each selected image and the text and then compute the average for 10 selected images. It's worth noting that when other methods are evaluated, 400 out of DreamFusion's 415 prompts are selected. This is because some generations failed, so our method is disadvantaged during evaluation on all 415 prompts from DreamFusion. We use two models, *ViT-L/14* from OpenAI [10] [1] and *ViT-bigG-14* from OpenCLIP [3, 13] [2], to calculate CLIP similarity. Our method is superior to all methods except ProlificDreamer, but it is 40 times faster than ProlificDreamer in generation speed. As shown in Fig 1, our method shows notably better quality and details than a concurrent work Instant3D but the CLIP similarity increases marginally.

Generation with Ground. When initializing, we add a layer of point clouds representing the ground at the bottom of the generated point clouds. The color of the ground is randomly initialized. Then, we use the point clouds with the added ground to initialize the 3D Gaussians. Fig. 2 shows the results of the final 3D Gaussian Splatting [5].

Diversity. In Fig. 3, we demonstrate the diversity of our method in generating 3D assets by using different random seeds for the same prompt.

Generation with More Fine-grained Prompts. More refined prompts are used to generate 3D assets, as shown in Fig. 4. It can be seen that Shap-E [4] generates similar results when given different descriptions of the word "axe" in the prompt. However, our method produces 3D assets that better match the prompt.

Automatically Select A Human Model. As shown in Fig 5, we attempt to use CLIP to guide the selection of the

[1]https://huggingface.co/openai/clip-vit-large-patch14
[2]https://github.com/mlfoundations/open_clip

Table 1. Quantitative comparisons on CLIP [11] similarity with other methods.

Methods	ViT-L/14 ↑	ViT-bigG-14 ↑	Generation Time ↓
Shap-E [4]	20.51	32.21	6 seconds
DreamFusion [9]	23.60	37.46	1.5 hours
ProlificDreamer [15]	27.39	42.98	10 hours
Instant3D [7]	26.87	41.77	20 seconds
Ours	27.23 $\pm$ 0.06	41.88 $\pm$ 0.04	15 minutes

Figure 2. Results of generation with ground.

Figure 3. Results of the diversity of our method.

initialized human body model, by computing the similarities between images rendered from the generated SMPL models and the text prompt. We can achieve good rendering effects on various human body models. It would also be a promising direction to extend the assets to dynamic ones with the sequence of generated human body models.

A.2. More Ablation Studies

2D Diffusion Model During the process of optimizing 3D Gaussians with a 2D diffusion model, we perform ablation on the 2D diffusion models we use, specifically *stabilityai/stable-diffusion-2-1-base* [12] [3] and *DeepFloyd/IF-I-XL-v1.0* [4]. Fig. 6 shows the results of the ablation experiment, where it can be seen that the 3D assets generated using the *stabilityai/stable-diffusion-2-1-base* have richer details.

[3]https://huggingface.co/stabilityai/stable-diffusion-2-1-base
[4]https://huggingface.co/DeepFloyd/IF-I-XL-v1.0

Box Size in Point Growth In Fig 7, we conduct an ablation experiment on the box size, where a larger box leads to a fatter asset along with a more blurry appearance.

A.3. More Discussions

Limitations Introduced by the 3D Datasets. Fig 8 shows the generation results of complex prompts. The domain-limited 3D diffusion model can only generate parts of the desired object with rough appearances. Our method completes the remaining part and provides finer details by bridging the domain-abundant 2D diffusion model.

Figure 4. Results of generation with more refined prompts.

Recent Works. We discuss with more related work. Our focus is to connect the 3D and 2D diffusion models, fusing the data capacity from both types of diffusion models and generating 3DGS-based assets directly from text. DreamGaussian [14] finally generates mesh-based 3D assets from an image or an image generated from text, which can be orthogonal to our method. There is a possibility of a combination in the future. NerfDiff [1] uses a 3D-aware conditional diffusion to enhance details. DiffRF [8] employs 3D-Unet to operate directly on the radiation field, achieving truthful 3D geometry and image synthesis. 3DDesigner [6] proposes a two-stream asynchronous diffusion module, which can improve 3D consistency.

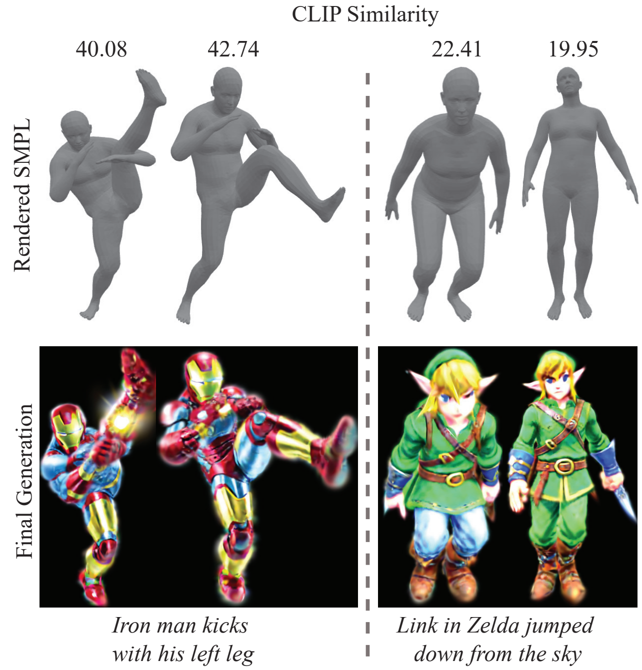


Figure 5. Avatar generation.

Figure 6. Ablation studies of optimizing 3D Gaussians with different 2D diffusion models.

References

[1] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, pages 11808–11826. PMLR, 2023. 3

[2] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 1

[3] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 1

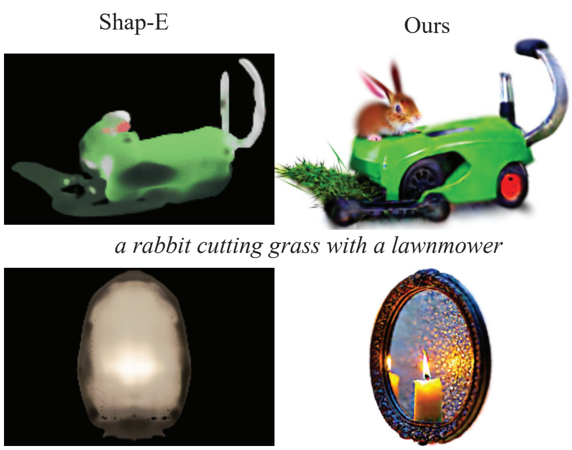
[4] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 2

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[6] Gang Li, Heliang Zheng, Chaoyue Wang, Chang Li, Chang-

Viking axe, fantasy, weapon...

Figure 7. Ablation on the size of the box.



A flickering candle casts a dim light on a dusty mirror

Figure 8. Generation with complex prompts.

wen Zheng, and Dacheng Tao. 3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2211.14108*, 2022. 3

[7] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 1, 2

[8] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4328–4338, 2023. 3

[9] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[11] 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 *ICML*, pages 8748–8763. PMLR, 2021. 1, 2

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[14] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[15] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 1, 2