Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details

Tencent Hunyuan3D

https://3d.hunyuan.tencent.com
https://github.com/Tencent/Hunyuan3D-2

Abstract

In this report, we present Hunyuan3D 2.5, a robust suite of 3D diffusion models aimed at generating high-fidelity and detailed textured 3D assets. Hunyuan3D 2.5 follows two-stages pipeline of its previous version Hunyuan3D 2.0, while demonstrating substantial advancements in both shape and texture generation. In terms of shape generation, we introduce a new shape foundation model – LATTICE, which is trained with scaled high-quality datasets, model-size, and compute. Our largest model reaches 10B parameters and generates sharp and detailed 3D shape with precise image-3D following while keeping mesh surface clean and smooth, significantly closing the gap between generated and handcrafted 3D shapes. In terms of texture generation, it is upgraded with phyiscal-based rendering (PBR) via a novel multi-view architecture extended from Hunyuan3D 2.0 Paint model. Our extensive evaluation shows that Hunyuan3D 2.5 significantly outperforms previous methods in both shape and end-to-end texture generation.

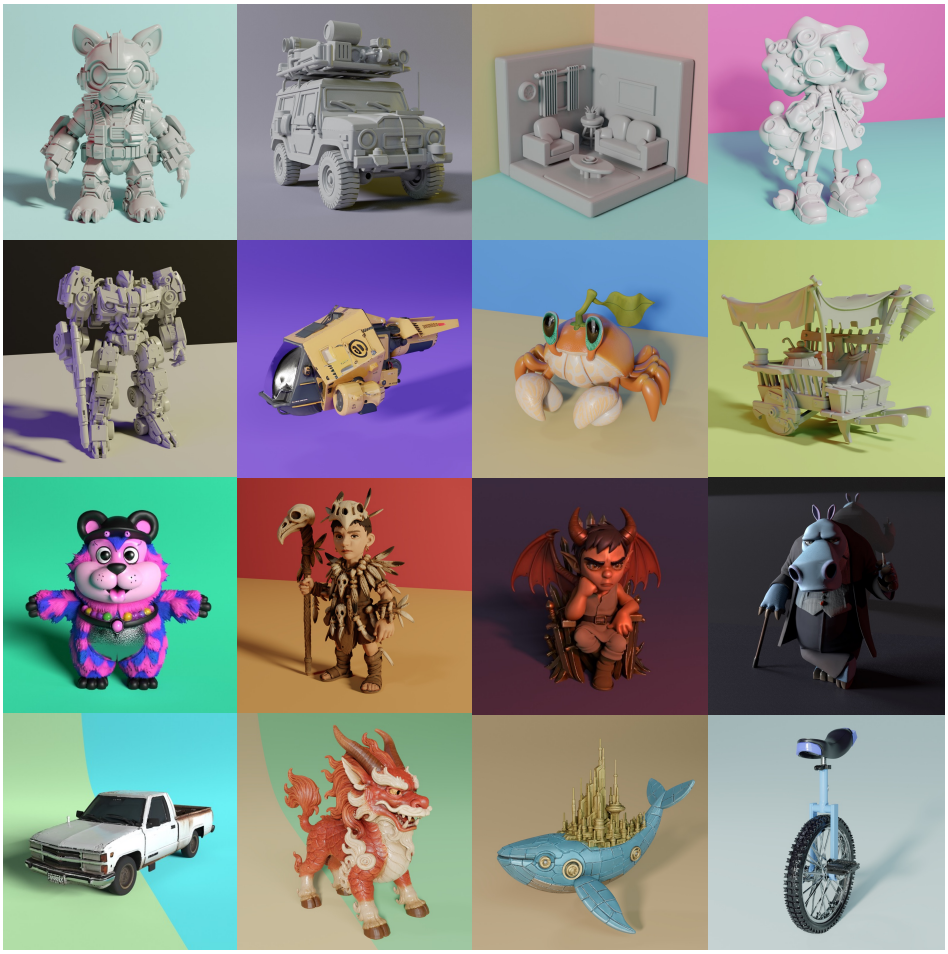

Figure 1: High quality 3D assets generated by Hunyuan3D 2.5.

1 Introduction

3D generation has rapidly developed in recent years, becoming a core driver of innovation and growth across various industries. From game development to embodied AI, from film special effects to virtual reality, the application scenarios of 3D models continue to expand, demonstrating their immense potential and value. With advancements in artificial intelligence, 3D generation has become more efficient and powerful, particularly in areas such as automated modeling and texturing, further simplifying the creative process and enhancing production efficiency.

Notably, recent 3D shape diffusion models based on 3dshape2vecset (Zhang et al., 2023) have pioneered a revolution in the 3D shape generation pipeline, as demonstrated by works such as CLAY (Zhang et al., 2024b), Hunyuan3D 2.0 (Zhao et al., 2025), and TripoSG (Li et al., 2025). Direct3D (Wu et al., 2024b),from another aspect, shown the potential of compressing and generating the shape via triplane. More recently, Trellis (Xiang et al., 2024) has emerged as a promising pipeline for high-quality textured 3D generation, leveraging its invented structured 3D latents as representations. Nevertheless, existing models are still limited for generating complex objects with finegrained details as demonstrated in figure 2. It remains an open problem how we could generate high-fidelity and detailed shape while maintaining smooth surface and sharp edges.

High-quality textures play a crucial role in enhancing the visual realism and detail representation of 3D assets. Recently, Numerous texture generation methods based on multi-view diffusion (Zhao et al., 2025; Huang et al., 2024b; Vainer et al., 2024b; Li et al., 2024a; Tang et al., 2025; Long et al., 2024; Wang & Shi, 2023; Shi et al., 2023b;a) have emerged, alleviating global consistency issue of inpainting-based methods (Huang et al., 2024a; Wu et al., 2024a; Zhang et al., 2024c; Ceylan et al., 2024; Zeng et al., 2024a; Chen et al., 2023a; Richardson et al., 2023) and synchronization techniques (Liu et al., 2025; Gao et al., 2024; Liu et al., 2024a; Zhang et al., 2024a). However, challenges remain in generating highly consistent multiview images, which can lead to artifacts and seams during the fusion and baking stages. Moreover, traditional RGB textures can no longer meet the demands for photorealistic 3D asset generation, while PBR material generation solution is not available in open source community.

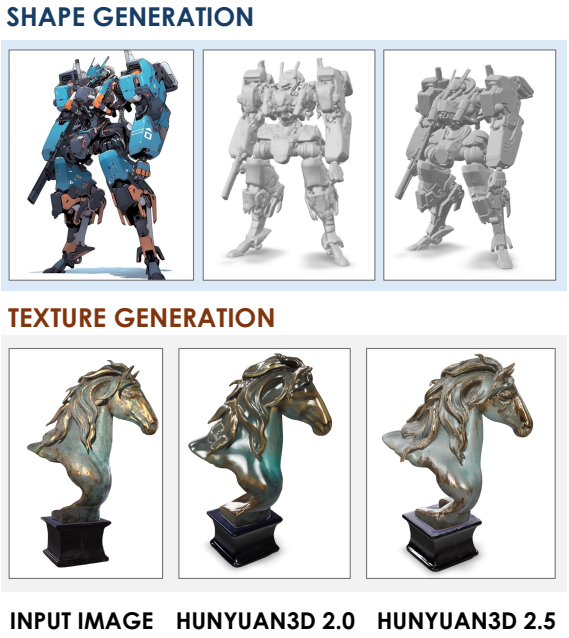

Figure 2: Drawbacks of existing methods: failure at detail generation and incorrect PBR.

This report presents **Hunyuan3D 2.5**, a robust suite of 3D diffusion models aimed at generating high-fidelity and detailed textured 3D assets. Hunyuan3D 2.5 builds upon the two-stage pipeline of its predecessor Hunyuan3D 2.0 (Zhao et al., 2025) and 2.1 (Hunyuan3D et al., 2025), while showcasing significant advancements in both shape generation and texture synthesis.

In the first stage of shape generation, we introduce a new shape foundation model – LATTICE, which has been trained on large-scale, high-quality datasets with increased model size and computational resources. We found that this new model exhibits stable improvement when scaling up the model. Benifit from these characteristics, our largest model generates detailed and sharp 3D shape with precise alignment to corresponding images while maintaining clean and smooth surfaces, significantly closing the gap between generated and handcrafted 3D shapes.

In the second stage of texture generation, we extend the Hunyuan3D 2.0 (Zhao et al., 2025) and 2.1 (Hunyuan3D et al., 2025) texture generation model into a high-fidelity material generation framework. Adhering to the principled BRDF model, our approach produces multi-view albedo, roughness, and metallic maps simultaneously. This aims to precisely describe the surface reflection properties of generated 3D assets and accurately simulate geometric microsurface distributions,

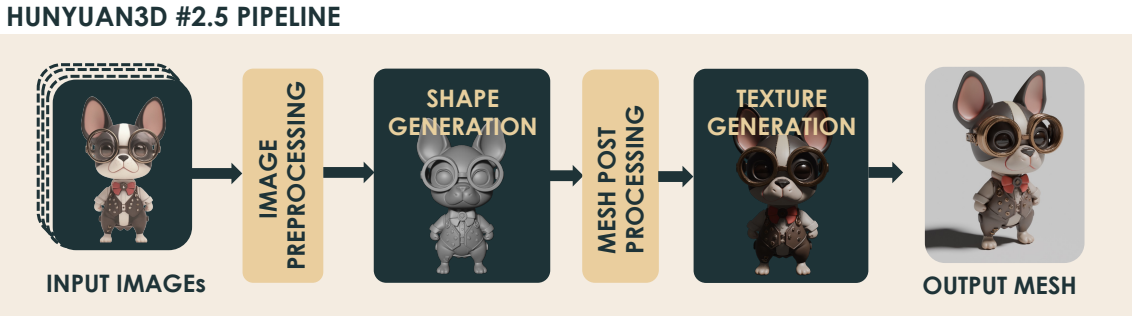

Figure 3: **Overview of Hunyuan3D 2.5 pipeline**. It separates the 3D asset generation into two stages: first, it generates the shape, and then it creates the texture based on that shape.

thereby achieving more realistic and detailed rendering results. Furthermore, we introduce a dual-phase resolution enhancement strategy to strengthen texture-geometry coordination, thereby improving end-to-end visual quality.

We perform extensive quantitative and qualitative evaluations as well as user studies in terms of shape generation and end-to-end texture generation, across diverse range of in-the-wild input images. The results demonstrate that Hunyuan3D 2.5 outperforms state-of-the-art open-source and closed-source commercial models.

2 Method

Hunyuan3D 2.5 is an image-to-3D generation model, which follows the same overall architecture of Hunyuan3D 2.0 (Zhao et al., 2025), as shown in figure 3. In a nutshell, the input image is first processed by an image processor to remove the background and perform proper resizing. Then, a shape generation model is conditioned on the input image and generates the 3D mesh without texture. The mesh is further processed to extract normal, UV map, and *etc*. After that, a texture generation model is called to generate the texture with previous outputs.

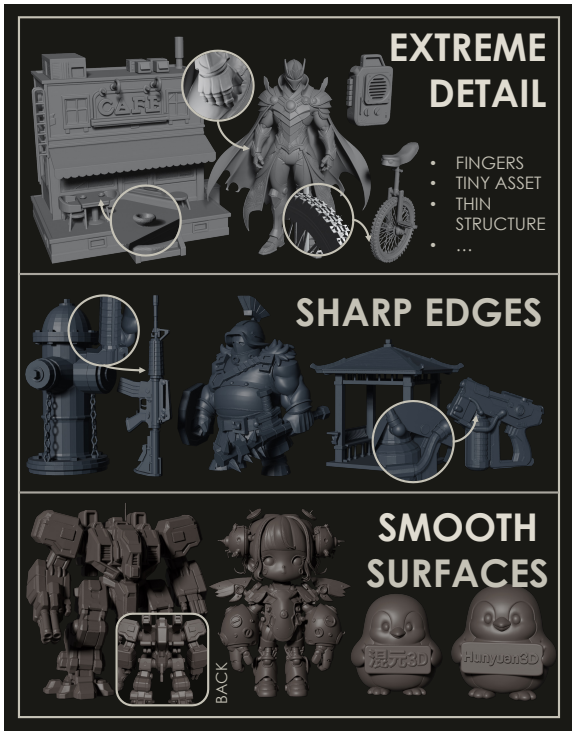

Figure 4: Illustration of major features of the new shape generation model in Hunyuan3D 2.5.

2.1 Detailed Shape Generation

Hunyuan3D 2.5 introduces a new shape generation model – LATTICE, which is a large-scale diffusion model capable of producing high-fidelity, detailed shapes with sharp edges and smooth surfaces from either a single image or four multi-view images. Trained on an extensive and high-quality 3D dataset featuring complex objects, the model is designed to generate exceptional detail. To ensure efficiency, we employ guidance and step distillation techniques to reduce inference time.

Extreme Detail. Benfit from scaling up, Hunyuan3D 2.5 can generate fine-grained details at an unprecedented level. In the first row of figure 4, we present several examples generated by our model. As shown, the model achieves a level of accuracy approaching that of handcrafted designs, such as the correct number of fingers, the detailed bicycle wheel pattern, and even a bowl within a large scene.

Smooth Surfaces & Shape Edges. Existing models (Zhao et al., 2025; Li et al., 2025; Xiang et al.,

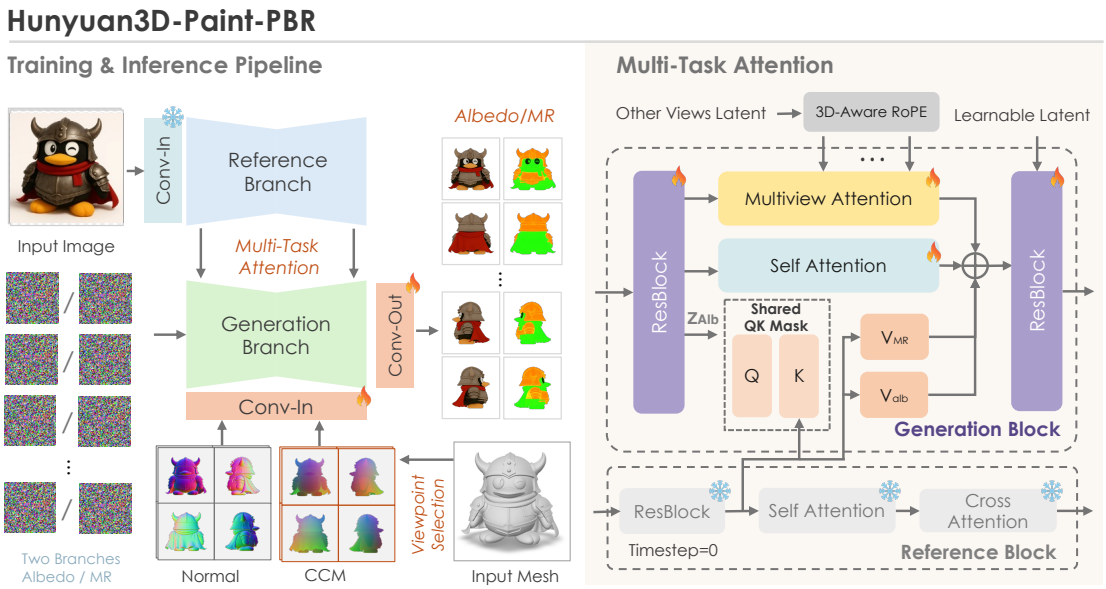

Figure 5: Overview of material generation framework.

2024) often struggle to generate sharp edges while maintaining smooth, clean surfaces, particularly for complex objects. In contrast, Hunyuan3D 2.5 strikes an excellent balance, as demonstrated in the second and third rows of figure 4.

2.2 Realistic Texture Generation

We propose a novel material generation framework in Hunyuan3D 2.5, which is extended on the foundation of multiview PBR texture generation architecture of hunyuan3D 2.1 (Hunyuan3D et al., 2025). As shown in figure 5, our model takes normal map and CCM rendered by 3D mesh as geometry conditions, and a reference image as guidance, generating high-quality PBR material maps as texture. We inherit 3D-aware RoPE in (Feng et al., 2025) to enhance cross-view consistency for seamless texture map generation.

Multi-Channel Material Generation. We introduce learnable embeddings for three material maps: albedo, MR, of which the MR channel is the combation expression of metallic and roughness. Specifically, three independent embeddings $\mathbf{E}_{\text{albedo}}$, $\mathbf{E}_{\text{mr}}$, and $\mathbf{E}_{\text{normal}} \in \mathbb{R}^{16\times1024}$ are initialized and subsequently injected into the respective channels via cross-attention layers. The embedding and attention modules are trainable, allowing the network to effectively model the distribution of three materials separately.

Although material channels exhibit significant domain gaps, maintaining spatial correspondence is crucial across different levels, from semantic to pixel-level alignment. To address this challenge, we propose a dual-channel attention mechanism that ensures spatial alignment among generated albedo and metallic-roughness (MR).

After systematically examining the reference attention module, we found that the main cause of multi-channel misalignment lies in the misaligned attention masks. Therefore, we intentionally share the attention mask among multiple channels while varying the value computation in the output calculation. Specifically, since the basecolor branch contains the most semantically similar information to the reference image (both exist in the common RGB color space), we utilize the attention mask calculated from the basecolor channel and apply it to guide the reference attention in the other two branches, as formulated below:

$$\mathrm{M}_{attn} = \text{Softmax}\left(\frac{Q_{albedo}K_{ref}^{T}}{\sqrt{d}}\right) \tag{1}$$

$$z_{albedo}^{new} = z_{albedo} + \mathrm{MLP}_{albedo}\left[\mathrm{M}_{attn} \cdot V_{albedo}\right], \quad z_{MR}^{new} = z_{MR} + \mathrm{MLP}_{MR}\left[\mathrm{M}_{attn} \cdot V_{MR}\right] \tag{2}$$

This design enables the generated albedo and MR features to maintain spatial coherence while being guided by the reference image's information. Building upon this framework, we incorporated an illumination-invariant consistency loss during training to enforce the disentanglement of material properties and illumination components. For more information, please refer to (He et al., 2025).

Geometric Alignment. The alignment of textures with geometry critically impacts the visual integrity and aesthetic quality of 3D assets. However, achieving precise texture-geometry alignment presents considerable challenges, particularly for complex, high-polygon geometry. A key observation from our analysis is that higher-resolution images preserve richer high-frequency geometric details while mitigating VAE compression losses, thereby significantly enhancing geometric conditioning. Nevertheless, training with high-resolution multi-view images demands substantial memory resources, which necessitates reducing the number of views during training and consequently deteriorates the model's capability for dense-view inference.

To address this challenge, we propose a dual-phase resolution enhancement strategy that progressively improves texture-geometry alignment quality while maintaining computational feasibility. In the first phase, we employ a conventional multi-view training approach using 6-view 512×512 images, following the methodology of Hunyuan3D-2.0 Zhao et al. (2025). This phase establishes a solid foundation for multi-view consistency and basic texture-geometry correspondence.

In the second phase, we implement a zoom-in training strategy that enables the model to capture high-quality details while preserving the multi-view training benefits from the first phase. Specifically, we randomly zoom into both the reference image and multi-view generated images during training. This approach allows the model to learn fine-grained texture details without requiring full high-resolution training from scratch, thereby circumventing the memory constraints associated with direct high-resolution multi-view training.

During inference, we leverage multi-view images at up to 768×768 resolution, accelerated by the UniPC sampler (Zhao et al., 2023) for efficient high-quality generation.

3 Evaluation

To comprehensively assess the performance of Hunyuan3D 2.5, we performed evaluations from two key perspectives: (1) 3D shape generation, and (2) textured 3D asset generation.

3.1 Shape Generation

Competing Methods. We compare with open-source baselines, Michelangelo (Zhao et al., 2024), Craftsman 1.5 (Li et al., 2024b), Trellis (Xiang et al., 2024), and Hunyuan3D-2 (Zhao et al., 2025), and closed-source baselines are Commerical Model 1, and Commerical Model 2.

Metrics. To assess the performance of shape generation, we utilize ULIP (Xue et al., 2023) and Uni3D (Zhou et al., 2023) to calculate the similarity between the generated mesh and the input images (ULIP-I and Uni3D-I) as well as image prompts synthesized by the vision-language model (Chen et al., 2024b) (ULIP-T and Uni3D-T).

Comparison. We show the numerical comparison in table.1 and visual comparison in figure 6. It can be observed that our method achieves the best image-shape and text-shape similarities in terms of ULIP-T and Uni3D-T and Uni3D-I. Nevertheless, we noted that these metrics could not fully reflect the model capabilities. As shown in figure 6, our model actually perform much better than all other open-sourced and commerical models.

Figure 6: Visual comparison of different methods in terms of shape generation.

Figure 7: Visual comparison of different methods in terms of texture generation. We compared the front and back of models generated by different methods, as well as the effects of the corresponding complete material maps and albedo maps.

3.2 Texture Generation

Competing Methods. We perform quantitive comparison with text- and image-conditioned methods, including Text2Tex Chen et al. (2023a), Paint3D Zeng et al. (2024a), Paint-it Youwang et al.

Table 1: Numerical comparisons of different shape generation models on ULIP-T/I, Uni3D-T/I.

	ULIP-T(↑)	ULIP-I(↑)	Uni3D-T(↑)	Uni3D-I(↑)
Michelangelo (Zhao et al., 2024)	0.0752	0.1152	0.2133	0.2611
Craftsman 1.5 (Li et al., 2024b)	0.0745	0.1296	0.2375	0.2987
Trellis (Xiang et al., 2024)	0.0769	0.1267	0.2496	0.3116
Commercial Model 1	0.0741	**0.1308**	0.2464	0.3106
Commercial Model 2	0.0746	0.1284	0.2516	0.3131
Hunyuan3D 2.0 (Zhao et al., 2025)	0.0771	0.1303	0.2519	0.3151
Hunyuan3D 2.5	**0.07853**	0.1306	**0.2542**	**0.3151**

Table 2: Quantitative comparison with state-of-the-art methods. We compare with two classes of methods, one conditioned on text only, and the other one based on image. Our method achieves the best performance compared with both classes.

Method	CLIP-FID↓	FID↓	CMMD↓	CLIP-I↑	LPIPS↓
Text2Tex Chen et al. (2023a) ICCV'23	31.83	187.7	2.738	-	0.1448
SyncMVD Liu et al. (2024a) SIGGRAPH Asia'24	29.93	189.2	2.584	-	0.1411
Paint-it Youwang et al. (2024) CVPR'24	33.54	179.1	2.629	-	0.1538
Paint3D Zeng et al. (2024a) CVPR'24	26.86	176.9	2.400	0.8871	0.1261
TexGen Yu et al. (2024) TOG'24	28.23	178.6	2.447	0.8818	0.1331
Ours	**23.97**	**165.8**	**2.064**	**0.9281**	**0.1231**

(2024), SyncMVD Liu et al. (2024a), and TexGen Yu et al. (2024). Furthermore, we show qualitive comparison on closed-source commercial models.

Metrics. We use Fréchet Inception Distance (FID), CLIP-based FID (CLIP-FID), and Learned Perceptual Image Patch Similarity (LPIPS) to measure the similarity between the generated textures and the ground truth. CLIP Maximum-Mean Discrepancy (CMMD) is used to assess the diversity and richness of the generated texture details. And CLIP-Image Similarity (CLIP-I) is employed to evaluate how well the generated textures semantically align with the input images (for image prompt methods).

Comparison. We show the numerical comparison in table.2 and visual comparison in figure 7, For an intuitive comparison, we directly show the end-to-end results. It can be observed that for PBR material generation, competing models struggle to accurately estimate the correct MR (metallic and roughness) values, and face challenges in decoupling the inherent illumination effects in the input images for the albedo component.

User Study. We also conducted a user study to evaluate human preferences for generated textured models using different methods. In this study, each participant was asked to rank each method for each sample in the testset. The testset included a diverse range of real-world images from various categories. As shown in figure 8, we compared our method with three different commercial models. The results clearly demonstrate that our method significantly outperforms the others. For instance, in the image-to-3D task, our method achieved a 72% win rate, which is 9 times higher than that of Commercial Model 1.

4 Related Work

Shape Generation. 3D shape generation has advanced rapidly in recent years. Early works (Wu et al., 2016; Sanghi et al., 2022; Yan et al., 2022; Yin et al., 2023) based on different generative models (Kingma, 2013; Goodfellow et al., 2014; Papamakarios et al., 2021) demonstrated the preliminary potential for generating specific categories of shapes. With the rise of diffusion models (Rombach et al., 2022; Ho et al., 2020), 3D shape generation methods based on score distillation (Poole et al., 2023) have been introduced, enabling text-to-3D generation by leveraging text-to-image models. Feedforward methods such as LRM (Hong et al., 2023), Hunyuan3D 1.0 (Yang et al., 2024) and LGM (Tang et al., 2024) represent another line of research focused on generating 3D assets in a single step. Recently, native 3D diffusion models have significantly improved generation quality by utilizing 3D data. Notable works in this area include Michelangelo (Zhao et al., 2024), CLAY (Zhang et al., 2024b), Trellis (Xiang et al., 2024), Hunyuan3D 2.0 Zhao et al. (2025), and

TripoSG (Li et al., 2025), among others. Although multi-step sampling is required, native 3D diffusion models based on vecset Zhang et al. (2023) can be accelerated via FlashVDM (Lai et al., 2025), achieving speeds that surpass even feedforward methods. On the other hand, autoregressive models, such as MeshGPT (Siddiqui et al., 2024) BPT (Weng et al., 2024), and Meshtron (Hao et al., 2024) have become popular for mesh generation with human-like topology.

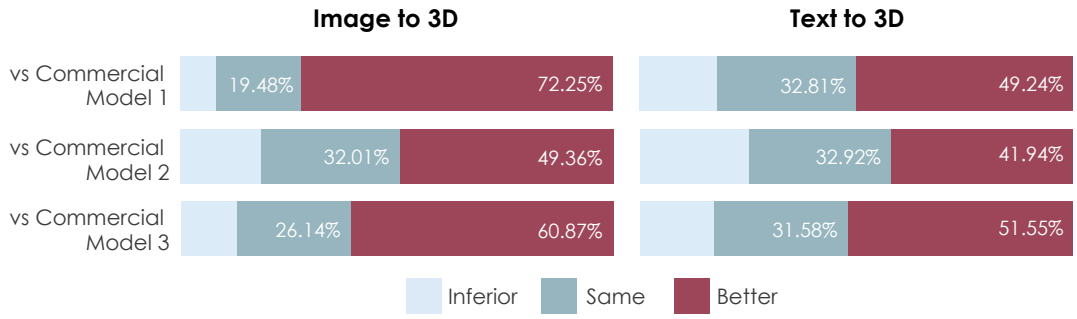

Figure 8: User study against three latest commerical models in terms of end-to-end textured results.

Texture Synthesis. Multiview diffusion (Zhao et al., 2025; Huang et al., 2024b; Vainer et al., 2024b; Li et al., 2024a; Tang et al., 2025; Long et al., 2024; Wang & Shi, 2023; Shi et al., 2023b;a; Liu et al., 2023), which primarily introduce cross-view attention mechanisms to modeling multi-view latent interactions, has opened up new avenues for addressing the global consistency issue of 3D textures. Zero123++ (Shi et al., 2023a) spatially concatenates multi-view images and utilizes the self-attention to build up cross-view interaction. Other works inject view constraints into the attention block using diverse attention masks (Tang et al., 2023; Huang et al., 2024b; Li et al., 2024a). For PBR material generation, existing methods mainly include three categories: Generation-based approaches Vainer et al. (2024a); Sartor & Peers (2023); Vecchio et al. (2024); Chen et al. (2024a); Zeng et al. (2024b) leverage diffusion models to learn material priors and recover PBR properties through physical rendering; retrieval-based techniques Zhang et al. (2024c); Fang et al. (2024) adapt pre-built material graphs from libraries to ensure visual consistency and editability; optimization-based methods Chen et al. (2023b); Zhang et al. (2024d); Wu et al. (2023); Xu et al. (2023); Yeh et al. (2024); Youwang et al. (2024); Liu et al. (2024b) first generate initial textures and then refine them through techniques like Score-Distillation Sampling Poole et al. (2023).

5 Conclusion

In this work, we presented Hunyuan3D 2.5, an advanced suite of 3D diffusion models for generating high-quality, detailed 3D assets. By introducing a new shape foundation model and extending texture generation with physical-based rendering (PBR), Hunyuan3D 2.5 achieves remarkable improvements in both shape fidelity and texture realism. Extensive evaluations show that Hunyuan3D 2.5 outperforms current state-of-the-art models in terms of shape detail, surface smoothness, and texture consistency. This work marks a significant advancement in the 3D generation field, providing a powerful tool for creating realistic and detailed 3D assets across various industries.

6 Contributors

Project Sponsors:

Jie Jiang, Linus

Project Leaders:

Chunchao Guo, Jingwei Huang, Zeqiang Lai

Core Contributors:

- *Shape Generation*: Zeqiang Lai, Yunfei Zhao, Jingwei Huang, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang
- *Texture Generation*: Mingxin Yang, Shuhui Yang, Yifei Feng, Sheng Zhang, Xin Huang

Contributors[1]**:**

Di Luo, Fan Yang, Fang Yang, Lifu Wang, Sicong Liu, Yixuan Tang, Yulin Cai, Zebin He, Tian Liu, Yuhong Liu

References

Duygu Ceylan, Valentin Deschaintre, Thibault Groueix, Rosalie Martin, Chun-Hao Huang, Romain Rouffet, Vladimir Kim, and Gaëtan Lassagne. Matatlas: Text-driven consistent geometry texturing and material assignment. *arXiv preprint arXiv:2404.02899*, 2024.

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18558–18568, 2023a.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *ICCV*, 2023b.

Xi Chen, Sida Peng, Dongchen Yang, Yuan Liu, Bowen Pan, Chengfei Lv, and Xiaowei Zhou. Intrinsicanything: Learning diffusion priors for inverse rendering under unknown illumination. In *Proceedings of European Conference on Computer Vision*, pp. 450–467, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Ye Fang, Zeyi Sun, Tong Wu, Jiaqi Wang, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. Make-it-real: Unleashing large multimodal model's ability for painting 3d objects with realistic materials. *arXiv preprint arXiv:2404.16829*, 2024.

Yifei Feng, Mingxin Yang, Shuhui Yang, Sheng Zhang, Jiaao Yu, Zibo Zhao, Yuhong Liu, Jie Jiang, and Chunchao Guo. Romantex: Decoupling 3d-aware rotary positional embedded multi-attention network for texture synthesis, 2025. URL https://arxiv.org/abs/2503.19011.

Chenjian Gao, Boyan Jiang, Xinghui Li, Yingpeng Zhang, and Qian Yu. Genesistex: adapting image denoising diffusion to texture space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4620–4629, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Zekun Hao, David W Romero, Tsung-Yi Lin, and Ming-Yu Liu. Meshtron: High-fidelity, artist-like 3d mesh generation at scale. *arXiv preprint arXiv:2412.09548*, 2024.

[1]Alphabetical order.

Zebin He, Mingxin Yang, Shuhui Yang, Yixuan Tang, Tao Wang, Kaihao Zhang, Guanying Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, and Wenhan Luo. Materialmvp: Illumination-invariant material generation via multi-view pbr diffusion, 2025. URL https://arxiv.org/abs/2503.10289.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

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. *arXiv preprint arXiv:2311.04400*, 2023.

Xin Huang, Tengfei Wang, Ziwei Liu, and Qing Wang. Material anything: Generating materials for any 3d object via diffusion. *arXiv preprint arXiv:2411.15138*, 2024a.

Zehuan Huang, Yuanchen Guo, Haoran Wang, Ran Yi, Lizhuang Ma, Yan-Pei Cao, and Lu Sheng. Mv-adapter: Multi-view consistent image generation made easy. *arXiv preprint arXiv:2412.03632*, 2024b.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, Qingxiang Lin, Zeqiang Lai, Xianghui Yang, Huiwen Shi, Zibo Zhao, Bowen Zhang, Hongyu Yan, Lifu Wang, Sicong Liu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Dongyuan Guo, Junlin Yu, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Shida Wei, Chao Zhang, Yonghao Tan, Yifu Sun, Lin Niu, Shirui Huang, Bojian Zheng, Shu Liu, Shilin Chen, Xiang Yuan, Xiaofeng Yang, Kai Liu, Jianchen Zhu, Peng Chen, Tian Liu, Di Wang, Yuhong Liu, Linus, Jie Jiang, Jingwei Huang, and Chunchao Guo. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025. URL https://arxiv.org/abs/2506.15442.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Haolin Liu, Fuyun Wang, Huiwen Shi, Xianghui Yang, Qinxiang Lin, Jinwei Huang, Yuhong Liu, Jie Jiang, Chunchao Guo, and Xiangyu Yue. Unleashing vecset diffusion model for fast shape generation, 2025. URL https://arxiv.org/abs/2503.16302.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024a.

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner, 2024b.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023.

Shang Liu, Chaohui Yu, Chenjie Cao, Wen Qian, and Fan Wang. Vcd-texture: Variance alignment based 3d-2d co-denoising for text-guided texturing. In *European Conference on Computer Vision*, pp. 373–389. Springer, 2025.

Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Zexiang Liu, Yangguang Li, Youtian Lin, Xin Yu, Sida Peng, Yan-Pei Cao, Xiaojuan Qi, Xiaoshui Huang, Ding Liang, and Wanli Ouyang. Unidream: Unifying diffusion priors for relightable text-to-3d generation. In *Proceedings of European Conference on Computer Vision*, pp. 74–91, 2024b.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9970–9980, 2024.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18603–18613, 2022.

Sam Sartor and Pieter Peers. Matfusion: a generative diffusion model for svbrdf capture. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10, 2023.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023a.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2023b.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. Meshgpt: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19615–19625, 2024.

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 *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023.

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction. In *European Conference on Computer Vision*, pp. 175–191. Springer, 2025.

Shimon Vainer, Mark Boss, Mathias Parger, Konstantin Kutsy, Dante De Nigris, Ciara Rowles, Nicolas Perony, and Simon Donné. Collaborative control for geometry-conditioned pbr image generation. In *Proceedings of European Conference on Computer Vision*, pp. 127–145, 2024a.

Shimon Vainer, Konstantin Kutsy, Dante De Nigris, Ciara Rowles, Slava Elizarov, and Simon Donné. Jointly generating multi-view consistent pbr textures using collaborative control, 2024b. URL https://arxiv.org/abs/2410.06985.

Giuseppe Vecchio, Renato Sortino, Simone Palazzo, and Concetto Spampinato. Matfuse: controllable material generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4429–4438, 2024.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Haohan Weng, Zibo Zhao, Biwen Lei, Xianghui Yang, Jian Liu, Zeqiang Lai, Zhuo Chen, Yuhong Liu, Jie Jiang, Chunchao Guo, et al. Scaling mesh generation via compressive tokenization. *arXiv preprint arXiv:2411.07025*, 2024.

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016.

Jinbo Wu, Xing Liu, Chenming Wu, Xiaobo Gao, Jialun Liu, Xinqi Liu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. Texro: generating delicate textures of 3d models by recursive optimization. *arXiv preprint arXiv:2403.15009*, 2024a.

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 preprint arXiv:2405.14832*, 2024b.

Tong Wu, Zhibing Li, Shuai Yang, Pan Zhang, Xingang Pan, Jiaqi Wang, Dahua Lin, and Ziwei Liu. Hyperdreamer: Hyper-realistic 3d content generation and editing from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10, 2023.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

Xudong Xu, Zhaoyang Lyu, Xingang Pan, and Bo Dai. Matlaber: Material-aware text-to-3d via latent brdf auto-encoder. *arXiv preprint arXiv:2308.09278*, 2023.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6239–6249, 2022.

Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d-1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024.

Yu-Ying Yeh, Jia-Bin Huang, Changil Kim, Lei Xiao, Thu Nguyen-Phuoc, Numair Khan, Cheng Zhang, Manmohan Chandraker, Carl S Marshall, Zhao Dong, et al. Texturedreamer: Image-guided texture synthesis through geometry-aware diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4304–4314, 2024.

Fukun Yin, Xin Chen, Chi Zhang, Biao Jiang, Zibo Zhao, Jiayuan Fan, Gang Yu, Taihao Li, and Tao Chen. Shapegpt: 3d shape generation with a unified multi-modal language model. *arXiv preprint arXiv:2311.17618*, 2023.

Kim Youwang, Tae-Hyun Oh, and Gerard Pons-Moll. Paint-it: Text-to-texture synthesis via deep convolutional texture map optimization and physically-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4347–4356, 2024.

Xin Yu, Ze Yuan, Yuan-Chen Guo, Ying-Tian Liu, Jianhui Liu, Yangguang Li, Yan-Pei Cao, Ding Liang, and Xiaojuan Qi. Texgen: a generative diffusion model for mesh textures. *ACM Transactions on Graphics (TOG)*, 43(6):1–14, 2024.

Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4252–4262, 2024a.

Zheng Zeng, Valentin Deschaintre, Iliyan Georgiev, Yannick Hold-Geoffroy, Yiwei Hu, Fujun Luan, Ling-Qi Yan, and Miloš Hašan. RGB↔X: Image decomposition and synthesis using material-and lighting-aware diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

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 (TOG)*, 42(4):1–16, 2023.

Hongkun Zhang, Zherong Pan, Congyi Zhang, Lifeng Zhu, and Xifeng Gao. Texpainter: Generative mesh texturing with multi-view consistency. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024a.

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, 2024b.

Shangzhan Zhang, Sida Peng, Tao Xu, Yuanbo Yang, Tianrun Chen, Nan Xue, Yujun Shen, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. Mapa: Text-driven photorealistic material painting for 3d shapes. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024c.

Yuqing Zhang, Yuan Liu, Zhiyu Xie, Lei Yang, Zhongyuan Liu, Mengzhou Yang, Runze Zhang, Qilong Kou, Cheng Lin, Wenping Wang, et al. Dreammat: High-quality pbr material generation with geometry-and light-aware diffusion models. *ACM Transactions on Graphics*, 43(4):1–18, 2024d.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *NeurIPS*, 2023.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.