Tencent Hunyuan3D-1.0: A Unified Framework for Text-to-3D and Image-to-3D Generation

Xianghui Yang[*], Huiwen Shi[*], Bowen Zhang[*], Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu,

Xinzhou Wang, Qingxiang Lin, Jiaao Yu, Lifu Wang, Zhuo Chen, Sicong Liu,

Yuhong Liu, Yong Yang, Di Wang, Jie Jiang, Chunchao Guo[†]

Tencent Hunyuan

Figure 1. 3D asset gallery. All the 3D assets are generated by Hunyuan3D-1.0 given a text prompt or a single image as input. Hunyuan3D-1.0 is a unified framework that supports high-quality text- and image-conditioned 3D generation both.

Abstract

While 3D generative models have greatly improved artists' workflows, the existing diffusion models for 3D generation suffer from slow generation and poor generalization. To address this issue, we propose a two-stage approach named Hunyuan3D-1.0 including a lite version and a standard version, that both support text- and image-conditioned generation. In the first stage, we employ a multi-view diffusion model that efficiently generates multi-view RGB in approximately 4 seconds. These multi-view images capture rich details of the 3D asset from different viewpoints, relaxing the tasks from single-view to multi-view reconstruction. In the second stage, we introduce a feed-forward reconstruction model that rapidly and faithfully reconstructs the 3D asset given the generated multi-view images in approximately 7 seconds. The reconstruction net-

---

[*]Equal contribution. [†]Corresponding author.

1

work learns to handle noises and in-consistency introduced by the multi-view diffusion and leverages the available information from the condition image to efficiently recover the 3D structure. Our framework involves the text-to-image model, i.e., Hunyuan-DiT, making it a unified framework to support both text- and image-conditioned 3D generation. Our standard version has $3\times$ more parameters than our lite and other existing model. Our Hunyuan3D-1.0 achieves an impressive balance between speed and quality, significantly reducing generation time while maintaining the quality and diversity of the produced assets.

1. Introduction

3D generation has long been an attractive and active topic in the fields of computer vision and computer graphics, with significant applications spanning gaming, film, e-commerce, and robotics. Creating high-quality 3D assets is a time-intensive process for artists, making automatic generation a long-term goal for researchers. Early efforts in this field focused on unconditional generation within specific categories, constrained by 3D representation and data limitations. The recent success of scaling laws in large language models (LLMs), as well as in image and video generation, has illuminated a path toward this long-term vision. However, achieving similar advancements in 3D asset generation remains challenging due to the expressive nature of 3D assets and the limited availability of comprehensive datasets. The largest existing 3D dataset, Objarverse-xl [7], comprises only 10 million assets, which pales in comparison to the large-scale datasets available for language, image, and video tasks. Leveraging priors from 2D generative models presents a promising approach to address this limitation.

To take advantage of 2D generative models, pioneering works have explored this problem and achieved notable advancements. Poole *et al*. [34] utilize Score Distillation Sampling (SDS) to distill a 3D representation, *i.e.*, Nerf [30], via 2D image diffusion models. Despite issues with over-saturation and significant time costs, this approach inspired subsequent 2D lifting research. Follow-up works have explored to improve sampling efficiency [51], fine-tune diffusion models into multi-view diffusion frameworks [1, 23, 40], and replace sampling losses with regular rendering losses [22, 25, 26, 60]. However, these optimization-based methods remain time-consuming, requiring anywhere from 5 minutes to an hour to optimize the 3D representation [30, 49, 57, 62]. In contrast, feed-forward methods [4, 11, 13, 43, 59] can generate 3D objects in mere seconds but often struggle with generalization to unseen objects and fail to generate thin, paper-like structures. Disentangling single-view generation tasks into generating multi-view images and completing sparse-view reconstruction via feed-forward methods is a promising path to mitigate generalization issues and eliminate the optimization problem in SDS.

Despite several works [58] in multi-view generation and sparse-view reconstruction, few have organized these approaches into a cohesive framework that addresses their combined challenges. First, widely used multi-view diffusion models are often criticized for multi-view inconsistency and slow denoising processes. Second, sparse-view reconstruction models typically rely solely on view-aware RGB images to predict 3D representations. Addressing these issues separately is challenging. Noticing the need to tackle these sub-tasks together, we propose Hunyuan3D-1.0, which integrates the strengths of multi-view diffusion models and sparse-view reconstruction models to achieve 3D generation in 10 seconds in the best-case scenario, achieving a subtle balance between generalization and quality. In the first stage, the multi-view diffusion model generates RGB to finish the 2D-to-3D lifting. We fine-tune a large-scale 2D diffusion model to generate multi-view images to enhance the model's understanding of 3D information. Additionally, we set the 0-elevation camera orbit for the generated views to maximize the visible area between generated views. In the second stage, the sparse-view reconstruction model utilizes the imperfectly consistent multi-view images to recover the underlying 3D shape. Unlike most sparse-view reconstruction models that only use RGB images with known poses, we incorporate the conditional image, without the known view pose, to provide additional view information as an auxiliary input to cover the unseen part in the generated multi-view images. Furthermore, we employ a linear unpatchify layer operation to enrich details in the latent space without incurring additional memory or computational costs.

Our contributions are summarized as follows:
- We introduce a unified framework Hunyuan3D-1.0, support text- and image- condition 3D generation both.
- We design the 0-elevation pose distribution in the multi-view generation, maximizing the visible area between generated views.
- We introduce a view-aware classifier-free guidance that balances the controllability and diversity for different view generations.
- We incorporate the hybrid input that involves the uncalibrated condition image as an auxiliary view in the sparse-view reconstruction process to compensate for the unseen part in the generated images.

2. Related Works

Recent advances in multi-view generation models and sparse-view reconstruction models have significantly improved the quality of image-to-3D generation. Here, we briefly summarize the related works.
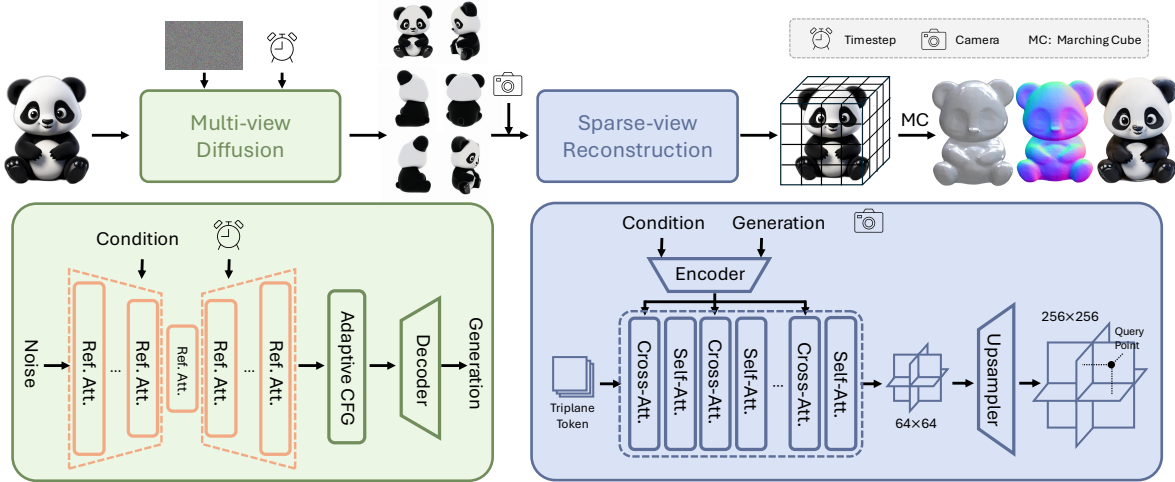
Figure 2. The overview of our Hunyuan3D-1.0. Given an input image, we first utilize a multi-view diffusion model to synthesize 6 novel views at fixed camera poses. Then we feed the generated multi-view images into a transformer-based sparse-view large reconstruction model to reconstruct a high-quality 3D mesh. The whole image-to-3D generation process takes only around 10 seconds.

Multi-view Generation. The potential of 2D diffusion models for novel-view generation has gained significant attention since the introduction of 3DiM [53] and Zero-1-to-3 [23]. A key challenge in this area is multi-view consistency, as the quality of downstream 3D reconstruction heavily relies on it to accurately estimate 3D structures. MVDiffusion [42] addresses this by generating multi-view images in parallel using correspondence-aware attention, which facilitates cross-view information interaction. MV-Dream [40]. Wonder3D [26] enhances multi-view consistency through the design of multi-view self-attention mechanisms. Zero123++ [39] tiles multi-views into a single image, which is also used in Direct2.5 [28] and Instant3D [20]. Syncdreamer [25] projects multi-view features into 3D volumes and enforces 3D alignment in the noise space. One significant issue with cross-view attention is its computational complexity, which increases quadratically with image size. Although some works [16, 44] introduce epipolar features into multi-view attention to enhance viewpoint fusion, the pre-computation of epipolar lines remains non-trivial. Era3D [21] proposes row-wise attention to reduce computational workloads by pre-defining the generated images with an elevation of 0. In this work, we propose two versions of multi-view generation models to balance efficiency and quality. The larger model has $3\times$ parameters than existing models, and both models are trained on a large-scale internal dataset, ensuring a more efficient and high-quality multi-view generation.

Sparse-view Reconstruction. Sparse-view reconstruction focuses on reconstructing target objects or scenes using only 2-10 input images, which is an extreme case in traditional Multi-View Stereo (MVS) tasks. Classical MVS methods often emphasize feature matching for depth estimation [2, 3] or voxel representations [5, 17, 33, 38, 45]. Learning-based MVS methods typically replace specific modules with learnable networks, such as feature matching [10, 18, 29, 46, 67], depth fusion [8, 35], and depth inference from multi-view images [14, 61, 63, 66]. In contrast to the explicit representations used by MVS, recent neural approaches [24, 31, 32, 49, 64, 65] represent implicit field via multi-layer perceptrons (MLPs). These methods often rely on camera parameter estimation obtained through complex calibration procedures, such as Structure-from-Motion approaches [15, 37]. However, in real-life scenarios, inaccuracies in pre-estimated camera parameters can be to the performance of these algorithms. Recent works [19, 50] propose directly predicting the geometry of visible surfaces without any explicit knowledge of the camera parameters. We notice most existing methods assume either purely posed images or purely uncalibrated images as inputs, neglecting the need for hybrid inputs. In this work, we address this gap by considering both calibrated inputs and uncalibrated images to achieve detailed reconstructions, thereby better integrating the sparse-view reconstruction framework into our 3D generation pipeline.

3. Methods

We present the two stages in our approach, Hunyuan3D-1.0, in this section. First, we introduce the multi-view diffusion model for 2D-to-3D lifting in Sec. 3.1. Second, we discuss pose-known and pose-unknown image fusion and the super-resolution layer within the sparse-view reconstruction framework in Sec. 3.2.

| Input | w/o super-resolution | w/ super-resolution |

Figure 3. Visual comparison of reconstruction using (a) low-resolution triplane vs (a) high-resolution triplane by super-resolution.

3.1. Multi-view Diffusion Model

Witnessing the huge success of diffusion models in 2D generation, their potential on novel-view generation models has also been explored. Most novel-view [23, 53] or multi-view [25, 40, 47, 48] generation models leverage the generalization ability of the diffusion model trained on a large amount of data. We further scale it up by training a larger model with $3\times$ parameters on a large-scale dataset.

Multi-view Generation. We simultaneously generate multi-view images by organizing the multi-view images as a grid. To achieve this, we follow Zero-1-to-3++ [39] and scale it up by replacing the model with a $3\times$ larger model [36]. We utilize reference attention as employed in Zero-1-to-3++ [39]. Reference attention guides the diffusion model to generate images that share similar semantic content and texture with a reference image. This involves running the denoising UNet model on an extra condition image and appending the self-attention key and value matrices from the condition image to the corresponding attention layers during the denoising process. Unlike the rendering settings of Zero-1-to-3++, we render target images with an elevation of $0°$, azimuth of $\{0°, 60°, 120°, 180°, 240°, 300°\}$ and a white background. The target images are arranged in a 3×2 grid, with the size of 960×640 for the lite model and 1536×1024 for the standard model.

Adaptive Classifier-free Guidance. Classifier-free guidance (CFG) [12] is a widely used sampling technique in diffusion models to balance controllability and diversity. In multi-view generation, it has been observed that a small CFG helps synthesize detailed textures but introduces unacceptable artifacts, while a large CFG ensures excellent object geometry at the expense of texture quality [55]. Additionally, the performance of different CFG scale values varies across different views, such as front and back views. A higher CFG scale retains more details from the condition image for front views, but it can result in darker back views. Based on these observations, we propose an Adaptive Classifier-Free Guidance schedule that sets different CFG scale values for different views and time steps. Intuitively, for front views and at early denoising time steps,

we set a higher CFG scale, which is then decreased as the denoising process progresses and as the view of the generated image diverges from the condition image. Specifically, we set the front view CFG scale following the curve:

$$w_t = 2 + 16 * (t/1000)^5 \qquad (1)$$

For other views, we apply scaled versions of this curve

$$w_{t,v} = w_t * \tau_v, \qquad (2)$$

where we define $\tau_v \in [0.5, 1]$ according to view distance from the front, and $\tau_{front} = 1$ and $\tau_{back} = 0.5$. This adaptive approach allows us to dynamically adjust the CFG scale, optimizing for both texture detail and geometric accuracy across different views and stages of the denoising process. By doing so, we achieve a more balanced and high-quality multi-view generation.

3.2. Sparse-view Reconstruction Model

In this section, we detail our sparse-view reconstruction model, a transformer-based approach designed to recover 3D shapes in a feed-forward manner within 2 seconds, using the generated multi-view images from the multi-view diffusion model. Unlike larger reconstruction models that rely on 1 or 3 RGB images [11, 13, 20], our method combines calibrated and un-calibrated inputs, lightweight super-resolution, and explicit 3D representation to achieve high-quality 3D reconstructions from sparse-view inputs. This approach addresses the limitations of existing methods and provides a robust solution for practical 3D generation tasks.

Hybrid Inputs. Our sparse-view reconstruction model utilizes a combination of calibrated and uncalibrated images (*i.e.*, the user inputs) for the reconstruction process. The calibrated images come with their corresponding camera embeddings, which are predefined during the training phase of the multi-view diffusion model. Since we constrain the multi-view generation to a 0-elevation orbit, the model has difficulty capturing information from top or bottom views, resulting in uncertainties in these perspectives.

To address this limitation, we propose incorporating information from the uncalibrated condition image into the reconstruction process. Specifically, we extract features from the condition image and create a dedicated view-agnostic branch to integrate this information. This branch takes a special full-zero embedding as the camera embedding in the attention module, allowing the model to distinguish the condition images from generated images and effectively incorporate the features from the condition image. This design minimizes uncertainties and improves the model's ability to accurately reconstruct 3D shapes, even from sparse views.

Super-resolution. While a higher feature resolution in transformer-based reconstruction enables the encoding of more detailed aspects of the 3D shape, we have noticed that

Method	CD↓	F-score$_{\tau=0.1}$ ↑	F-score$_{\tau=0.2}$ ↓	F-score$_{\tau=0.5}$ ↑
SyncDreamer [25]	0.518	0.306	0.543	0.852
TripoSR [43]	0.356	0.511	0.727	0.920
Wonder3D [26]	0.573	0.277	0.489	0.809
CRM [52]	0.262	0.538	0.800	0.977
LGM [41]	0.409	0.442	0.658	0.881
OpenLRM [11]	0.214	0.605	0.840	**0.997**
InstantMesh [58]	0.216	0.670	0.862	0.977
Ours-lite	0.199	0.661	0.877	0.986
Ours-std	**0.175**	**0.735**	**0.910**	0.987

Table 1. Comparison on GSO [9]. Our Hunyuan3D-1.0 achieve new state-of-the-art performance on GSO [9] in terms of CD and F-score metrics.

Method	CD↓	F-score$_{\tau=0.1}$ ↑	F-score$_{\tau=0.2}$ ↓	F-score$_{\tau=0.5}$ ↑
SyncDreamer [25]	0.202	0.632	0.884	0.995
TripoSR [43]	0.157	0.776	0.915	**0.999**
Wonder3D [26]	0.249	0.554	0.815	0.976
OpenLRM [11]	0.158	0.754	0.940	0.992
CRM [52]	0.245	0.568	0.830	0.979
LGM [41]	0.269	0.533	0.769	0.967
InstantMesh [58]	0.187	0.678	0.897	0.990
Ours-lite	0.150	0.786	0.938	0.997
Ours-std	**0.136**	**0.814**	**0.948**	0.998

Table 2. Comparison on OminiObject3D [56]. Our Hunyuan3D-1.0 achieve new state-of-the-art performance on OmniObject3D [56] in terms of CD and F-score metrics.

most existing works predominantly use low-resolution triplanes. These artifacts are directly linked to the triplane resolution, and we identify this as an aliasing issue that can be alleviated by increasing the resolution. The enhanced capacity also improves the geometry. However, it is not straightforward to increase the resolution, as it follows a quadratic complexity with the size. Drawing inspiration from the recent works [54, 68], we propose an upsampling module for triplane super-resolution. This approach maintains linear complexity with respect to the input size by avoiding self-attention on the higher-resolution triplane tokens. With this modification, we initially produced 64×64 resolution triplanes with 1024 channels. We further increase the triplane resolution by decoding one low-resolution triplane token into 4×4 high-resolution triplane tokens using a linear layer, resulting in 120-channel triplane features at a 256×256 resolution. Fig. 3 demonstrates richer details captured by the model with higher-resolution triplanes.

3D Representation. While most existing 3D generation models end with implicit representations, *e.g.*, NeRF or Gaussian Splatting, we argue that implicit representations are not the final goal of 3D generation. Only explicit representations can be seamlessly utilized by artists or users in practical applications. Therefore, we adopt the Signed Distance Function (SDF) from NeuS [49] in our reconstruction model to represent the shape via implicit representation and convert it into explicit meshes by marching cube [27]. Given the generated meshes, we extract their UV maps by unwarping. The final outputs are ready for texture mapping and further artistic refinement, which can be directly used in various applications.

4. Implementation

Training datasets. We train the multi-view diffusion model and the sparse-view reconstruction model using an internal dataset analogous to Objaverse [6, 7]. To ensure the quality and relevance of the training data, we filtered out 3D data that contained complex scenes, lacked meaningful textures, or exhibited unreasonable distortions. Additionally, all 3D objects in the dataset were scaled to fit within a unit sphere

before rendering.

For rendering the condition images, we employed a random sampling strategy for camera poses. Specifically, we sampled the camera elevation from a range of [-20, 60] degrees and the azimuth from [0, 360] degrees. The HDR is randomly sampled from the a HDR set and field of view (FOV) were sampled from a uniform distribution $U(47, 0.01)$, and the camera distance was sampled from $U(1.5, 0.1)$. For rendering the target images, we fix the camera parameters for model learning. We render 24 images with azimuth angles uniformly sampled from the set $\{0, 15, 30, 45, ..., 330, 345\}$ degrees, and a fixed elevation of 0 degrees. The FOV was set to 47.9 degrees, and the camera distance was fixed at 1.5 units. Uniform lighting conditions were applied to ensure consistency across the target images. All renderings were completed using Blender with a fixed rendering resolution of 1024×1024.

Training details. We train the multi-view diffusion model and sparse-view reconstruction model separately. For the multi-view diffusion model, our lite verison adopts the SD-2.1 as the backbone and our standard version takes SD-XL as the backbone. The RGB images are organized as a 3×2 grid. The condition image is randomly resized with $[256, 512]$ during training, while fixed with size 512 during inference. The target images are all resized into 320×320. For the sparse-view reconstruction model, we extract the image features via DINO encoder and adopt the tri-plane as the intermediate latent representation. The reconstruction model is fisrt trained with 256×256 multiview input images and then finetuned with 512×512 multiview input images. All training is completed on 64 A100 GPUs.

Evaluation. We evaluate our models against existing approaches using two public datasets: GSO [9] and OmniObject3D [56] with randomly sampled approximately 70 objects. To convert implicit 3D representations into meshes, we utilized the Marching Cubes algorithm [27] to extract iso-surfaces. We then sampled 10,000 points from these surfaces to compute the Chamfer Distance (CD) and F-score (FS), which are standard metrics for evaluating the accuracy of 3D shape reconstructions. Since some methods

Figure 4. Qualitative comparisons of single-view generation. Our Hunyuan3D-1.0 achieves better visual quality compared to existing methods

require manual recalibration to align the predicted shapes with the ground truth, we applied the Iterative Closest Point (ICP) method for alignment in cases where the generation pose was unknown.

5. Results

We quantitatively and qualitatively compare Hunyuan3D-1.0 to previous state-of-the-art methods using two different datasets with 3D reconstruction metrics.

Quantitative Comparisons. We compare Hunyuan3D-1.0 with the existing state-of-the-art baselines on 3D reconstruction that use feed-forward techniques, including OpenLRM [11], SyncDreamer [25], TripoSR [43], Wonder3D [26], CRM [52], LGM [41] and InstantMesh [58]. As shown in Table 1 and Table 2, our Hunyuan3D-1.0, especially our standard version, outperforms all the baselines,
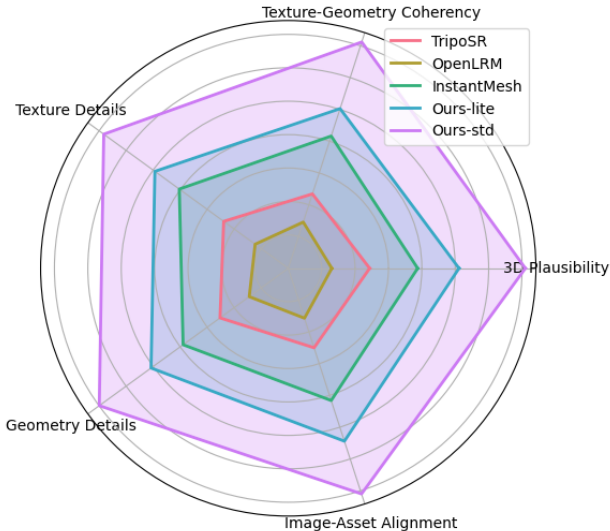
Figure 5. User study. Our Hunyuan3D-1.0 received the highest user preference across 5 metrics.
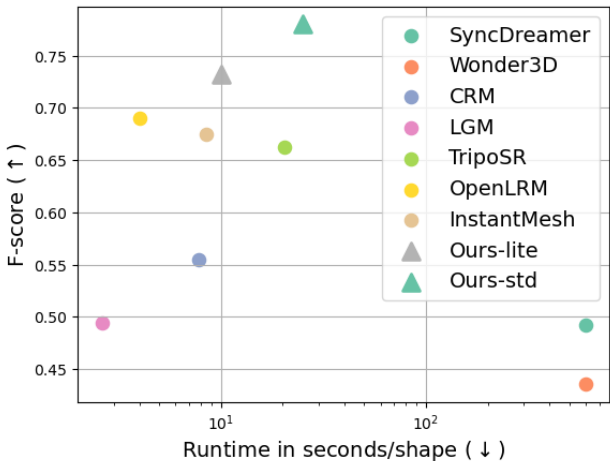

Figure 6. Performance vs Runtime. Our Hunyuan3D-1.0 balance the quality and efficiency well.

both in terms of CD and F-score metrics, achieving new state-of-the-art performance on this task.

Qualitative Comparisons. We present qualitative results of existing methods in Fig. 4. The figure illustrates that OperLRM [11] and TripoSR [43] struggle with geometric shapes, such as the soap and the box, and often generate blurred textures, as seen with the chair and the shoes. InstantMesh [20] captures more surface details but still exhibits some artifacts in certain areas, such as the seat of the chair, the logo on the cup, and the corners of the soap and box. In contrast, our model demonstrates superior reconstruction quality for both shape and texture. They not only capture the more accurate overall 3D structures of the
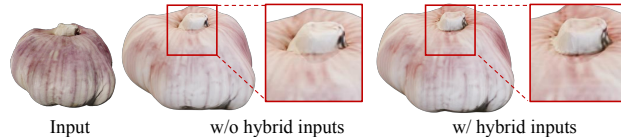

Figure 7. Adaptive CFG vs Fixed CFG.

Figure 8. Reconstruction on generated images only vs hybrid inputs.

objects but also excel in modeling intricate details. Our Hunyuan3D-1.0 received the highest user preference across 5 metrics as shown in 5.

Performance vs. Runtime. Another key advantage of Hunyuan3D-1.0 is its inference speed. The lite model takes around 10 seconds to produce a 3D mesh from a single image on an NVIDIA A100 GPU, while the standard model takes roughly 25 seconds. Note that these times do not include UV map unwrapping and texture baking, which takes approximately 15 seconds. Fig. 6 presents a 2D plot comparing our method to existing approaches, with inference times on the x-axis and the average F-Score on the y-axis. The plot demonstrates that Hunyuan3D-1.0 achieves an optimal balance between quality and efficiency.

6. Ablation Studies.

We single out the effectiveness of our proposed techniques, *i.e.*, adaptive CFG, and hybrid inputs to the generation speed and quality in this section.

Adaptive CFG. We verify the effectiveness of adaptive classifier-free guidance (CFG) on generated multi-view images in Fig. 7. Traditional fixed CFG throughout the denoising process often results in dark shadows in the back views. While the time-adaptive CFG introduced by Consistent123 [55] helps mitigate the shadow issue, it overlooks the relationships between views. In our camera orbit settings, the condition image has more area visible from the front view. The slow CFG would reduce the condition control for the front view, while the high CFG leaves excessive control over the back view, causing the model to replicate details from the front, such as the logo on the back of the cup. By dynamically adjusting the CFG during the generation process, we achieve a significant improvement in image quality. The adaptive CFG mechanism effectively prevents oversaturation and enables the model to generate more coherent and realistic multi-view images.

We evaluate the effectiveness of adaptive classifier-free guidance (CFG) on the generated multi-view images, as shown in Fig. 7. Traditional fixed CFG throughout the denoising process often tends to generate dark shadows in the back views. Although the time-adaptive CFG introduced by Consistent123 [55] helps mitigate this shadowing issue, it ignores the relationships between views. In our camera orbit settings, the condition image has a larger visible area from the front view. A low CFG reduces the control for the front view generation, while a high CFG exerts too much control over the back view generation, causing the model to replicate details from the front, such as the copied logo on the back of the cup. By dynamically adjusting the CFG during the generation process, we achieve a balance between controllability and diversity across different views, enabling the model to produce more coherent and realistic multi-view images.

Hybrid Inputs. The hybrid input technique was designed to enhance the reconstruction of unseen parts of 3D shapes. To evaluate its effectiveness, we compare the shapes generated w/o vs w/ hybrid input. As shown in Fig. 8, the generated garlic exhibits a flat top due to the lack of top-view information in our 0-evaluation orbit. By incorporating top-view information, the reconstruction model can accurately recover the dent around the garlic root. This demonstrates that the hybrid input approach significantly enhances the reconstruction accuracy of unseen regions and confirms that it produces more complete and accurate 3D shapes, especially in areas that are not directly visible in the generated views.

7. Conclusion.

This work introduces Hunyuan3D-1.0, a two-stage 3D generation pipeline capable of creating high-quality 3D shapes. The pipeline consists of a multi-view generation model that produces multi-view images rich in texture and geometry details and a feed-forward sparse-view reconstruction model that recovers the underlying 3D shape with explicit representations. We incorporate several innovative designs to enhance the speed and quality of the 3D generation process, including adaptive classifier-free guidance to balance the controllability and diversity for multi-view diffusion, hybrid inputs to address the unseen part reconstruction, and a lightweight super-resolution module to enhance the representation of details. Extensive evaluations on benchmark tasks demonstrate that Hunyuan3D-1.0 achieves state-of-the-art performance in 3D generation. Our method consistently outperforms existing approaches, highlighting its effectiveness in addressing the inherent challenges of 3D generation. These results validate the robustness and efficiency of our proposed pipeline, making substantial contributions to the 3D Generative community.

References

[1] Stablezero123. https://huggingface.co/stabilityai/stable-zero123. Accessed: 2024-02-22. 2

[2] M. Agrawal and L.S. Davis. A probabilistic framework for surface reconstruction from multiple images. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pages II–II, 2001. 3

[3] Jeremy S. De Bonet. Poxels: Probabilistic voxelized volume reconstruction. In *CVPR*, 1999. 3

[4] Mark Boss, Zixuan Huang, Aaryaman Vasishta, and Varun Jampani. Sf3d: Stable fast 3d mesh reconstruction with uv-unwrapping and illumination disentanglement. *arXiv preprint*, 2024. 2

[5] A. Broadhurst, T.W. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 388–393 vol.1, 2001. 3

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 5

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5

[8] Simon Donne and Andreas Geiger. Defusr: Learning non-volumetric depth fusion using successive reprojections. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 5

[10] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE international conference on computer vision*, pages 1586–1594, 2017. 3

[11] Zexin He and Tengfei Wang. Openlrm: Open-source large reconstruction models. https://github.com/3DTopia/OpenLRM, 2023. 2, 4, 5, 6, 7

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[13] 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. 2, 4

[14] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view

stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[15] Nianjuan Jiang, Zhaopeng Cui, and Ping Tan. A global linear method for camera pose registration. In *2013 IEEE International Conference on Computer Vision*, pages 481–488, 2013. 3

[16] Yash Kant, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, Igor Gilitschenski, and Aliaksandr Siarohin. Spad : Spatially aware multiview diffusers, 2024. 3

[17] K.N. Kutulakos and S.M. Seitz. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 307–314 vol.1, 1999. 3

[18] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *European Conference on Computer Vision*, 2018. 3

[19] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 3

[20] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3, 4, 7

[21] 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*, 2024. 3

[22] 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*, 36, 2024. 2

[23] 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*, pages 9298–9309, 2023. 2, 3, 4

[24] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. DIST: Rendering Deep Implicit Signed Distance Function With Differentiable Sphere Tracing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 3

[25] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3, 4, 5, 6

[26] 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. *arXiv preprint arXiv:2310.15008*, 2023. 2, 3, 5, 6

[27] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, 1987. 5

[28] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2.5: Diverse text-to-3d generation via multi-view 2.5d diffusion. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8744–8753, 2023. 3

[29] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5695–5703, 2016. 3

[30] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[31] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[32] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[33] Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[34] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[35] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. In *2017 International Conference on 3D Vision (3DV)*, pages 57–66. IEEE, 2017. 3

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4

[37] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 3

[38] S.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1067–1073, 1997. 3

[39] 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, 2023. 3, 4

[40] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3, 4

[41] 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. *arXiv preprint arXiv:2402.05054*, 2024. 5, 6

[42] Shitao Tang, Fuayng Zhang, Jiacheng Chen, Peng Wang, and Furukawa Yasutaka. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint 2307.01097*, 2023. 3

[43] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, , Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024. 2, 5, 6, 7

[44] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 3

[45] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2634, 2017. 3

[46] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 3

[47] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii 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 (ECCV)*, 2024. 4

[48] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 4

[49] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 2, 3, 5

[50] 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. 3

[51] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[52] 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. *arXiv preprint arXiv:2403.05034*, 2024. 5, 6

[53] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models, 2022. 3, 4

[54] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 5

[55] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 4, 7, 8

[56] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 5

[57] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, 2022. 2

[58] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models, 2024. 2, 5, 6

[59] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. *arXiv preprint arXiv:2311.09217*, 2023. 2

[60] Xianghui Yang, Yan Zuo, Sameera Ramasinghe, Loris Bazzani, Gil Avraham, and Anton van den Hengel. Viewfusion: Towards multi-view consistency via interpolated denoising, 2024. 2

[61] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. *European Conference on Computer Vision (ECCV)*, 2018. 3

[62] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, 2018. 2

[63] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[64] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 3

[65] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 3

[66] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *CVPR*, 2020. 3

[67] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015. 3

[68] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. 5