

Sharp-It: A Multi-view to Multi-view Diffusion Model for 3D Synthesis and Manipulation

Supplementary Material

6. Additional Results

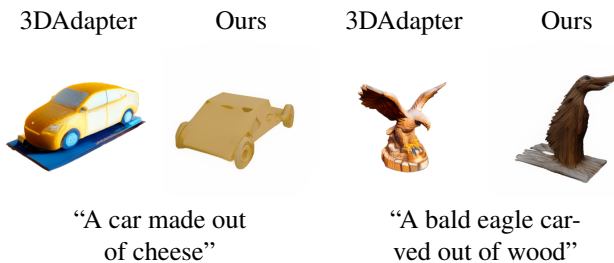
Quantitative Comparison Against Zero123++ Many state-of-the-art text-to-3D methods follow a two-stage pipeline: generating a multi-view set followed by a sparse view reconstruction method. Our work improves the first stage and is compatible with any sparse-view reconstruction method. Since we focus on the first stage of the pipeline, Figure 6 in the main paper qualitatively demonstrates the advantage of our approach over Zero123++ [61], which is a common model for multi-view image synthesis.

In Table 3, we present a quantitative comparison, using prompts from Objaverse’s test set. To generate our results, we first used Shap-E and then refine with Sharp-It, while for Zero123++, we generate an image from the prompt, and then use the model conditioned on this generated image. We measure FID and CLIP similarity between the input text and output images. As shown in the table, our method outperforms Zero123++ in FID while achieving comparable text alignment. This shows that our method does not degrade text alignment, even though it inherits Shap-E’s limitations.

Table 3. Quantitative comparison of our method with Zero123++.

	FID (↓)	CLIP (↑)
Zero123++	51.95	0.262
Ours	38.14	0.264

Comparison Against 3D-Adapter Here we provide a qualitative comparison against a recent text-to-3D work, 3D-Adapter [8]. 3D-Adapter does not follow the common pipeline of first generation a multi-view image set and then reconstruct it.



Sharp-It Reliance on Shap-E Our method relies on Shap-E to produce an initial reasonable result. Yet, our en-

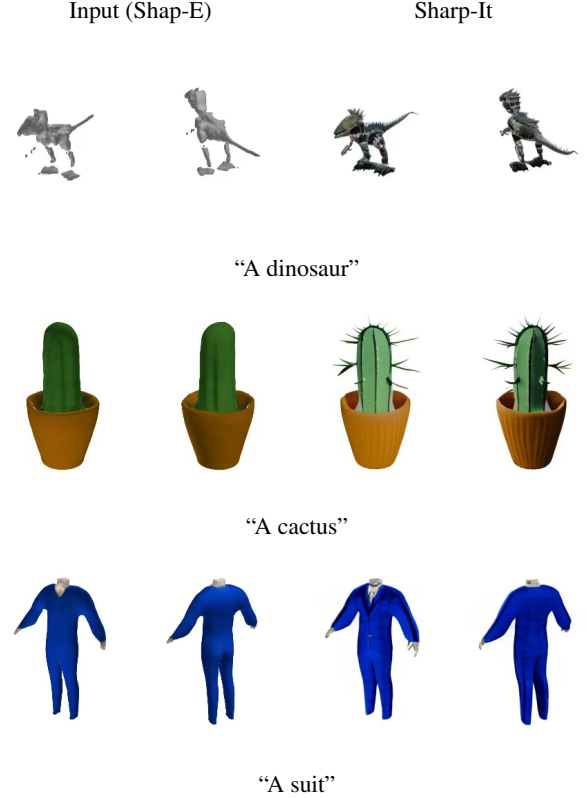


Figure 10. Refinement results of Sharp-It, Resulting from Shap-E Generation.

hancement evaluation (Table 1 in the main paper) was conducted on a test set from the standard Objaverse dataset to illustrate the performance of our method in a general case. All these test objects were encoded into Shap-E’s space and were enhanced with different methods. Moreover, the bust example in Figure 4 in the main paper is also an Objaverse object encoded into this space. More such examples are shown in Figure 12.



Figure 11. A failure case of shap-e generation, Prompt: a man wearing a suit

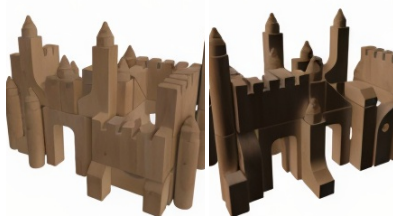
Objaverse original object

Shap-E

Sharp-It



"A jade dragon statue"



"A sand castle"



"A wooden dog statue"

Figure 12. Example of high refinement quality of objects, from objaverse test set, where we encode the object into the Shap-E latent space, and refine it using Sharp-It