

Dora: Sampling and Benchmarking for 3D Shape Variational Auto-Encoders

Supplementary Material

This supplementary material provides additional details and results to complement our main paper. We first present implementation details (Appendix A), followed by extensive comparisons of VAE performance (Appendix B) and 3D generation comparisons between our method and baselines (Appendix C). We conclude with a discussion of the limitations and future work in Appendix D.

A. More implementation details

Data Processing. Our training data consists of approximately 400,000 3D meshes carefully filtered from Objaverse [8]. Following CLAY [22], we preprocess all meshes to ensure watertight geometry. The dataset is randomly split into training and test sets, where the test set is further utilized to construct our Dora-bench benchmark.

Dora-bench Construction. We introduce Dora-bench, a comprehensive benchmark designed to evaluate 3D reconstruction quality across different levels of geometric complexity. The benchmark integrates data from multiple sources: ABO [7], GSO [10], Meta [3], and Objaverse [8] test set. The benchmark categorizes models into four detail levels (Level 1 to Level 4), with approximately 800 samples per level. Due to the scarcity of highly detailed models in ABO, GSO, and Meta datasets, Level 4 samples are predominantly sourced from the Objaverse test set.

Evaluation Metrics. We employ multiple complementary metrics to comprehensively evaluate reconstruction quality. To assess fine-grained geometric details, we compute the Sharp Normal Error (SNE) by rendering normal maps from 22 fixed, evenly spaced viewpoints around each object using nvdiffrast [13]. For quantitative evaluation of overall geometric accuracy, we utilize the Kaolin library [11] to compute two additional metrics: F-score, which measures the coverage and completeness of the reconstructed shape, and Chamfer Distance (CD), which evaluates the bi-directional similarity between the reconstructed and ground truth point clouds.

VAE Architecture and Training. Our VAE architecture follows recent successful designs [14, 23], with 8 self-attention layers in the encoder and 16 in the decoder. For sharp edge sampling, we set the number of sampled points $N_d = 32768$, target sharp points $N_{\text{desired}} = 16384$ and angle threshold $\tau = 30$. Following 3DShape2VecSet [21], we construct Q_{space} by combining two types of point sampling: points randomly sampled near the mesh surface and points uniformly sampled within the spatial range of $[-1, 1]$.

We adopt the multi-resolution training strategy proposed in CLAY [22], where the latent code length (LCL) N_s is

randomly selected between 256 and 1280 during training. This approach facilitates progressive training in the subsequent diffusion stage. The KL divergence weight is set to 0.001. We train our Dora-VAE on the Objaverse [8] training set using 32 A100 GPUs with a batch size of 2048 for two days.

Diffusion Model for Image-to-3D. We apply our Dora-VAE to the downstream image-to-3D task. Specifically, we implement a conditional diffusion model based on the DiT architecture [5, 17], similar to Direct3D [20] and CLAY [22]. The model conditions on image features extracted by DINOv2 [16] from single-view images rendered using BlenderProc [9]. Our diffusion model contains 0.39 billion parameters and is trained on 32 A100 GPUs for three days.

B. More comparison of VAE

We present comprehensive quantitative and qualitative comparisons of our Dora-VAE against existing methods on the Dora-bench dataset. In addition to the baselines discussed in the main paper, we include 3DShape2VecSet [21], which was trained on ShapeNet [4] rather than the larger Objaverse [8] dataset.

Quantitative Results. We see in Table S1, 3DShape2VecSet [21] consistently underperforms across all detail levels, primarily due to its limited training data affecting generalization capability.

Qualitative Evaluation. Figures S1 and S2 present visual comparisons for Level 3 and 4 examples (specific data sources listed in Table S2). For XCube [18], we present only its fine-tuned version (XCube^f) as it slightly outperforms the original version. We see our Dora-VAE outperforms all other baselines. While XCube demonstrates rich visual details, we observe that its geometry sometimes deviates from the ground truth mesh. We attribute this to quantization errors introduced during mesh extraction using NKSR [12], which explains its lower performance in metrics like chamfer distance (CD) and SNE despite visually appealing results.

C. Image-to-3D Generation Comparison

We evaluate our Dora-VAE-based latent diffusion model against state-of-the-art methods for single-image 3D generation. Our comparison includes 1) LRM-based methods: MeshFormer [15] and CRM [19], as well as 2) industry solution: Tripo v2.0 [2]. We use the official code and model provided by CRM [19] for inference and obtain the results of MeshFormer [15] and Tripo v2.0 [2] from their hugging-

Methods	LCL	\uparrow F-score(0.01) \times 100				\uparrow F-score(0.005) \times 100				\downarrow CD \times 10000				\downarrow SNE \times 100				
		L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	L1	L2	L3	L4	
Xcube [18]	>10000	98.968	98.799	98.615	98.226	95.525	93.872	92.322	85.365	6.315	6.288	7.935	9.926	1.579	1.432	1.430	1.679	
Xcube [†] [18]	>10000	99.393	99.794	99.824	99.079	96.753	95.535	93.422	87.365	4.015	4.142	5.740	7.627	1.543	1.408	1.259	1.639	
VecSet [21]	512	94.768	88.890	80.126	59.347	77.545	67.929	55.516	34.619	27.380	42.075	100.975	159.151	2.939	3.056	3.470	6.034	
Craftsman [14]	256	98.016	95.874	91.756	81.739	87.994	82.549	73.000	57.379	4.389	9.129	14.530	33.441	1.906	1.873	2.191	3.933	
Ours	w/o DCA	1280	99.964	99.925	99.678	97.890	96.561	95.975	91.618	83.124	2.236	2.506	4.444	6.432	1.448	1.215	1.205	1.828
	w/o SES,DCA	1280	99.944	99.814	97.294	96.779	95.977	94.623	88.406	79.240	2.422	2.983	3.980	6.196	1.496	1.313	1.352	2.207
Ours	full	256	99.507	98.986	96.669	89.577	93.272	90.466	82.386	68.669	3.356	5.202	10.276	24.527	1.555	1.410	1.618	3.035
		1280	99.988	99.955	99.880	99.170	97.038	96.831	93.458	87.473	2.097	2.500	3.945	5.265	1.433	1.186	1.137	1.579

Table S1. Quantitative comparison in Dora-bench. [†] indicates the fine-tuning model that uses the same training data as ours.

```

Objaverse/000-002/0aecee43ac2749499a16ab4388a0baa2
ABO/B07MF1TQYR
Meta/meta_ADT_1.0_WireShelvingUnitBlackSmall_1.3d
Objaverse/000-009/yC4xcavDg89GZgqjJtUs5KYbrgz
ABO/B07V4FNHCD
Objaverse/000-149/61ace9488e1c45718530e2d8f9da4a9d
GSO/Sootheze_Cold.Therapy.Elephant
ABO/B07JYN8DBM
ABO/B07N6Q9JB1
ABO/B07XJB28C7
Meta/meta_DTC_1.0_BasketPlasticRectangular_3d
Objaverse/000-008/8f0d1d4df2d64d1aa7c062868ca09535
GSO/Rubbermaid.Large.Drainer
Meta/meta_DTC_1.0_Pottery_B0CJJ59SLH_BlueHairFairy_3d

```

Table S2. Data sources for Figure S1 and S2

face demo and product website. Note that CLAY [22] and Rodin Gen-1 [1] are excluded due to implementation unavailability and usage limitations at submission time.

As demonstrated in Figures S3 and S4, our method achieves superior results compared to LRM-based approaches in terms of both geometric detail and fidelity. The performance limitations of MeshFormer and CRM can be attributed to their lack of explicit geometric constraints, leading to unstable or lower-quality reconstructions.

Our method achieves comparable geometric quality to Tripo v2.0, a leading commercial solution, while using significantly more constrained resources. Specifically, we achieve these results with only three days of training on 32 A100 GPUs and approximately 400,000 training samples. This remarkable performance, achieved with limited computational resources and training data compared to commercial solutions, demonstrates the effectiveness of Dora-VAE in enhancing geometric detail and improving diffusion model performance.

D. Limitations and Future Directions

While Dora-VAE achieves state-of-the-art reconstruction quality with 1,280 latent code tokens, we identify several

limitations and promising directions for future research.

Current Limitations. The primary limitation of our approach lies in maintaining high-quality reconstructions when further reducing the number of latent tokens. This challenge becomes particularly evident when comparing with recent advances in 2D domain, such as Deep Compression Autoencoder (DC-AE) [6], which has achieved remarkable compression rates while preserving reconstruction quality.

Future Directions. We envision two main directions for future work: 1) *Enhanced Compression Efficiency*: We aim to explore novel techniques for increasing the compression rate of 3D VAEs while maintaining reconstruction quality. This research direction could potentially bridge the efficiency gap between 2D and 3D compression methods. 2) *Advanced Diffusion Models*: Building upon Dora-VAE’s superior reconstruction capabilities, we plan to develop more powerful image-to-3D diffusion models. We believe that the improved reconstruction quality offered by Dora-VAE can directly boost the performance ceiling of diffusion models, enabling higher-quality generation results under the same training conditions.

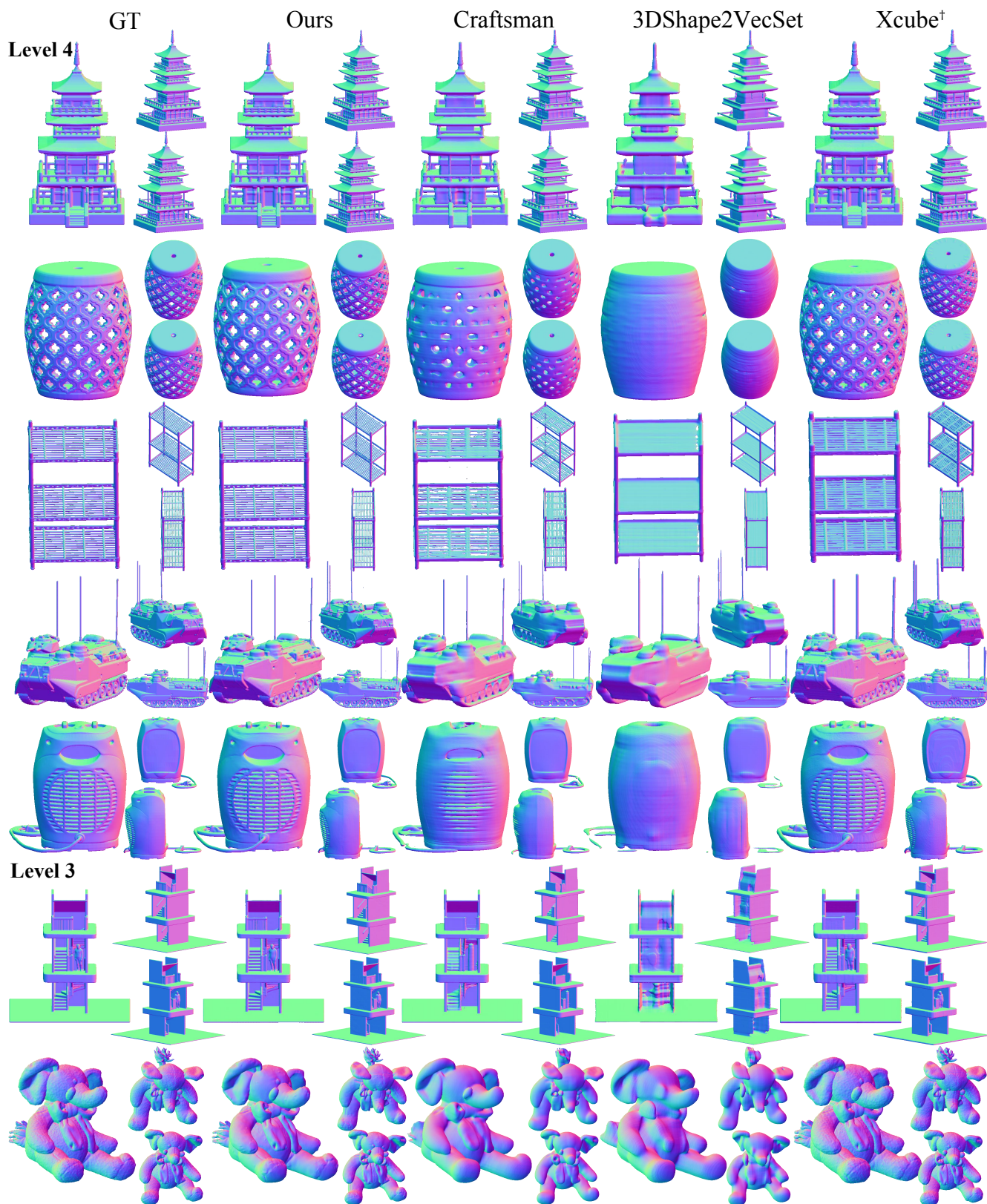


Figure S1. Qualitative comparison of the VAE reconstruction results. [†] indicates the fine-tuning model that uses the same training data as ours.

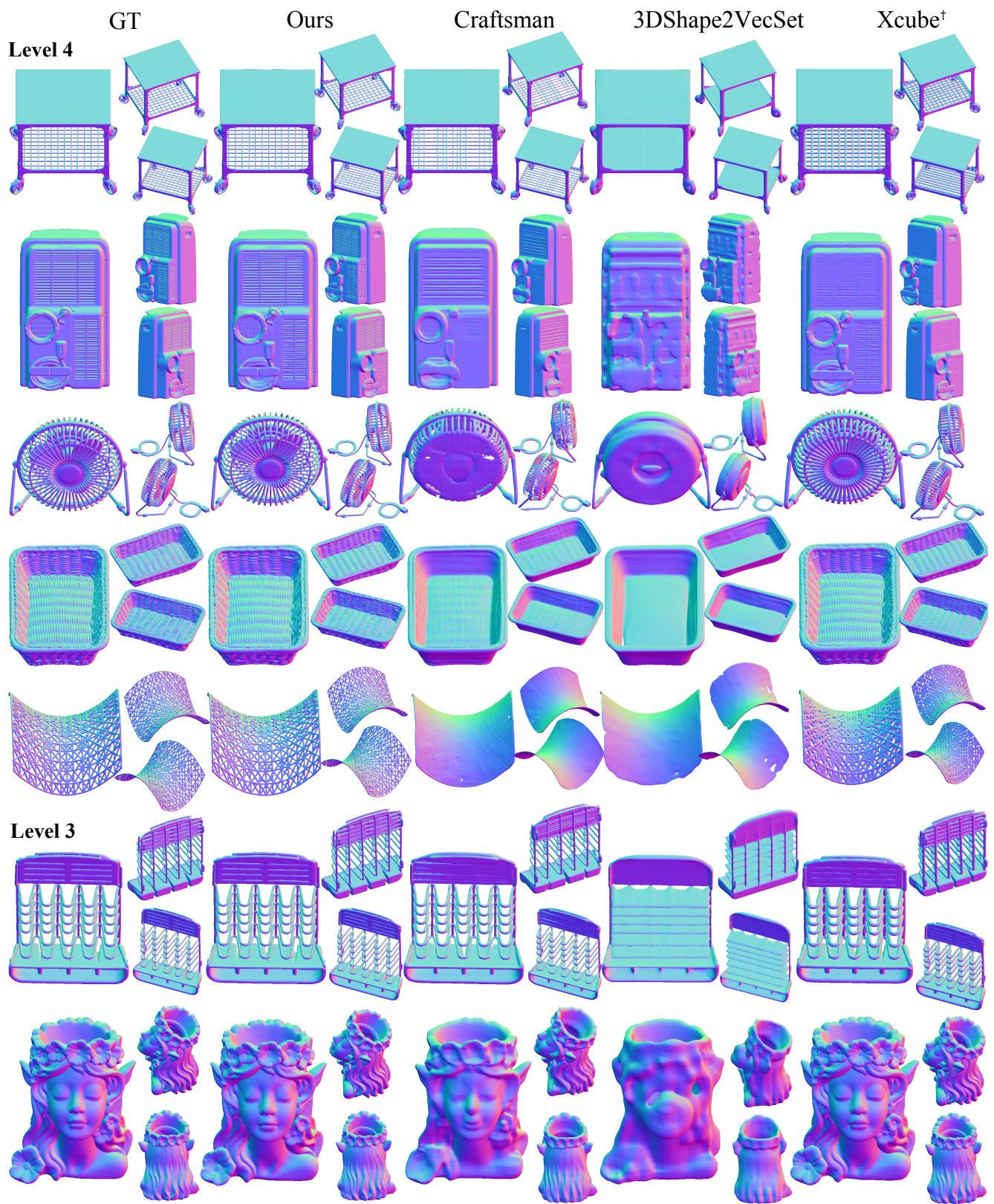


Figure S2. Qualitative comparison of the VAE reconstruction results. [†] indicates the fine-tuning model that uses the same training data as ours.

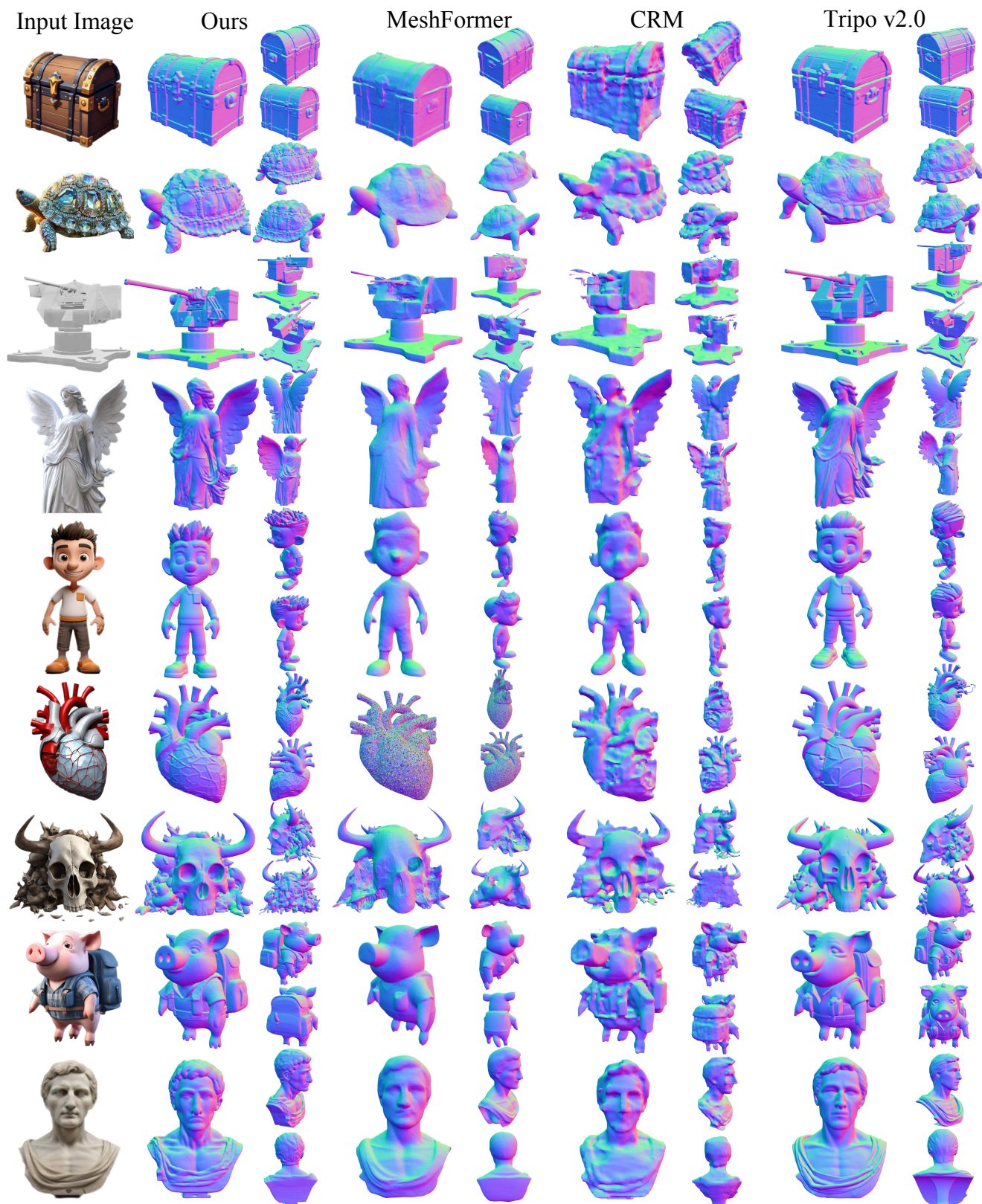


Figure S3. Qualitative comparison of the Image-to-3D results.

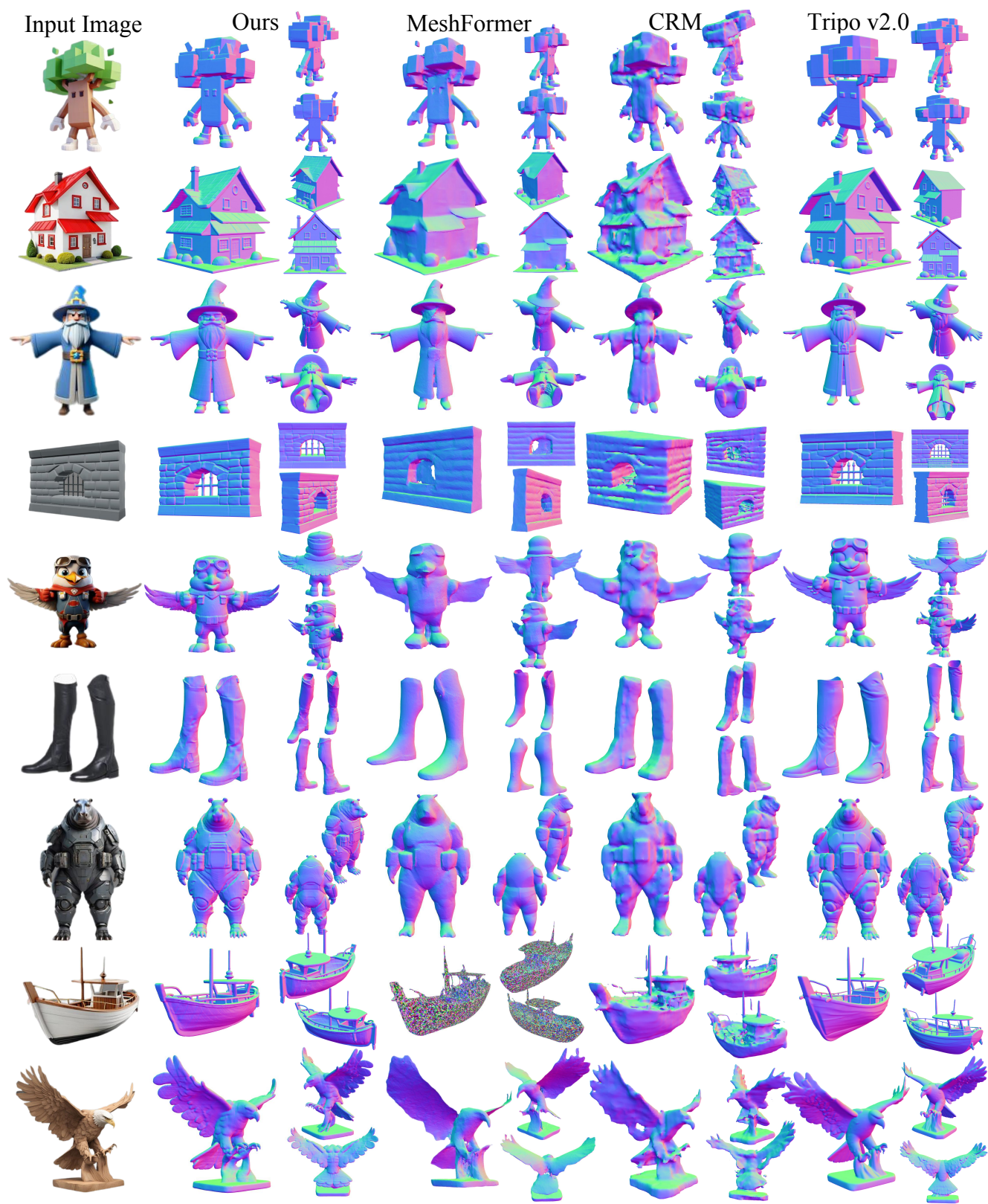


Figure S4. Qualitative comparison of the Image-to-3D results.

References

- [1] Rodin gen-1, 2024. <https://hyperhuman.deemos.com/rodin/>. 2
- [2] Tripo ai, 2024. <https://www.tripo3d.ai/>. 1
- [3] Digital twin catalog. META, 2024. <https://www.projectaria.com/datasets/dtc/>. 1
- [4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1
- [5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1
- [6] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muiyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 2
- [7] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 1
- [8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 1
- [9] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 1
- [10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 1
- [11] Clement Fuji Tsang, Maria Shugrina, Jean Francois Lafleche, Or Perel, Charles Loop, Towaki Takikawa, Vismay Modi, Alexander Zook, Jiehan Wang, Wenzheng Chen, Tianchang Shen, Jun Gao, Krishna Murthy Jatavallabhula, Edward Smith, Artem Rozantsev, Sanja Fidler, Gavriel State, Jason Gorski, Tommy Xiang, Jianing Li, Michael Li, and Rev Lebaredian. Kaolin: A pytorch library for accelerating 3d deep learning research. 1
- [12] Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, and Francis Williams. Neural kernel surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4369–4379, 2023. 1
- [13] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 1
- [14] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 1, 2
- [15] Minghua Liu, Chong Zeng, Xinyue Wei, Ruoxi Shi, Linghao Chen, Chao Xu, Mengqi Zhang, Zhaoning Wang, Xiaoshuai Zhang, Isabella Liu, Hongzhi Wu, and Hao Su. Meshformer: High-quality mesh generation with 3d-guided reconstruction model. *arXiv preprint arXiv:2408.10198*, 2024. 1
- [16] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1
- [17] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1
- [18] Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4209–4219, 2024. 1, 2
- [19] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xi-ang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 1
- [20] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv:2405.14832*, 2024. 1
- [21] Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 1, 2
- [22] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 1, 2
- [23] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang YU, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1