Supplemental Materials for SAR3D: Autoregressive 3D Object Generation and Understanding via Multi-scale 3D VQVAE

Yongwei Chen[1] Yushi Lan[1] Shangchen Zhou[1] Tengfei Wang[2] Xingang Pan[1]
[1]S-Lab, Nanyang Technological University
[2]Shanghai Artificial Intelligence Laboratory
https://cyw-3d.github.io/projects/SAR3D/

A. Effectiveness of 3D VQVAE

As shown in Fig. S1, our 3D Multi-scale Vector Quantized Variational Autoencoder (VQVAE) exhibits a strong capability in reconstructing complex 3D objects with diverse topologies. By leveraging a multiscale design, our model captures both global and local features, allowing it to reconstruct objects with textures, multiple holes, and diverse surface patterns.

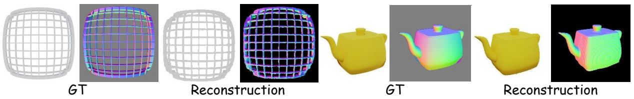


Figure S1. **Reconstruction Results of Our Multiscale VQVAE.** Our VQVAE can reconstruct both complex geometries that contain multiple holes and textures.

B. Multi-scale quantization and interpolation

Similar to VAR [10], we employ quantization and interpolation in a residual design on the latent tri-plane feature map, as described in Algorithm 1 and Algorithm 2. In particular, they demonstrate that all scales share the same codebook, and each plane of the latent tri-plane is quantized independently based on the corresponding plane's previous scales. To upsample z_k^i to the resolution of $h_K \times w_K$, we utilize convolutional layers $\phi_k^i(\cdot)$. For interpolating z_k^i to resolution $h_K \times w_K$, we don't use any network.

C. Transformer blocks

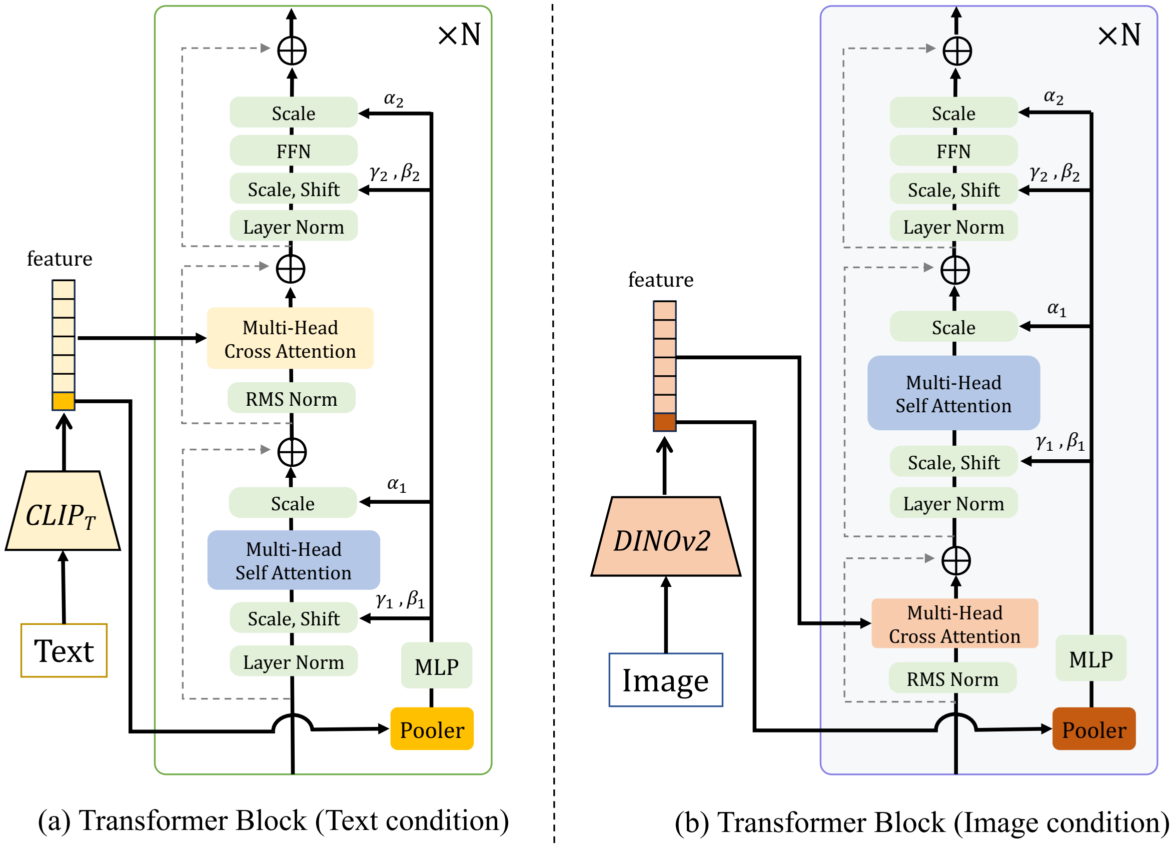
The architecture of our transformer block for 3D generation is illustrated in Fig. S2. We utilize the CLIP text encoder or the DINOv2 image encoder to process text and image embeddings, respectively. The pooled tokens are then passed through an MLP to compute the scale and shift parameters for the multi-head self-attention and feedforward network (FFN) modules. Additionally, the feature vectors are incorporated into multi-head cross-attention blocks to facilitate cross-modal attention. To enhance the integration of cross-modal information into the model, similar to [6], we modify the structure of the transformer blocks by rearranging the order of self-attention and cross-attention in the text-conditioned and image-conditioned transformer blocks.

Algorithm 1 Multi-scale 3D VQVAE Encoding

```
Require: multiview renderings $\tilde{M}$
Require: steps $K$, resolutions $(3, h_k, w_k)_{k=1}^{K}$
$f \leftarrow \mathcal{E}(\tilde{M})$, $R \leftarrow []$;
for $k = 1, \ldots, K$ do
    for $i = 1, \ldots, 3$ do
        $r_k^i \leftarrow \mathcal{Q}(\text{interpolate}(f, h_k, w_k))$
        $R \leftarrow \text{queue\_push}(R, r_k^i)$
        $z_k^i \leftarrow \text{lookup}(Z, r_k^i)$
        $z_k^i \leftarrow \text{interpolate}(z_k^i, h_K, w_K)$
        $f^i \leftarrow f^i - \phi_k^i(z_k^i)$
    end for
end for
return multi-scale latent tri-plane tokens $R$
```

Algorithm 2 Multi-scale 3D VQVAE Reconstruction

```
Require: multi-scale latent tri-plane token maps $R$
Require: steps $K$, resolutions $(3, h_k, w_k)_{k=1}^{K}$
$\hat{f} \leftarrow 0$
for $k = 1, \ldots, K$ do
    for $i = 1, \ldots, 3$ do
        $r_k^i \leftarrow \text{queue\_pop}(R)$
        $z_k^i \leftarrow \text{lookup}(Z, r_k^i)$
        $z_k^i \leftarrow \text{interpolate}(z_k^i, h_K, w_K)$
        $\hat{f}^i \leftarrow \hat{f}^i + \phi_k^i(z_k)$
    end for
end for
$\hat{T} \leftarrow \mathcal{D}(\hat{f})$
return reconstructed triplane representation $\hat{T}$
```



(a) Transformer Block (Text condition)

(b) Transformer Block (Image condition)

Figure S2. **Transformer Blocks in Our 3D Generation Transformer.** The CLIP text encoder (CLIP$_T$) or the DINOv2 image encoder processes text and image embeddings, respectively. The pooled tokens are passed through an MLP to compute the scale and shift parameters for the multi-head self-attention and feedforward network (FFN) modules. Additionally, feature vectors are incorporated into multi-head cross-attention blocks to enable cross-modal attention.

Table S1. **Statistics of Training Data.**

Splatter-Image	OpenLRM	One-2-3-45	Lara	CRM	LGM	Shap-E	LN3Diff	Ours
44K	580K	46K	240K	376K	80K	2M - 9M	170K	170K

D. More image-to-3D comparison

As illustrated in Fig. S3, we show more reults to compare our SAR3D with three categories of methods: *single-image to 3D methods* (Splatter-Image [8], OpenLRM [2, 3]), *multi-view image to 3D methods* (One-2-3-45 [7], Lara [1], CRM [11], LGM [9]), and *native 3D diffusion models* (Shap-E [4], LN3Diff-image [5]). We use pretrained models for these baseline methods, and the data statistics are provided in the Tab. S1. Compared to baseline methods, our SAR3D generates intact, distortion-free results and delivers high-quality visual effects in both reference and novel views.

E. More 3D captioning results

Additional 3D captioning results are presented in Fig. S4. Given a 3D model, our SAR3D-LLM is capable of generating detailed captions. For instance, in the case of the *skateboard ramp*, our method can describe specific details about its shape, such as *curved, flat top, sloping bottom*, as well as its functionality, like *performing tricks and jumps.*

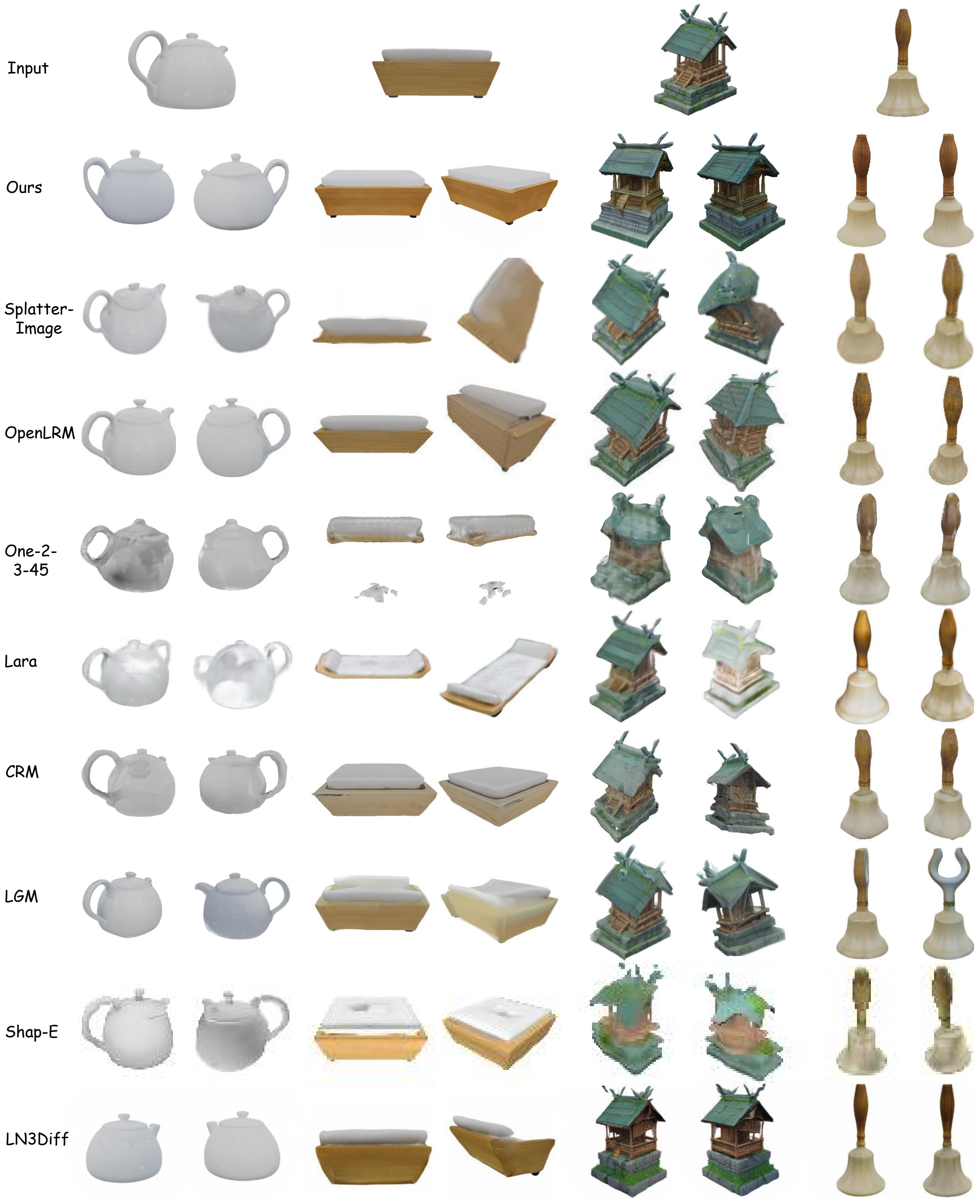
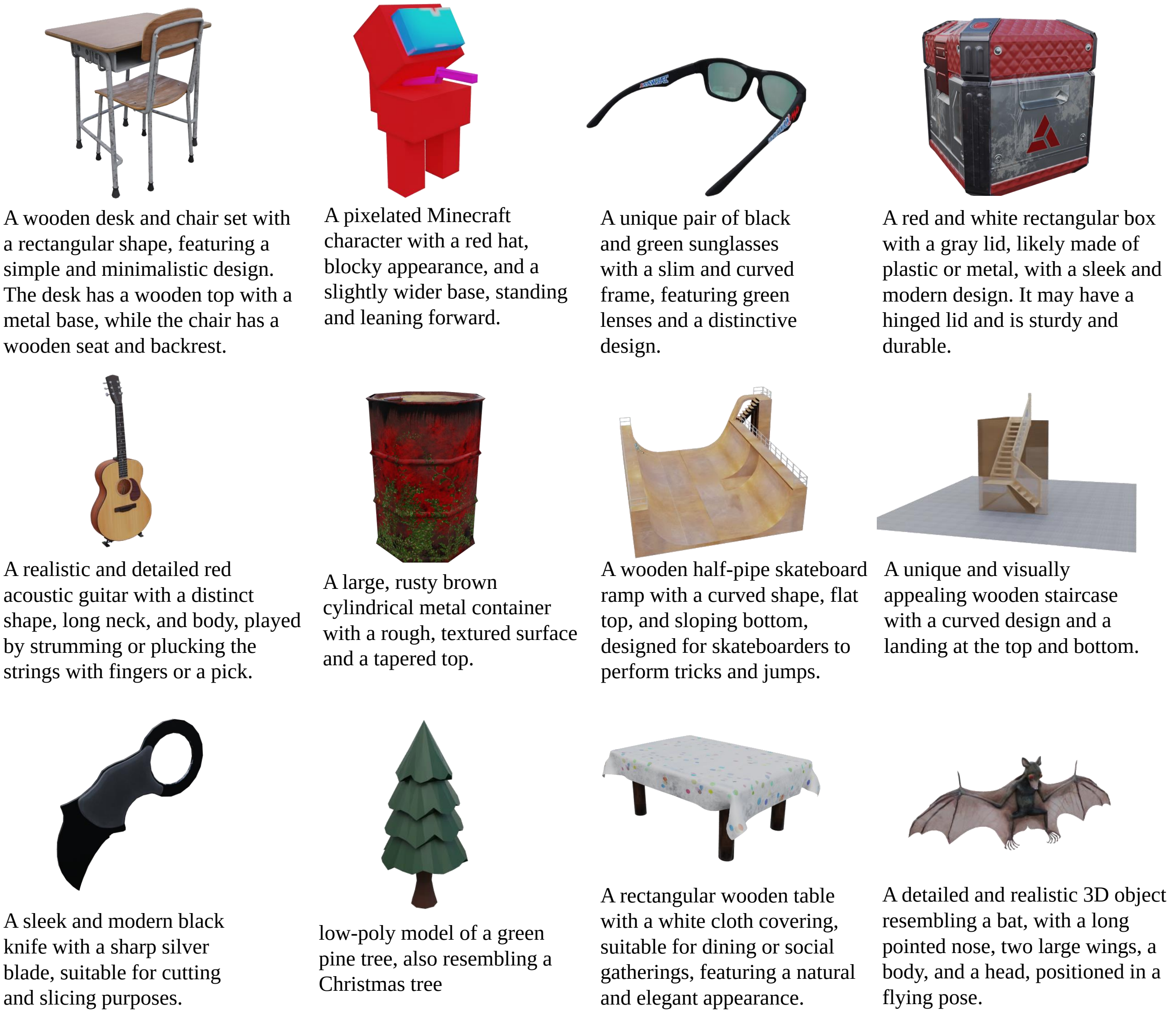


Figure S3. **More Comparisons of Image-to-3D Generation.** Our method consistently produces higher-quality 3D objects without distortion from a single image, excelling in both reference and novel views.



A wooden desk and chair set with a rectangular shape, featuring a simple and minimalistic design. The desk has a wooden top with a metal base, while the chair has a wooden seat and backrest.

A pixelated Minecraft character with a red hat, blocky appearance, and a slightly wider base, standing and leaning forward.

A unique pair of black and green sunglasses with a slim and curved frame, featuring green lenses and a distinctive design.

A red and white rectangular box with a gray lid, likely made of plastic or metal, with a sleek and modern design. It may have a hinged lid and is sturdy and durable.

A realistic and detailed red acoustic guitar with a distinct shape, long neck, and body, played by strumming or plucking the strings with fingers or a pick.

A large, rusty brown cylindrical metal container with a rough, textured surface and a tapered top.

A wooden half-pipe skateboard ramp with a curved shape, flat top, and sloping bottom, designed for skateboarders to perform tricks and jumps.

A unique and visually appealing wooden staircase with a curved design and a landing at the top and bottom.

A sleek and modern black knife with a sharp silver blade, suitable for cutting and slicing purposes.

low-poly model of a green pine tree, also resembling a Christmas tree

A rectangular wooden table with a white cloth covering, suitable for dining or social gatherings, featuring a natural and elegant appearance.

A detailed and realistic 3D object resembling a bat, with a long pointed nose, two large wings, a body, and a head, positioned in a flying pose.

Figure S4. **Additional 3D Captioning Results.** Our method generates detailed descriptions based on the input of 8 scales of latent tri-plane tokens.

References

[1] Anpei Chen, Haofei Xu, Stefano Esposito, Siyu Tang, and Andreas Geiger. Lara: Efficient large-baseline radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 338–355. Springer, 2024. 2

[2] Zexin He and Tengfei Wang. OpenLRM: Open-source large reconstruction models. `https://github.com/3DTopia/OpenLRM`, 2023. 2

[3] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024. 2

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

[5] Yushi Lan, Fangzhou Hong, Shuai Yang, Shangchen Zhou, Xuyi Meng, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff: Scalable latent neural fields diffusion for speedy 3d generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 112–130. Springer, 2024. 2

[6] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[7] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund

Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:22226–22246, 2023. 2

[8] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10208–10217, 2024. 2

[9] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–18. Springer, 2024. 2

[10] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:84839–84865, 2024. 1

[11] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 57–74. Springer, 2024. 2