

## A. Background for 3D Foundation Models

Foundation models for learning object geometry priors have been widely studied via image representations, such as depth [2, 19], normal [1, 9, 15] or point [26, 27] maps. While these models mostly describe object structures for input view only, recent *3D foundation models* learn to directly encode object geometry in a native 3D latent space, e.g. constructed by tokenized point clouds [4, 34] or voxels [17, 20], thus can model complete object shapes. Specifically, [16, 35, 36] adopt a 3DShape2VecSet [34] architecture to compress sampled object points as a latent vector, which can be efficiently denoised via rectified flow models [8, 18]. Alternatively, [17, 30] propose a two-stage diffusion process using sparse voxels to support detailed and localized shape generation. Moreover, to avoid time-consuming mesh extraction and support real-time rendering, several works [22, 28, 31] adopt Gaussian Splatting [14] as the 3D representation and learns to decode corresponding parameters during inference, which enables the model to capture more realistic lighting effects in the rendered results. Unlike previous pipelines for 3D shape generation [10, 11, 21, 24, 29] that rely on an intermediate multi-view synthesis step, above methods fully operate on 3D object representations with explicit 3D geometry supervision, thus effectively model realistic object shape distributions as foundation priors.

## B. Data Processing Pipeline

We use a subset of Objaverse-XL [7] with 160K assets as our training data and follow [25] to render synthetic orbital videos as ground truths. Specifically, we first normalize each object inside a unit cube and center it at the origin. Note that such normalization is also assumed by the 3D foundation model. We then use Blender’s CYCLES engine to render 84-frame orbits of 576×576 pixels images with white background. During training, we randomly select a frame as the starting frame with a step of 4 to construct the target 21-frame orbit. For lighting, we follow the script from [30] to illuminate the object with curated HDRI environment maps. Finally, we follow [25] to set the camera FoV as 33.8 degrees. For static orbits, we sample azimuth angles that equally divide 360 degrees, and set the elevation to 0. For dynamic orbits, we smoothly transition the elevation values in the range [-60, 60] using a random set of sinusoids. In summary, our data processing pipeline closely follows [25], so that the finetuned base model from SVD shares a comparable performance.

## C. More Ablations for 3D Foundation Priors

In Figure 1, we show generated samples using the same input image and shape prior but different random seeds. We observe that incorporating 3D adapter preserves the

stochastic nature of the base video generative model, thus enabling the generation of videos with different object appearances. Moreover, all generated results close resemble the shape prior, *i.e.* as visualized in the rightmost column, which demonstrates its capability of shape control. In addition, we observe that the shape priors only act as a *soft* constraint, where generated images can differ in fine-grained object parts. We attribute this as two reasons: (i) both images and 3D shapes are processed in *patches* compressed by their respective VAEs, making conditioning on individual pixels challenging, (ii) the image and shape priors are conditioned in parallel via cross attention layers, hence the base video model has its own capability of balancing from multiple condition signals. Similar to other adapters [33], we observe that achieving strict identity consistency is challenging for 3D shape priors as well. Nevertheless, by incorporating shape conditions as an auxiliary constraint, our method can effectively regularize the overall structure and robustly ensure a plausible output.

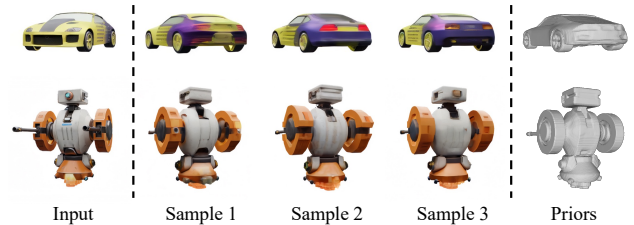


Figure 1. **Effects of random seeds for shape priors.** Our method can generate different object appearances following the same prior, preserving the stochastic nature of generative model.

In Figure 2, we visualize decoded shape priors under different seeds for the 3D foundation model. Since the foundation model uses a large CFG scale by default [36], we observe it produces a consistent shape prior with minor variance, which helps to accelerate the convergence for training and ensures a consistent behavior during inference.

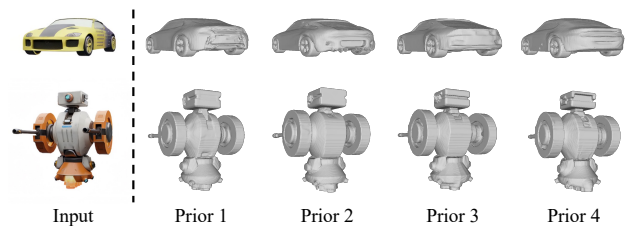


Figure 2. **Effects of random seeds for the 3D foundation model.** By incorporating a large CFG scale, the 3D foundation model demonstrate a consistent behavior on resulting priors

## D. More Results

We show in Figure 6 for results on more unseen examples with more views. Our method consistently generalizes to



Figure 3. **Qualitative Comparisons with NVS methods [5, 13, 37] for completeness.** Our method produces more realistic results with improved fidelity in shape and appearances details. Moreover, we directly output smooth video results leveraging temporal priors from general video model [3], as shown in supplementary videos.

diverse types of objects and produces consistent frames with realistic object shapes.

### E. More Comparisons with NVS Methods

We show more comparisons with several recent NVS baselines: Free3D [37], 3D-Adapter [5, 6], and MV-Adapter [13] in Figure 3. We obtain all results using their official code and weights. Since these methods either generate individual or sparse frames during inference, they are not directly comparable with our method or baselines methods [25, 32] that target on producing temporal consistent orbital videos with dense frames. Nevertheless, qualitative comparison show that our method achieves superior fidelity in terms of both shape and appearances details, leading to results with noticeably improved realism.

### F. More Comparison with Foundation Models

We show more comparisons with several recent 3D foundation models: Ouroboros3D [28], Gen-3Diffusion (GEN3D)

[31], and an improved version of Hunyuan3D (Hunyuan2.1) [23] in Figure 7. Both [28] and [31] learn to generate 3DGS as the 3D representation and renders results with view-dependent lighting effects via Gaussian Splatting. In addition, unlike Hunyuan2.0 [36] that bakes the input lighting into the textures, Hunyuan2.1 [23] first delight the input image and estimate PBR textures from the intermediate inputs, thus allows relighting for the generated meshes to produce more complex light effects such as reflection for mental materials. In the comparison, we show that our method remain superior in visual quality, especially in terms of the texture fidelity in rendered results as well as base color alignment to the ground truth. To this end, this further verifies the value of orbital video generation methods for visual aesthetic.

### G. 3D Reconstruction from Video Results

We show in Figure 4 that similar to baseline works [25, 32], our method can be extended to generate 3D meshes from the video results, which also demonstrates the multi-view con-

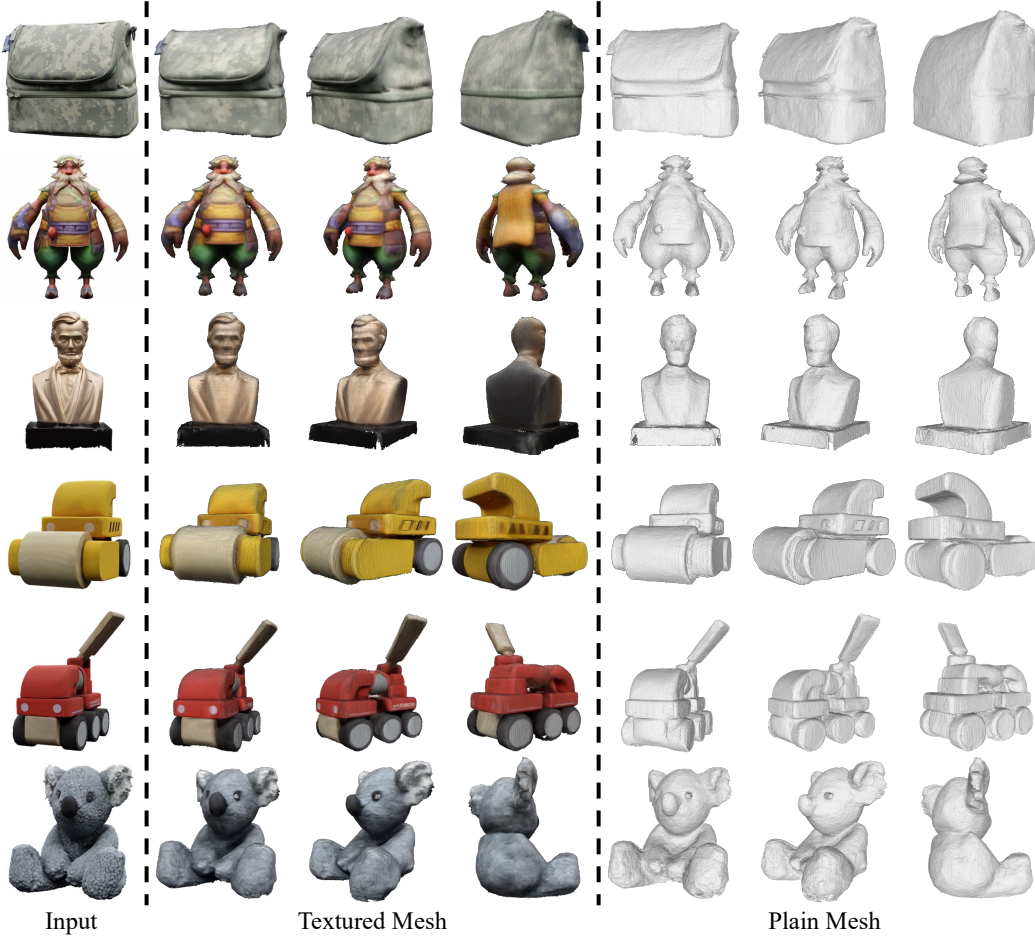


Figure 4. **Reconstructed textured meshes from generated video results.** Similar to [25, 32], our method can be extended to reconstruct textured 3D meshes, which also demonstrates the multi-view consistency of generated videos.

sistency. Specifically, we use a simple method that adopts 2DGS [12] to directly reconstruct textured meshes using the predicted RGB images, and observe that the obtained 3D models faithfully recover the input images.

## H. Failure Cases

We show examples of failure cases in Figure 5. Due to the limited geometry resolution, the 3D foundation model fails to generate accurate shape priors for complex scenes and tiny object regions. Moreover, the visual fidelity of the results is also bounded by the capacity of the base video model, *e.g.* SVD used in our method, which makes synthesizing and controlling for detailed object textures and components challenging. Nevertheless, we compare with [32] with higher spatial resolution and find simply scaling the base video model without incorporating 3D shape priors does not yield improved results. In view of this, future research works are encouraged to combine the shape priors with more powerful base video models to enhance visual fidelity.



Figure 5. **Examples of failure cases.** Due to the limited resolution, both base video model and 3D foundation model fail to recover fine-grained details on complex scenes, leading to blurry results and ineffective shape control particular when the shape priors disagree with ground truth videos.

## I. Supplementary Video

We refer readers with more results in the supplementary videos, tested on a wide range of objects and real world examples to demonstrate the efficacy and generalizability of the proposed method.





Figure 6. We show more results of generated frames on diverse input views and unseen objects.

## References

- [1] Gwangbin Bae and Andrew J Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9535–9545, 2024. 1
- [2] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1
- [3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages





Figure 7. **Qualitative Comparisons with 3D generation methods** [23, 28, 31]. Our method produces results with higher fidelity compared to 3DGS based methods [28, 31] and ensures better aligned base colors over foundation models that explicitly estimate mesh texture maps [23, 36]. This further verifies the efficacy of native video generation methods in producing visually improved results.

22563–22575, 2023. 2

- [4] Jen-Hao Rick Chang, Yuyang Wang, Miguel Angel Bautista Martin, Jiatao Gu, Josh Susskind, and Oncel Tuzel. 3d shape tokenization. *arXiv preprint arXiv:2412.15618*, 2024. 1
- [5] Hansheng Chen, Bokui Shen, Yulin Liu, Ruoxi Shi, Linqi Zhou, Connor Z. Lin, Jiayuan Gu, Hao Su, Gordon Wetzstein, and Leonidas Guibas. 3d-adapter: Geometry-consistent multi-view diffusion for high-quality 3d generation, 2024. 2
- [6] Hansheng Chen, Ruoxi Shi, Yulin Liu, Bokui Shen, Jiayuan Gu, Gordon Wetzstein, Hao Su, and Leonidas Guibas. Generic 3d diffusion adapter using controlled multi-view

editing. *arXiv preprint arXiv:2403.12032*, 2024. 2

- [7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 1
- [8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis.

- In *Forty-first international conference on machine learning*, 2024. 1
- [9] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 1
  - [10] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 1
  - [11] 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. 1
  - [12] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 3
  - [13] Zehuan Huang, Yuan-Chen 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*, 2024. 2
  - [14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 1
  - [15] Rawal Khrodgar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 1
  - [16] 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
  - [17] Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Tripo3r: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025. 1
  - [18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1
  - [19] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 1
  - [20] 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
  - [21] 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*, pages 1–18. Springer, 2024. 1
  - [22] Zhenyu Tang, Junwu Zhang, Xinhua Cheng, Wangbo Yu, Chaoran Feng, Yatian Pang, Bin Lin, and Li Yuan. Cycle3d: High-quality and consistent image-to-3d generation via generation-reconstruction cycle, 2024. 1
  - [23] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material, 2025. 2, 5
  - [24] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Tripo3r: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 1
  - [25] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 1, 2, 3
  - [26] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025. 1
  - [27] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1
  - [28] Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Yu Qiao, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. *arXiv preprint arXiv:2406.03184*, 2024. 1, 2, 5
  - [29] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1
  - [30] 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. 1
  - [31] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard Pons-Moll. Gen-3Diffusion: Realistic Image-to-3D Generation via 2D 3D Diffusion Synergy. 2024. 1, 2, 5
  - [32] Haibo Yang, Yang Chen, Yingwei Pan, Ting Yao, Zhineng Chen, Chong-Wah Ngo, and Tao Mei. Hi3d: Pursuing high-resolution image-to-3d generation with video diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6870–6879, 2024. 2, 3
  - [33] Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1
  - [34] 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. 1

- [35] 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](#)
- [36] 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. [1](#), [2](#), [5](#)
- [37] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. *arXiv preprint arXiv:2312.04551*, 2023. [2](#)