HexaGen3D: StableDiffusion is One Step Away from Fast and Diverse Text-to-3D Generation (Supplementary Material)

Antoine Mercier Ramin Nakhli* Mahesh Reddy Rajeev Yasarla
Hong Cai Fatih Porikli Guillaume Berger
Qualcomm AI Research†
{amercier, mahkri, ryasarla, hongcai, fporikli, guilberg}@qti.qualcomm.com

1. More Implementation Details

Objaverse Dataset Filtering In line with the methodologies proposed in [3, 5], we curate the Objaverse dataset to improve the quality of 3D assets considered during training. We used the following filtering criteria:

- *Number of Faces.* To ensure a focus on single-object generations, we selected assets comprising fewer than 400,000 triangles, effectively filtering out complex scenes containing multiple objects.
- *Number of Geometries (Mesh Parts).* Similarly, we limited our selection to objects containing fewer than 200 distinct geometries.
- *Presence of Additional Texture Maps.* Assets lacking a metallic-roughness texture map were excluded. This decision was based on our finding that such textures are usually indicative of higher-quality 3D models.

While these filtering criteria served our research well, we recognize the possibility of further enhancements. We thus direct readers interested in more sophisticated curation methods to the approach detailed in [5].

VAE losses We follow 3DGen and combine a mask silhouette loss, a depth loss, a laplacian smoothness loss and a KL divergence loss to supervise the geometry VAE: $L_{geometry} = \alpha L_{mask} + \phi L_{depth} + \lambda L_{smooth} - \gamma D_{KL}$, with $\alpha = 3$, $\phi = 10$, $\lambda = 0.01$, and $\gamma = 10^{-7}$. We use a sum of L1 and L2 losses to supervise the color VAE.

VAE optimization hyperparameters We train our VAE models for 15 epochs, using a batch size of 16, an initial learning rate of 3×10^{-5} and a cosine annealing schedule down to a minimum learning rate of 10^{-6}. This takes roughly a week on 8 A100s.

Stage 2 optimization hyperparameters We use a batch of 192 hexaviews for the hexaview diffusion and hexaview-to-triplanar mapping tasks, and a batch of 32 regular images for 2D regularization. We down-weigh the diffusion loss coming from the 2D regularization batches by a factor 0.25. We train our latent generative models for $50,000$ iterations using a learning rate of 3×10^{-5}. This takes roughly 4 days on 8 A100s.

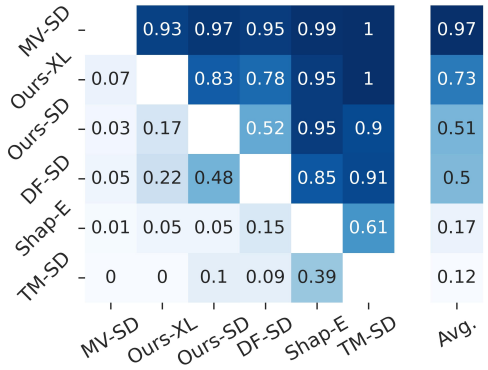
Baselines Implementation We select a range of recent text-to-3D approaches for comparison with HexaGen3D, including a feedforward approach, Shap-E [4], and three SDS-based approaches, DreamFusion [8], TextMesh [11], and MVDream [10]. While we use the official implementation of Shap-E* and MVDream†, we leverage the DreamFusion and TextMesh implementations available within the threestudio framework [2], which include some notable deviations from their original papers. Specifically, our DreamFusion setup uses the open-source StableDiffusion v2.1 as the guidance model instead of Imagen [9], and a hash grid [7] as the 3D representation, diverging from Mip-Nerf360 [1]. TextMesh similarly utilizes StableDiffusion v2.1 for guidance, with NeuS [12] replacing VolSDF [13] for the 3D signed distance field representation.

2. Evaluation of 3D Generations Through User Studies

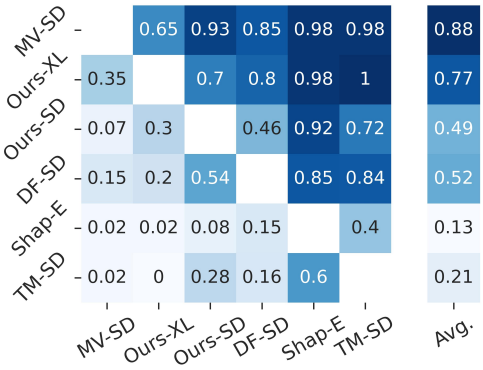
In our investigation aimed at assessing the visual quality and adherence to prompts of text-to-3D generations, we utilized 67 prompts sourced from DreamFusion [8]. For each

*Work done at Qualcomm AI Research during an internship.
†Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

*https://github.com/openai/shap-e
†https://github.com/bytedance/MVDream-threestudio



(a) Evaluating the approaches for "visual quality".



(b) Evaluating the approaches on "text prompt fidelity".

Figure 1. User study comparing all the text-to-3D approaches on (a) visual quality and (b) text prompt fidelity. Each cell indicates the user preference score (%) for an approach (row) over another (column). The approaches are: MVDream-SDv2.1 (MV-SD), DreamFusion-SDv2.1 (DF-SD), Shape-E, TextMesh-SDv2.1 (TM-SD), HexaGen3D-SDv1.5 (Ours-SD), and HexaGen3D-SDXL (Ours-XL).

prompt, across various methods, we generated four images with a rendering angle of 45° azimuth and 30° elevation. The evaluated methods included MVDream-SDv2.1 [10], TextMesh-SDv2.1 [11], DreamFusionv2.1 [8], Shap-E [4], HexaGen3D-SDv1.5, and HexaGen3D-SDXL. Participants were involved in pairwise evaluations of these methods, focusing on either visual quality or prompt fidelity. For each evaluation, participants were presented with two anonymized sets of four renderings from the same prompt, generated by two different methods (A and B), and asked to express their preference using a five-point scale: "A is significantly better", "A is slightly better", "No Preference", "B is slightly better", and "B is significantly better". The user responses were binarized to compute the final metrics.

User Study 1: Visual Quality Assessment. The primary focus of the first user study was on the visual quality of the 3D renderings. Twelve participants were instructed to evaluate and compare the visual quality of the outcomes produced by the different methods. Each participant assessed average 90 random result pairs, generating more than 360 data points per method and on average 72 comparisons between each possible method pair.

User study 2: Prompt Fidelity. The second user study focused on how well the generated 3D assets aligned with the textual prompts provided by users. In this phase, nine participants were asked to evaluate the outcomes based on their adherence to the original text prompts.

The comprehensive results of these studies are illustrated in Fig. 1 where we report the pair-wise preference score of any method (rows) over any other method (columns).

3. Extended Results from HexaGen3D-SDXL

More Results from MS-COCO Prompts. Leveraging a 2D pre-trained diffusion model significantly enhances generalization to uncommon objects or combinations thereof, which were not encountered during the fine-tuning phase. This allows HexaGen3D to handle a broad range of textual prompts effectively. To illustrate this capability, we present additional results using random captions from the MS-COCO dataset [6], showcased in Figs. 2 and 3.

Intermediate Hexaview Visualizations. We visualize the six-sided orthographic projections generated by HexaGen3D-SDXL for a variety of prompts from MS-COCO [6] prompts along with the corresponding final 3D object in Fig. 4. The generated hexaviews are highly detailed and multi-view consistent.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5470–5479, 2022. 1

[2] Yuan-Chen Guo, Ying-Tian Liu, Ruizhi Shao, Christian Laforte, Vikram Voleti, Guan Luo, Chia-Hao Chen, Zi-Xin Zou, Chen Wang, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation. https://github.com/threestudio-project/threestudio, 2023. 1

[3] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. arXiv preprint arXiv:2303.05371, 2023. 1

[4] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463, 2023. 1, 2

[5] 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. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 2, 4, 5

[7] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics (ToG), 41(4):1–15, 2022. 1

[8] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988, 2022. 1, 2

[9] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems, 35:36479–36494, 2022. 1

[10] 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. 1, 2

[11] Christina Tsalicoglou, Fabian Manhardt, Alessio Tonioni, Michael Niemeyer, and Federico Tombari. Textmesh: Generation of realistic 3d meshes from text prompts. arXiv preprint arXiv:2304.12439, 2023. 1, 2

[12] 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. arXiv preprint arXiv:2106.10689, 2021. 1

[13] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems, 34:4805–4815, 2021. 1

Figure 2. HexaGen3D-SDXL generations using random captions from the MSCOCO [6] dataset.

Figure 3. HexaGen3D-SDXL generations using random captions from the MSCOCO [6] dataset.

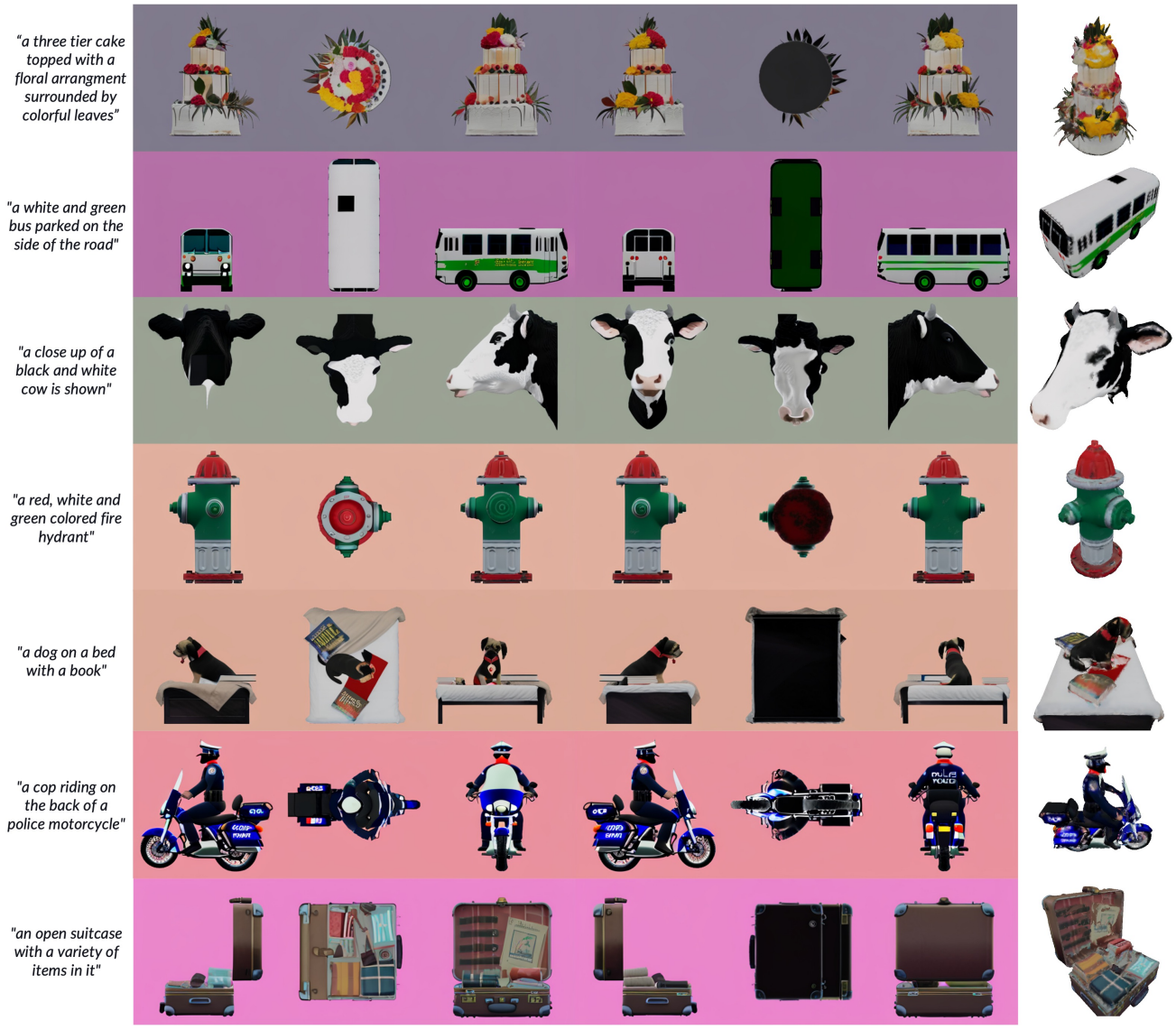


Figure 4. HexaGen3D produces six detailed and consistent orthographic projections as an intermediate step to 3D object generation.