

We present additional ablation study and visual results in the **Appendix manuscript**. Please also refer to our **website demo** in the supplementary material for a comprehensive overview.

A ABLATION STUDY

DiNO encoder. We conduct experiments using different encoders. In Fig. 12 we present the validation PSNR curves during training, when using either our convolutional image encoder or a pre-trained DiNO ViT (Caron et al., 2021). Specifically, our convolutional encoder is a single layer that downsamples input images from 512 to 32. The triplane generator is a self-attention transformer, identical in both settings. We train both models on 8 80G A100 GPUs. We observe that the DiNO experiment did not show improved convergence during the initial iterations. Alternatively, more careful designs could optimize the use of DiNO ViT, which we leave for future study.

Objaverse training dataset. In Fig. 12, we also show the training process with a dataset consisting solely of 100k Objaverse (Deitke et al., 2023) images. We did not observe a performance drop in the early stage compared to the other experiments in the figure, which were trained on our mixed dataset.

Vae encoder. In Fig.13, we show preliminary results using a pretrained VAE encoder¹ from an SD model (Rombach et al., 2022). To enable the VAE encoder to handle multi-channel input, we separately provide images, masks, and the camera rays to the encoder, then assemble the output features using a convolution layer. Experiments are run on 32 80G A100 GPUs. We observe that using a pretrained VAE encoder leads to better convergence in the early training stage. We attribute this to the good initialization provided by VAEs compared to training the convolutional encoder from scratch.

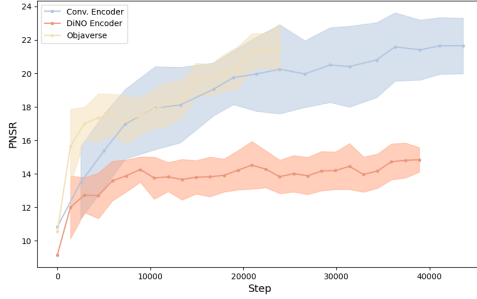


Figure 12: Ablation study on image encoders (Conv. vs DiNO) and dataset.

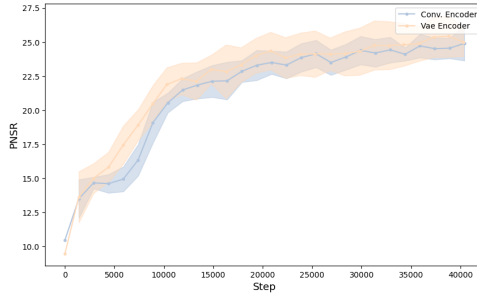


Figure 13: Ablation study on image encoders (Conv. vs VAE).

B ADDITIONAL RESULTS

We show additional results generated by our approach in Fig. 14- 16.

¹In practice, we use the pretrained SD VAE from <https://huggingface.co/madebyollin/taesd>

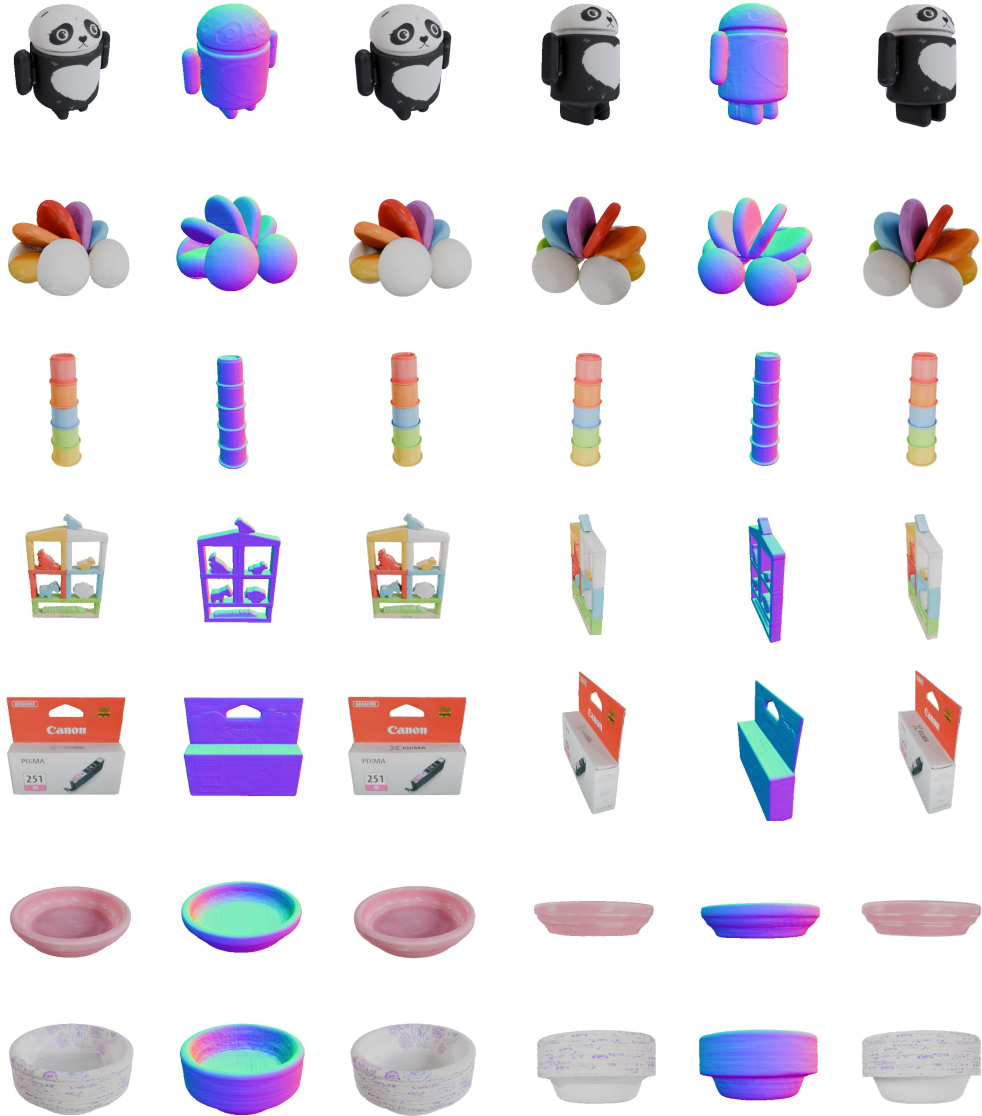


Figure 14: Additional visual 3D reconstruction results on GSO (Downs et al., 2022) dataset. The input of our model is 4 orthogonal views. We show the novel view generated RGB images (column 1, 4) and normal images (column 2, 5), and the ground-truth images (column 3, 6).

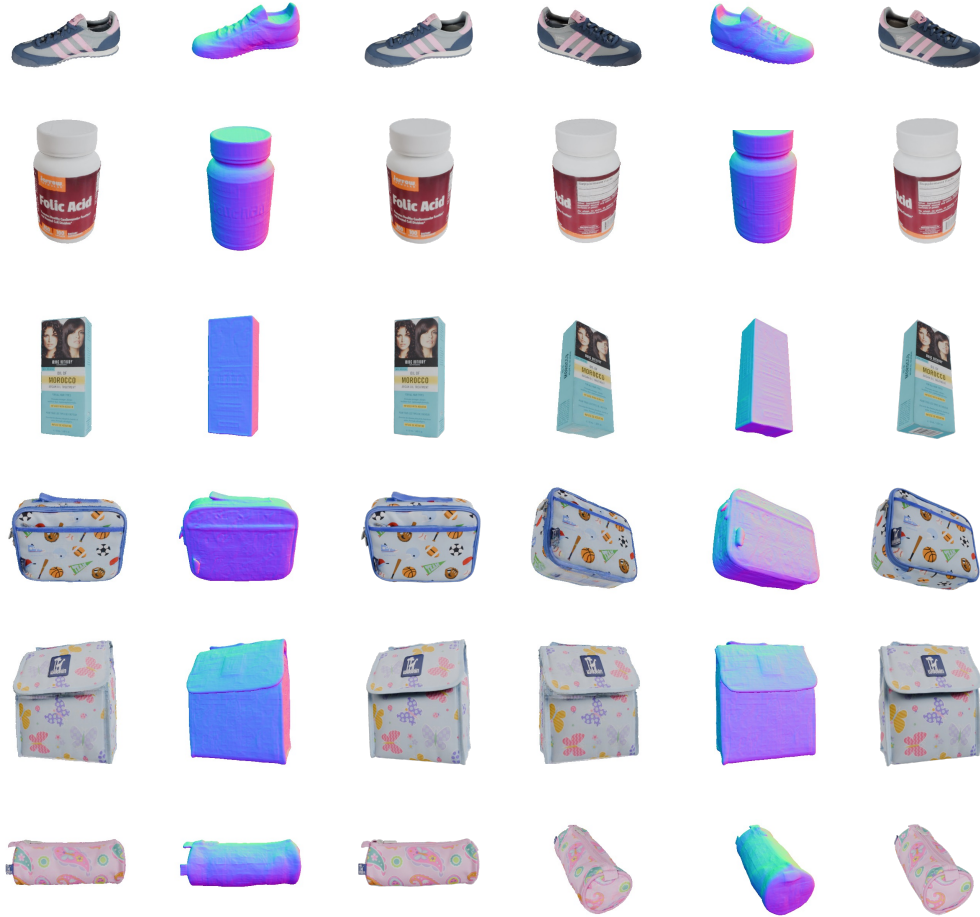


Figure 15: **Additional visual 3D reconstruction results on GSO (Downs et al., 2022) dataset.** The input of our model is 4 orthogonal views. We show the novel view generated RGB images (*column 1, 4*) and normal images (*column 2, 5*), and the ground-truth images (*column 3, 6*).

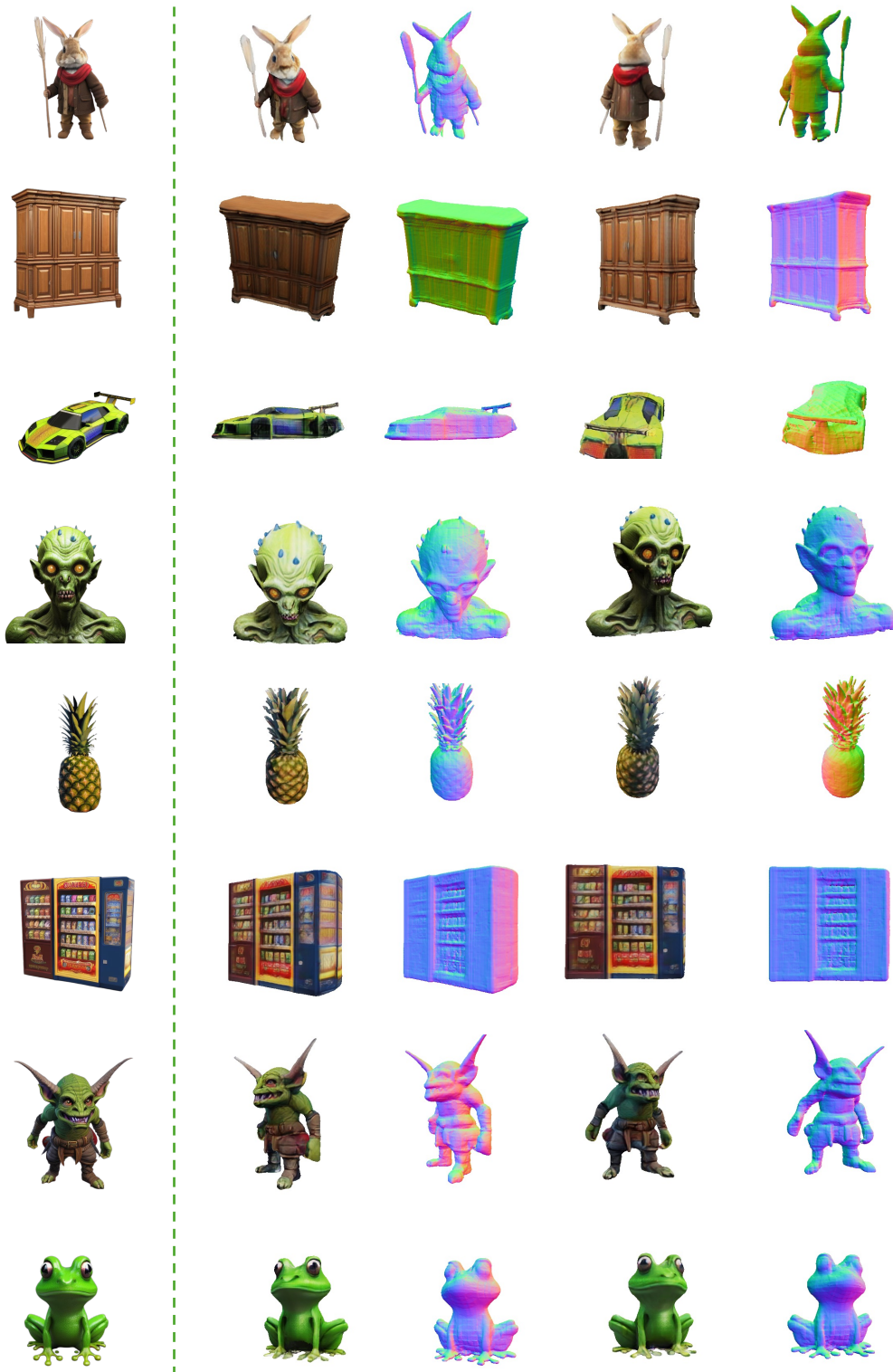


Figure 16: **Additional visual 3D asset generation results.** The input images (*column 1*) are either generated from text using pre-trained text-to-image diffusion models (Rombach et al., 2022) or online generated images.