Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 145–154
https://doi.org/10.54941/ahfe1008249

Updated Comparative Analysis Between Recent 3D Object Generation Models. Adaptation to Immersive Technologies

Geovana Amorim Abensur, Agustín Alejandro Ortiz Diaz, Sergio Cleger Tamayo, Delrick Nunes de Oliveira, Paulo Sergio Da Silva Farias, Gilberto Oliveira Neto, and Gabriel Rossi Zanini

Sidia Institute of Science and Technology. Manaus, Amazonas, 69055-035, Brazil

ABSTRACT

Artificial Intelligence-based Content Generation (AIGC) and Extended Reality (XR) enable scalable creation of immersive 3D assets, but single-image-to-3D models still involve trade-offs between quality, responsiveness, and runtime performance. This study evaluates two generative models, Trellis and TripoSR, across Animals, Objects, Humanoids, and Places categories. The main comparison uses original generated meshes, while humanoid assets are additionally analysed after polygon reduction through original, intermediate, and low-polygon versions. Metrics are grouped into generation complexity, geometric accuracy, texture fidelity, perceptual realism and semantic consistency, and VR performance. Results show that Trellis achieves the best reconstruction quality but requires approximately 30 seconds for generation. TripoSR provides the fastest generation and stable XR performance, while Trellis reaches the highest CLIP-Similarity, indicating stronger semantic alignment but heavier meshes. These findings show that model selection for XR depends on the target application, balancing reconstruction quality, semantic consistency, and runtime constraints.

Keywords: Artificial intelligence generated content, 3D generative content, Extended reality, Quantitative metrics, 3D real-time rendering

INTRODUCTION

One of the application areas that has benefited from recent advances in artificial intelligence is the development of models capable of generating three-dimensional representations from a single two-dimensional image. This approach seeks to infer the geometry, texture, and spatial structure of an object using techniques primarily based on deep learning (DL) and computer vision. These models learn to estimate information that is not explicitly present in the input image, such as hidden parts of the object, by training with large datasets (Wu, 2024; Cleger, 2025).

However, this remains a complex problem due to the inherent ambiguity of reconstructing complete three-dimensional information from limited visual data. Therefore, current research focuses on improving the generalizability of models, integrating more accurate geometric representations, and leveraging advanced DL techniques. These advances aim to generate more realistic

Received October 7, 2026; Revised November 8, 2026; Accepted November 14, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

and consistent reconstructions, expanding the possibilities for using these technologies in various scientific and technological fields (Xiang, 2025; Zhao, 2025).

Generating 3D objects from a single image has relevant applications in areas such as extended reality (XR), computer-aided design, and digital content creation. A recent study presented a comprehensive evaluation of three 3D mesh generation models from a single image and their application in immersive environments (Cleger, 2025). These three models, Unique3D (Wu, 2024), CRM (Wang, 2024), and InstantMesh (Xu, 2024), analysed follow the same strategy for generating 3D meshes, this strategy is structured in two well-defined stages. First, the input image is fed into a diffusion model capable of generating a set of multi-view images. Then, these multi-view images are fed into another model (e.g., convolutional U-Net or transformers) to finally reconstruct the 3D mesh (Cleger, 2025).

However other promising strategies have been published, achieving excellent results in the automatic generation of 3D meshes from a single image. Two interesting proposals are:

- **Trellis** (Xiang, 2025) is a large-scale 3D asset generation model that stands out for its flexible output format selection (e.g., radiation fields, 3D Gaussians, and meshes) and advanced local 3D editing capabilities.
- **TripoSR** (Tochilkin, 2024) focuses on real-time reconstruction from a single image using a TriPlane-NeRF architecture based on transformers; generates textured meshes in less than 0.5 seconds.

Different types of quantitative metrics have been used to evaluate each of the existing models. Some of the most used metrics to ensure that the generated 3D assets maintain geometric integrity and visual quality are Chamfer Distance, Volumetric IoU, F-score, and Clip-Similarity (Xu, 2024; Grande, 2024). The potential application of these 3D resources in immersive environments using XR technologies has been validated through the integration of the generated meshes in embedded applications on head-mounted displays (HMDs) (Cleger, 2025). Some of the key performance indicators that have been used as metrics are Frames Per Second, App GPU Time, and Average Prediction (Grande, 2024).

This work aims to perform a comparative analysis of two of the most prominent and recent models for generating 3D assets from a single image. To carry out this comparative analysis, we selected a set of relevant metrics used in various modern studies as comparison parameters. The presented study provides relevant information that can help developers improve the user's experience in immersive technologies such as XR.

RELATED WORK

Significant advances in automatic 3D content generation are continuously revolutionizing the development of immersive extended reality (XR) environments (Cleger, 2025). Current deep learning methods enable the automatic reconstruction of textured 3D objects from single-view images,

thereby reducing production time and facilitating scalable content creation for immersive environments (Wu, 2023). This contrasts with traditional approaches—such as computer-aided design (CAD), photogrammetry, and LiDAR scanning—which offer high geometric precision but require considerable manual effort and specialized expertise, while also offering limited scalability (Wu, 2024; Liu, 2024).

Current AI-based 3D generation methods can be classified into three categories (Samavati, 2023):

- **Optimization-based methods**, such as DreamFusion (Poole, 2022), Magic3D (Lin, 2022), and Repaint123 (Tao, 2023), generate high-quality 3D assets through iterative optimization. Although they achieve excellent photorealism, their high computational cost and long inference times limit their suitability for real-time VR applications.
- **Direct reconstruction models (feedforward),** such as CRM (Wang, 2024), Unique3D (Wu, 2024), and InstantMesh (Xu, 2024). These types of models predict textured 3D meshes from a single image in a single forward pass, making them better suited for interactive XR applications.
- **Hybrid approaches,** such as GET3D (Gao, 2022), EG3D (Chan, 2022) and Wonder3D++ (Long, 2023), combine optimization and feedforward techniques to improve reconstruction quality while simultaneously reducing computational cost. Despite producing high-quality results, their inference speed and resource requirements continue to pose challenges for real-time immersive applications.

However, other recent models also designed for AI-driven 3D generation from a single image, have introduced increasingly efficient and accurate architectures for reconstructing high-quality assets.

- **TripoSR** (Tochilkin, 2024) focuses on real-time single-image reconstruction through a transformer-based TriPlane-NeRF architecture, generating textured meshes in less than 0.5 seconds and outperforming previous feed-forward approaches in both reconstruction quality and computational efficiency.

Some more recent models, published in 2025, are:

- **Trellis** (Xiang, 2025) proposes a unified Structured LATent (SLAT) representation combined with Rectified Flow Transformers, enabling the generation of multiple 3D representations, including meshes, radiance fields, and Gaussian splats, while achieving state-of-the-art geometry and perceptual quality.
- **Hunyuan3D** (Zhao, 2025) adopts a two-stage pipeline that combines multi-view diffusion with transformer-based feed-forward reconstruction, improving multi-view consistency and reconstruction accuracy while maintaining practical inference times.
- **TripoSG** (Tochilkin, 2024), prioritizes high-fidelity 3D shape synthesis using large-scale Rectified Flow Transformers trained on millions of

3D assets and a Mixture-of-Experts architecture, producing highly detailed geometry at the expense of substantially higher computational requirements.

- **SAM3D** (Chen, 2025) introduces a foundation model that jointly predicts object geometry, texture, and spatial layout from a single image. Its main contribution is a large-scale human- and model-in-the-loop data engine that combines synthetic pretraining, semi-synthetic mid-training, and real-world preference-based alignment, effectively overcoming the scarcity of annotated 3D data.

Our work focuses on a comparative evaluation of two recent models Trellis and TripoSR. Due to access problems to the original implementations, we did not include the TripoSG, Hunyuan3D, and SAM3D models in the analysis.

PROPOSED METHODOLOGY

This work follows the methodology used by Cleger et al. (2025), which was structured to evaluate generative models designed to convert a single image into 3D assets for immersive XR applications. The methodology focuses on how generated 3D meshes should be compared in terms of reconstruction quality, visual fidelity, semantic consistency, generation cost, and standalone VR feasibility. The evaluation protocol is organized into four stages, as shown in Figure 1.

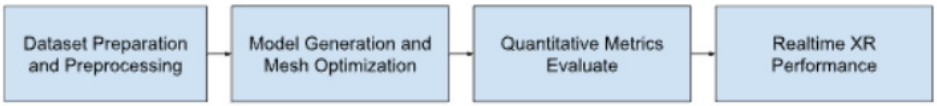


Figure 1: Methodology framework of the proposed Image-to-3D evaluation pipeline.

Dataset Preparation and Preprocessing

The evaluation protocol assumes a set of samples organized into representative four categories for immersive applications, as shown in Figure 2. Each sample must contain a single 2D input image, a paired reference 3D mesh, and the generated meshes produced by the evaluated Image-to-3D models (Cleger, 2025).



Figure 2: Selected object categories (Objects, Animals, Humanoids, Places).

To ensure fair comparison, the same 2D image should be used as input for all evaluated models within each category. Before quantitative evaluation, generated and reference meshes must be normalized and aligned into a common coordinate system. This preprocessing reduces differences in scale, orientation, and position, allowing the metrics to focus on reconstruction quality instead of transformation inconsistencies.

The preprocessing procedure consists of input standardization, mesh loading, scale normalization, orientation correction, and alignment between generated and reference meshes. Alignment should combine automatic transformations and refinement methods such as Iterative Closest Point, ensuring that all models are evaluated under the same spatial reference (Cleger, 2025; Gao, 2026).

3D Mesh Generation and Optimization

The methodology evaluates the direct output of each Image-to-3D model as the original mesh. These original meshes are used in the main comparison because they represent the native quality, polygon count, file size, and texture characteristics produced by each model (Fukaya, 2024). The two evaluated models are Trellis (Xiang, 2025) and TripoSR (Tochilkin, 2024).

For categories where interactive inspection is important, such as humanoid assets, the methodology also supports mesh optimization through polygon reduction (Cleger, 2025). In this case, the original mesh can be simplified into two additional variants: an intermediate reduced version, intended to balance visual preservation and rendering efficiency, and a low-polygon version, intended to maximize runtime performance, Figure 3.

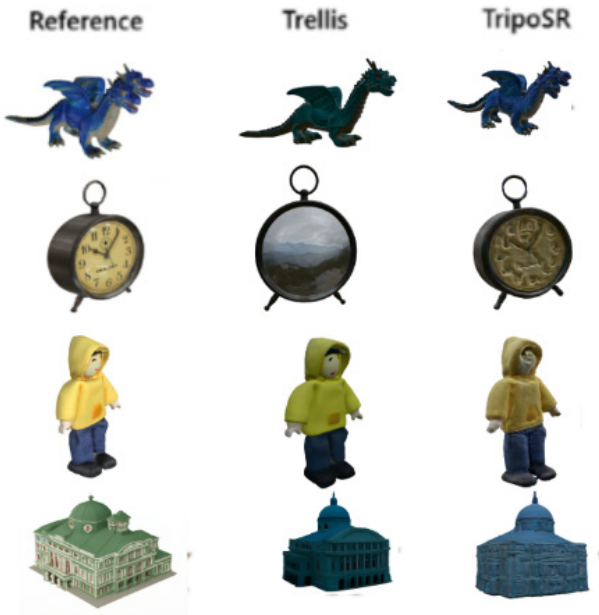


Figure 3: Generated and ground-truth meshes after alignment.

The inclusion of Trellis is also motivated by the recent plant reconstruction study by Gao et al., which evaluated Hunyuan3D_2.0 and Trellis among other contemporary one-image 3D reconstruction techniques (Gao, 2026). In that study, Trellis showed coherent spatial behaviour but did not always outperform the strongest models in geometric and perceptual metrics (Gao, 2026).

QUANTITATIVE EVALUATION METRICS

The proposed protocol groups the quantitative metrics into five dimensions (Samavati; 2023; Fukaya, 2024; Grande, 2024; Cascarano, 2025).

- **Generation complexity** evaluates the structural and computational cost of each generated asset through file size, polygon count, and generation time.
- **Geometric accuracy** evaluates the structural similarity between generated, and reference meshes using IoU, Chamfer Distance, Hausdorff Distance, and F-Score@0.05. These metrics capture global volume overlaps, average surface distance, worst-case deviation, and surface-level precision/recall.
- **Texture fidelity** evaluates rendered image similarity using PSNR and SSIM. These metrics provide pixel-level and structural comparisons between rendered views of generated and reference meshes.
- **Perceptual realism and semantic consistency** are assessed using LPIPS and CLIP-Similarity. LPIPS estimates learned perceptual similarity between generated and reference renders, while CLIP-Similarity measures semantic alignment between the original 2D input image and rendered views of the generated mesh.
- **XR performance** evaluates runtime feasibility through FPS, GPU utilization, and stale frames. These metrics indicate whether the generated asset can be rendered smoothly in standalone XR scenarios.

XR Performance Evaluation

The methodology evaluates XR feasibility by inserting the generated meshes into a controlled Unity-based VR scene and executing them on a standalone headset. The scene should be minimal, so that performance measurements reflect the rendering cost of the mesh itself rather than scene complexity.

Two conditions are considered. The first is the baseline scenario, where each original generated mesh is tested individually. The second is the optimization scenario, where the original, intermediate reduced, and low-polygon humanoid variants are compared, Figure 4. Additionally, runtime metrics are collected to determine how mesh complexity affects frame rate, GPU usage, and stale frames. This follows the baseline and optimization structure used in previous XR evaluation work, where mesh complexity was analysed in relation to runtime behaviour (Cleger, 2025).

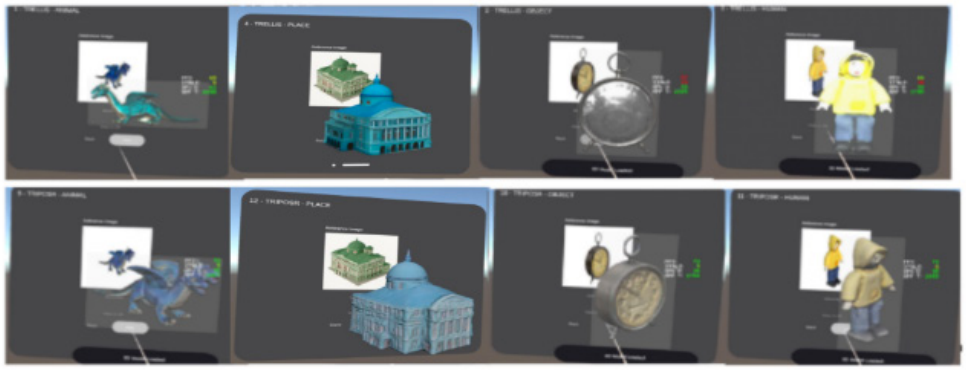


Figure 4: Unity VR baseline scene with original poly meshes.

EXPERIMENTS AND RESULTS

The experiments use two Image-to-3D generative models: Trellis and TripoSR. The main comparison used the original generated meshes.
The quantitative evaluation was conducted using a Python-based pipeline. The XR performance evaluation was performed on a Meta Quest 3 standalone headset using a Unity-based VR environment. The scene was designed to isolate the rendering cost of each mesh. Runtime behavior was measured using FPS, GPU utilization, and stale frames.

For each category, the same 2D input image was provided to two models. The generated meshes were loaded, normalized, and aligned to their paired reference meshes. After alignment, the quantitative metrics were computed and organized into three final tables, Tables 1, 2, and 3.

Table 1: Generation metrics and geometric accuracy for all methods.

Model	Category	Generation Metrics			Geometric Accuracy			
		Size ↓ (MB)	Poly ↓	Time↓ (s)	IoU ↑	Chamf ↓	Haus↓	F-Sc↑
Trellis	Animals	9.9862	288410	29.20	0.2718	0.0044	0.2305	0.7701
	Objects	11.7146	299044	28.80	0.3306	0.0035	0.1350	0.8565
	Humanoids	9.0135	275962	29.10	0.3449	0.0024	0.1101	0.8795
	Places	11.0507	298546	28.90	0.1549	0.0109	0.3261	0.5570
TripoSR	Animals	1.5219	79691	3.20	0.2468	0.0119	0.5235	0.7663
	Objects	1.2464	65247	10.30	0.1249	0.0056	0.3696	0.7744
	Humanoids	1.1663	61063	5.10	0.3148	0.0022	0.1155	0.8997
	Places	4.7180	247027	3.40	0.1353	0.0208	0.3662	0.4211

Generation Metrics and Geometric Accuracy

Table 1 shows that the models present distinct trade-offs in generation cost and geometric reconstruction. Trellis produced the heaviest meshes, with

polygon counts close to 300k in most categories and file sizes between approximately 9 MB and 12 MB. Its generation time remained stable at around 29 seconds. However, this higher polygon counts later affected XR performance, especially for Objects.

TripoSR generated the lightest and fastest outputs, with file sizes between 1.16 MB and 4.72 MB and generation times between 3.2 and 10.3 seconds. This makes it the most responsive model in the experiment.

Table 2: Texture fidelity and perceptual realism for all methods.

Model	Category	Texture Fidelity		Perceptual Realism	
		PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-Sim ↑
Trellis	Animals	5.7352	0.5110	0.6620	0.7037
	Objects	5.2066	0.2912	0.6290	0.6186
	Humanoids	4.6580	0.3830	0.5849	0.5745
	Places	4.9492	0.3989	0.7115	0.4887
TripoSR	Animals	7.1874	0.4697	0.6430	0.6772
	Objects	7.8375	0.2836	0.6775	0.5794
	Humanoids	7.9320	0.3847	0.4487	0.5515
	Places	4.9471	0.3755	0.6989	0.4805

Table 3: XR performance metrics for all methods.

Model	Category	VR Performance		
		FPS ↑	GPU Util (%) ↓	Stale (%) ↓
Trellis	Animals	71	52	5
	Objects	57	33	23
	Humanoids	66	33	15
	Places	65	62	6
TripoSR	Animals	72	36	4
	Objects	72	35	4
	Humanoids	72	35	6
	Places	71	35	7

TripoSR achieved competitive F-Score in Humanoids but showed weaker geometry in Objects and Places. Trellis performed competitively in Objects and Humanoids, but its results declined in Places.

Texture Fidelity and Perceptual Realism

Table 2 shows that TripoSR achieved better PSNR values in three of the categories (Animals, Objects, Humanoids), and nearly equivalent results in the remaining one (Places); however, the SSIM values reflect the opposite. These two metrics indicate stronger pixel-level similarity between generated and reference renders. TripoSR's main strength is not visual fidelity, but rather speed and runtime efficiency.

However, Trellis obtained the highest CLIP-Similarity across all categories. This indicates stronger semantic alignment with the original 2D input image, even though it did not always achieve the best geometric or perceptual reconstruction quality. This result shows that semantic consistency and geometric fidelity do not necessarily improve together.

XR Performance

Table 3 shows that TripoSR achieved the most stable XR performance, reaching between 71 and 72 FPS in all categories, and with stable frames between 4% and 7%. TripoSR also performed well, with low GPU utilization between 35% and 36%. Combined with its short generation time, this makes it the best option for fast and interactive XR workflows.

Finally, Trellis showed the weakest runtime behaviour. Although it reached 71 FPS in Animals, it dropped to 57 FPS in Objects and reached 23% stale frames. This limitation is associated with its heavier meshes and high polygon counts.

CONCLUSION

This study presented a comprehensive evaluation of two single-image-to-3D mesh generation models: Trellis and TripoSR targeting real-time use in immersive Virtual Reality (VR) environments. The models were assessed across four key object categories (Animals, Objects, Humanoids, and Places) using a set of quantitative metrics grouped under geometric accuracy, texture fidelity, perceptual realism and VR performance. The results show that no single model optimizes all evaluation dimensions.

Trellis is useful when semantic consistency with the input image is the main requirement, as it achieved the highest CLIP-Similarity. Nevertheless, its heavier meshes reduce standalone XR performance and may require polygon optimization before deployment.

TripoSR is the most suitable model for fast and interactive XR scenarios because it combines short generation time with stable runtime performance. Its lower geometric accuracy in complex categories may be acceptable when responsiveness is more important than detailed reconstruction.

In general, we can conclude that none of the two models analysed dominates all the dimensions evaluated for selecting a 3D generation model from a single image for XR. Each model has its own strengths and weaknesses and should be selected based on the objective of the task to be solved.

ACKNOWLEDGMENT

This paper was presented as part of the results of the Project “SIDIA-M_VST_PLATFORM_AND_APPLICATIONS”, carried out by the Institute of Science and Technology – SIDIA, in partnership with Samsung Electrônica da Amazônia LTDA, in accordance with the Information Technology Law n.8387/91 and article at the 39 of Decree 10,521/2020.

REFERENCES

Cascarano, P., De Vivo, F., D'Addona, D., et al. (2025) "A Comparative Analysis of 3D Modeling Methods for Integration into an Extended Reality Platform". IEEE International Conference on AI and eXtended and Virtual Reality, pp. 213–217. IEEE.

Chan, E.R., Lin, C.Z., Chan, M.A., et al. (2022) "EG3D: Efficient Geometry-aware 3D Generative Adversarial Networks". In: CVPR, pp. 8498–8508.

Chen, X., Chu, F., Gleize, P. et al. (2025). "SAM 3D: 3Dfy Anything in Images". https://arxiv.org/abs/2511.16624.

Cleger, S., Abensur, G., Ortiz, A., et al. (2025) "Comparative Analysis Between Different 3D Object Generation Models". Adaptation to Immersive Technologies. International Conference on HCI, pp. 18–35. Springer Nature Switzerland.

Fukaya, K., Daylamani-Zad, D., Agius, H. et al. (2024). "Evaluation Metrics for Intelligent Generation of Graphical Game Assets: A Systematic Survey-Based Framework". ACM Comput. Surv.

Gao, J., Yang, X., Qiu, Y., et al. (2022) "GET3D: A Generative Model of High-Quality 3D Textured Shapes Learned from Images". In: NeurIPS.

Gao, Z., Hartley, and Z., French, A. (2026). "Evaluation of one-image 3D reconstruction for plant model generation". Plant Methods, 22, 18.

Grande, R., Albusac, J., Vallejo, D., et al. (2024). "Performance Evaluation and Optimization of 3D Models from Low-Cost 3D Scanning Technologies for Virtual Reality". Appl. Sci. 14(14), 6037.

Li, Y., Z., Liu, Z. et al. (2025) "TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models". https://arxiv.org/abs/2502.06608.

Lin, C., Liu, L., Yang, J., et al. (2022) "Magic3D: High-Resolution Text-to-3D Generation". arXiv preprint arXiv:2211.10440.

Liu, J., Huang, X., Huang, T., et al. (2024) "A Comprehensive Survey on 3D Content Generation". arXiv preprint arXiv:2402.01166.

Long, X., Guo, Y., Lin, C., et al. (2023) "Wonder3D++: Single Image to 3D using Cross-Domain Diffusion". arXiv preprint arXiv:2310.15008.

Poole, B., Saharia, C., Chan, W., et al. (2022) "DreamFusion: Text-to-3D using 2D Diffusion". arXiv preprint arXiv:2210.02303.

Samavati, T., Soryani, M. (2023) "Deep Learning-Based 3D Reconstruction: A Survey". Artif. Intell. Rev. 56, 9175–9219.

Tao, M., Chen, J., et al. (2023) "Repaint123: Attentive Multi-view Text-to-3D Generation using Repainting Diffusion". arXiv preprint arXiv:2311.05712.

Tochilkin, D., Pankratz, D., Liu, Z. et al. (2024) "TripoSR: Fast 3D Object Reconstruction from a Single Image", https://arxiv.org/abs/2403.02151.

Wang, Z., Wang, Y., Chen, Y., et al. (2024) "CRM: Single Image to 3D Textured Mesh with Convolutional Reconstruction Model". In: European Conference on Computer Vision (ECCV), pp. 57–74. Springer, Cham.

Wu, J., Gan, W., Chen, Z., et al. (2023) "AI-Generated Content (AIGC): A Survey". arXiv preprint arXiv:2304.06632.

Wu, K., Liu, F., Cai, Z., et al. (2024). "Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image". https://arxiv.org/abs/2405.20343.

Xiang, J., Lv, Z., Xu, S. et al. (2025) "Structured 3D Latents for Scalable and Versatile 3D Generation", https://arxiv.org/abs/2412.01506.

Xu, J., Cheng, W., Gao, Y., et al. (2024) "InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models". arXiv preprint arXiv:2404.07191.

Zhao, Z., Lai, Z., Lin, Q. et al. (2025) "Hunyuan3D_2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation", https://arxiv.org/abs/2501.12202.