A. Limitations and Discussion

Due to computational resource limitations, we are not able to process and train our model on the full Objaverse dataset. Currently, the meshes from Objaverse we use only consist of 5% of Objaverse and 0.4% of Objaverse-XL objects. Based on the promising scaling properties of recent foundation models [3, 4, 10], we believe it will be valuable to explore the scaling properties of method.

Another limitation of our work is that we have not considered the modeling of object texture. Predicting textures of unseen surfaces is highly ill-posed and can greatly benefit from a strong 2D prior. Given the recent success of 2D diffusion models [8] and their application in optimization-based 3D generation methods [1, 2, 5–7, 9], we think it will be promising to initialize or regularize these methods with our shape prior, potentially boosting both the optimization efficiency and generation quality.

B. Additional Comparison

Additional qualitative results. We show additional qualitative results on OmniObject3D, Ocrtoc3D and Pix3D in Fig. 1, Fig. 2 and Fig. 3, respectively. Comparing with prior arts, the reconstruction of ZeroShape better captures the global shape structure and visible geometric details.

Additional quantitative comparison. We additionally compare ZeroShape to the optimization-based approach, SDS w/ Zero123 [1]. Due to the low efficiency of optimization-based approaches, we randomly sampled 30 objects on OmniObject3D for evaluation. In this evaluation, ZeroShape achieves an FS@2 of 0.4630, significantly higher than that of Zero123+SDS, 0.3826.

Additional ablation. We further ablate the performance by training MCC with our curated data. We improve the coordinate system in MCC to be view-centric for stable convergence. Tab. 1 shows that training MCC on our curated data (MCC-Zero) significantly improves its performance, verifying our data contribution. On the other hand, the performance gap between MCC-Zero and ZeroShape, and our 5× higher inference speed, demonstrate the efficacy of our design choices.

Method	MCC-CO3D	MCC-Zero	ZeroShape
FS@2↑	0.3215	0.4283	**0.4927**

Table 1. Additional ablation on OmniObject3D.

C. Inference on AI-generated Images

We present additional results of ZeroShape using images generated with DALL·E 3. To test the out-of-domain generalization ability, we generate images of imaginary objects as the input to our model (see Fig. 4). Despite the domain gap to realistic or rendered images, ZeroShape can faithfully recover the global shape structure and accurately follow the local geometry cues from the input image. These results also demonstrate the potential of using ZeroShape in a text-based 3D generation workflow.

D. Data Curation Details

In this section we describe our data generation procedure for training and for rendering the object scans from OmniObject3D to generate one of our benchmark test sets.

D.1. Synthetic Training Dataset Generation

Image Rendering. For an arbitrary 3D mesh asset, our Blender-based rendering pipeline first loads it into a scene and normalizes it to fit inside a unit cube. Our scene consists of a large rectangular bowl with a flat bottom, a common scene setup that 3D artists use for rendering to allow for realistic shading, and 4 point light sources and one area light source. We randomly place cameras around the object with 30mm to 70mm focal length for a 35mm sensor size equivalent. We randomly vary the distance, elevation (from 5 to 65 degrees), the LookAt point of the camera and generate images of 600×600 resolution (see Fig. 5). This variation in object/camera geometry allows capturing the variability of projective geometry in real world scenarios, coming from different capture devices and camera poses. This is in contrast with prior work that uses fixed intrinsics, fixed distance, and LookAt pointed at the center of the object.

In addition to RGB images, we extract segmentation masks, depth maps, intrinsics, extrinsics and object pose. We center crop the objects, mask out the background, resize images to 224×224 and process the additional annotations to account for the crop, segmentation and resize.

Ground Truth Occupancy Extraction. To obtain ground truth occupancy, we first extract watertight meshes using the code from occupancy networks [2], and then extract SDF for 32^3 query points per mesh following DISN [3]. The SDF is converted to occupancy during training.

D.2. Generating the OmniObject3D Testing Set

The original videos released by the OmniObject3D dataset have noisy foreground masks and are mostly taken indoor on a tabletop. To improve the lighting variability and ensure accurate segmentations, we follow the rendering procedure described in the previous section to generate testing data. Different from our training set generation, we use HDRI environment maps to generate scene lighting, which results in high lighting quality and diversity (see Fig. 6).

[1] https : / / github . com / threestudio - project / threestudio#zero-1-to-3-

[2] https://github.com/autonomousvision/occupancy_networks

[3] https://github.com/laughtervv/DISN

Figure 1. Additional qualitative results and comparison on OmniObject3D.

References

[1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[2] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[5] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1

[6] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and

Figure 2. Additional qualitative results and comparison on Ocrtoc3D.

Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023.

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

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[9] 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

[10] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

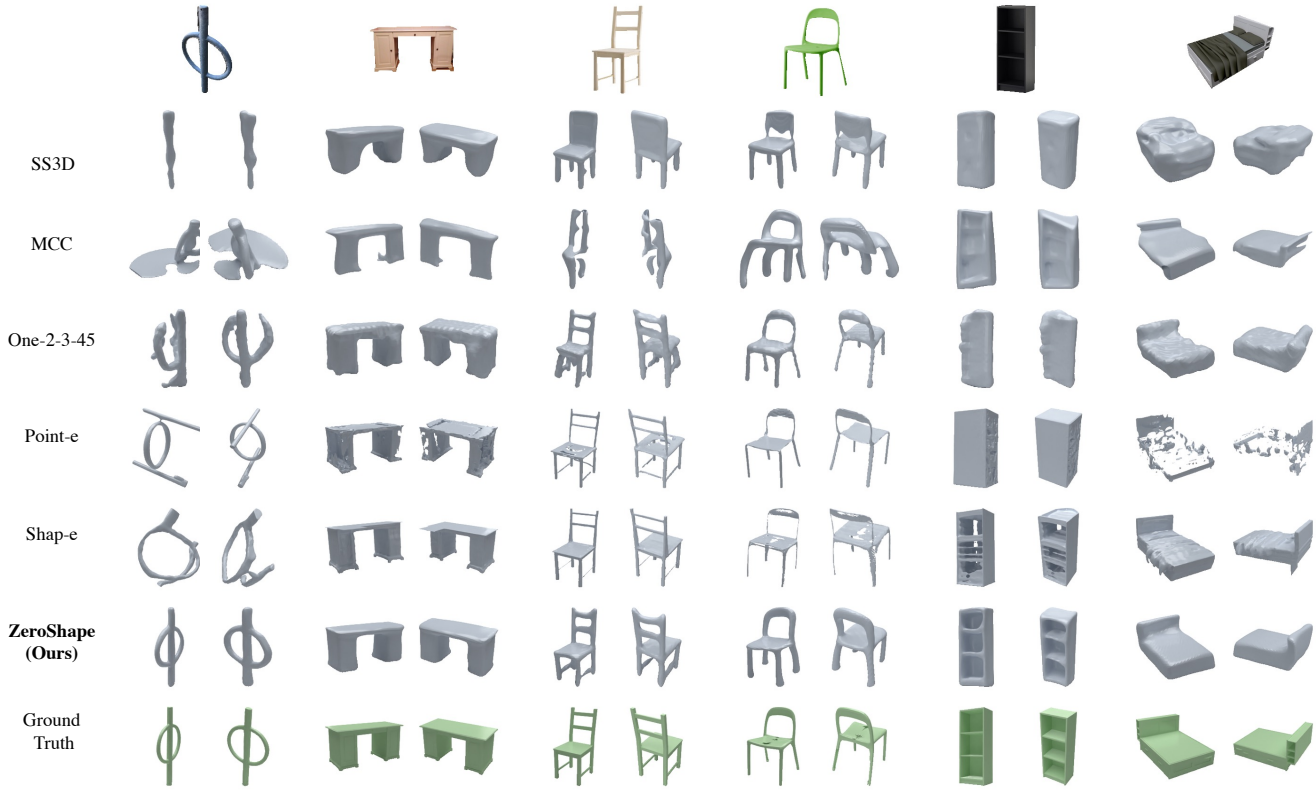

Figure 3. Qualitative results and comparison on Pix3D.

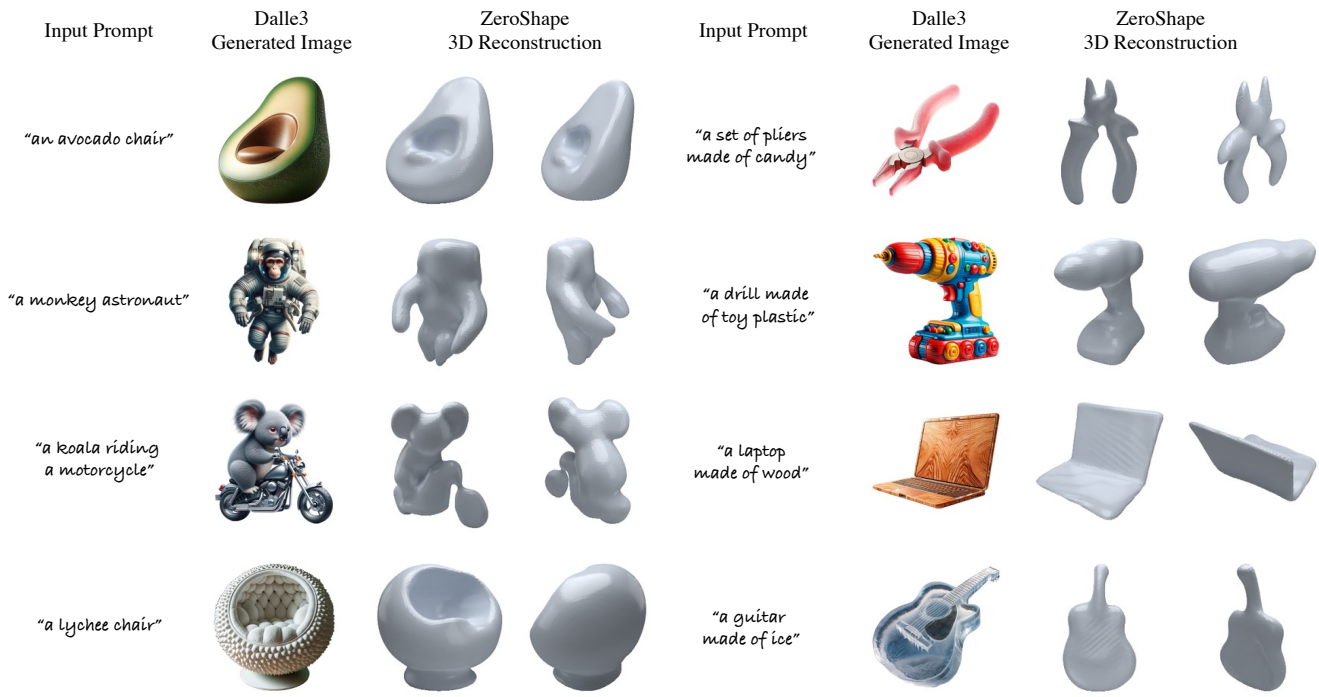

Figure 4. Qualitative results on images generated with DALL·E 3. These results demonstrate the zero-shot generalization ability of ZeroShape to complex novel images.

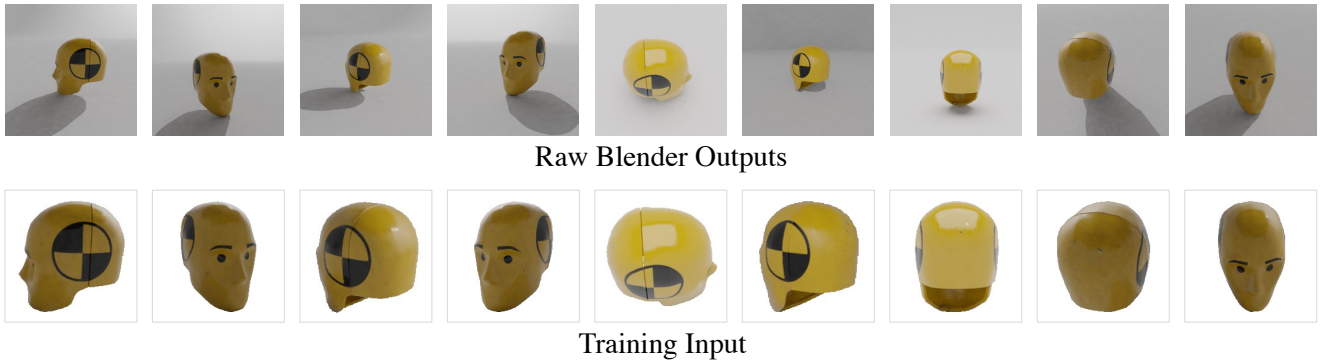

Figure 5. **Synthetic Training Data Generation.** We render training images with varying lighting, camera intrinsics and extrinsics. The images are center-cropped, foreground-segmented and resized before being used as training input.

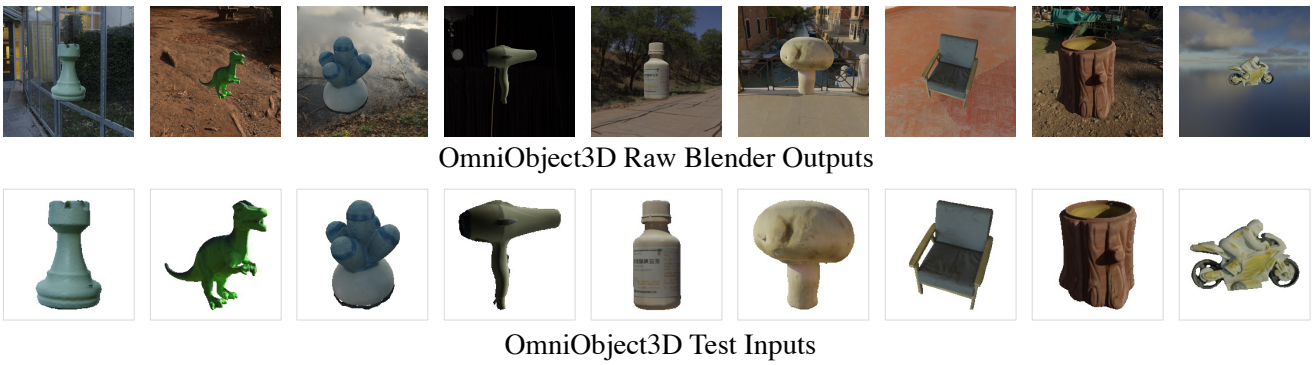

Figure 6. **OmniObject3D Testing Data Generation.** For OmniObject3D, we generate realistic testing images with varying lighting, camera intrinsics and extrinsics. To increase rendering realism and diversity, we use diverse HDRI environment maps for scene lighting.