Appendix

This supplementary material provides details regarding the implementation and additional results of the proposed Depth Guided Branched Diffusion (DGBD) method for controllable multi-view generation to showcase its effectiveness. Appendix A offers training and inference specifics. Additional visual results of the perspective branch of the proposed DGBD method are presented in Appendix B. Furthermore, additional results of the DGBD framework are demonstrated in Appendix C. In Appendix D we show additional results regarding the depth control scale of our framework.

A. Implementation details

The first branch of the proposed DGBD pipeline is fine-tuned from Stable-Diffusion-v1-5 after applying the corresponding modifications related to the self-attention mechanism and geometry control to the U-Net. The regular self-attention blocks are replaced with the proposed regularized batch-aware self-attention (RBA) modules to ensure multi-view consistency and a perceptive projection layer is added to inform the model about the geometry. The text embedding is concatenated with the given camera location and transformed to the dimensionality of the CLIP [3] embedding space (768) through a perspective projection layer. The camera location is encoded as in Zero-1-to-3 [1] and the perceptive projection layer is a linear layer ($772 \rightarrow 768$). The same prompt is used for the views of the same object. The model is trained for 18K iterations using a total batch size of 3072 on 8×A6000-40GB using the AdamW [2] optimizer with a learning rate of 10^{-4}. The training takes approximately a week. The training is conducted on image resolution 256×256 to achieve faster convergence similar to Zero-1-to-3. The second branch of DGBD is trained for 8K iterations (about four days) with the same hyper-parameters and machine with AdamW optimizer using 10^{-5} learning rate. At inference, p of the proposed RBA block is set to one. In all our experiments, the DDIM [4] scheduler is used with 50 denoising steps and a guidance scale of 7.5.

B. Qualitative results of the Perspective Branch

Fig. 1 demonstrates results of the perspective branch of the DGBD pipeline where the depth control scale is set to zero to conduct pure multi-view generation. The generated views align with the given text prompt, have high-quality details, and showcase light and reflections on the object. Moreover, the generated images are consistent across views.

C. Qualitative results of the DGBD method

Additional results of the DGBD framework are provided in Fig. 2. Given a textual caption, a depth map from one view, and the camera location of other views, the proposed method generates output views conforming to the prompt, transferring shape and size features from the depth map and aligning with the given perspective. The generated views are high-fidelity with great detail, light, and shadows.

D. Qualitative results of the influence of depth control scale

In Fig. 3, additional results regarding the depth control scale of the DGBD pipeline are displayed. Given a prompt, a depth map from one view, and perceptive information of another view, the method generates a view aligning with the caption, propagating structural information from the depth map and conforming with the viewpoint. The value of the depth control scale determines the strength of the propagation of shape and size attributes from the input depth map. Increasing the depth control scale increases the strength of the propagation level of the structural features from the depth map and produces results that carry more information regarding shape and size from the depth map while setting it to zero mandates the method to conduct pure multi-view generation.

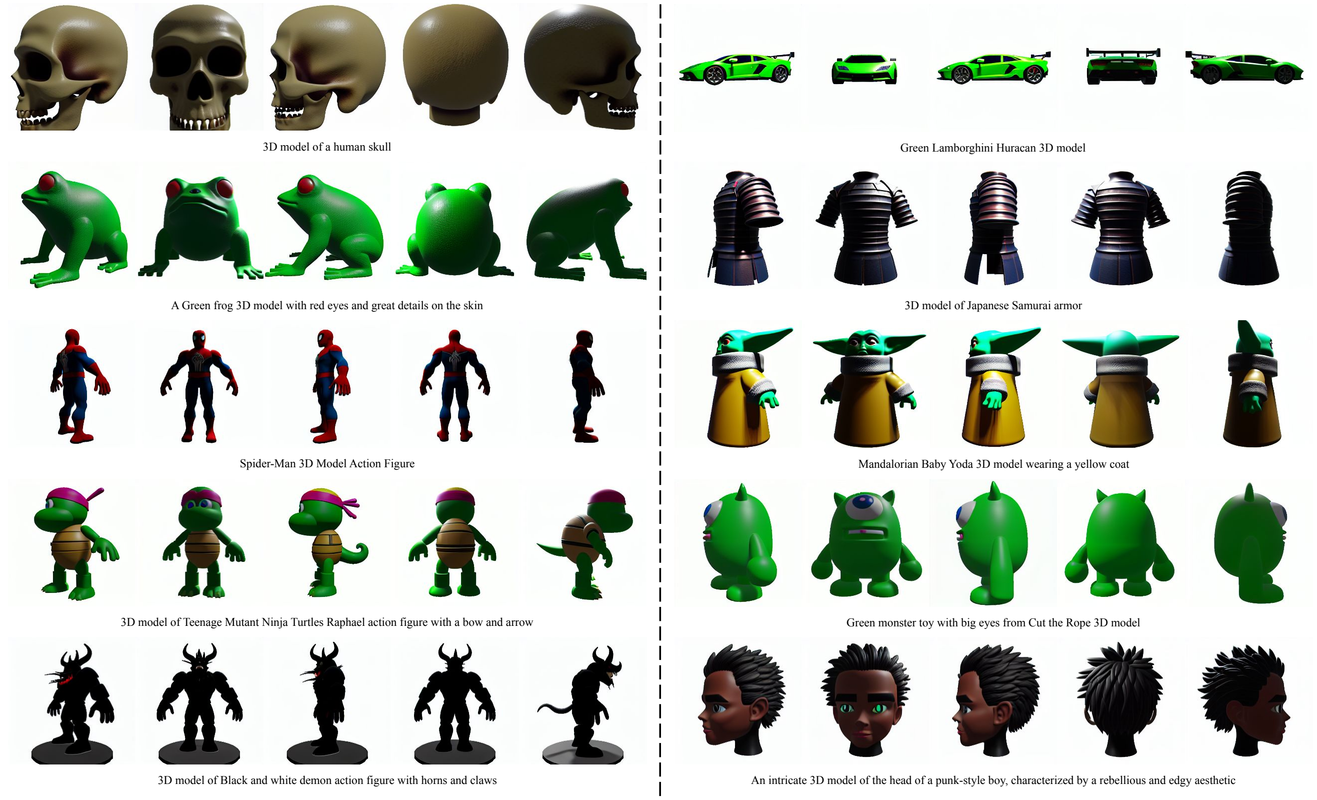
References

[1] 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 (ICCV)*, pages 9298–9309, 2023. 1

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 1



3D model of a human skull

A Green frog 3D model with red eyes and great details on the skin

Spider-Man 3D Model Action Figure

3D model of Teenage Mutant Ninja Turtles Raphael action figure with a bow and arrow

3D model of Black and white demon action figure with horns and claws

Green Lamborghini Huracan 3D model

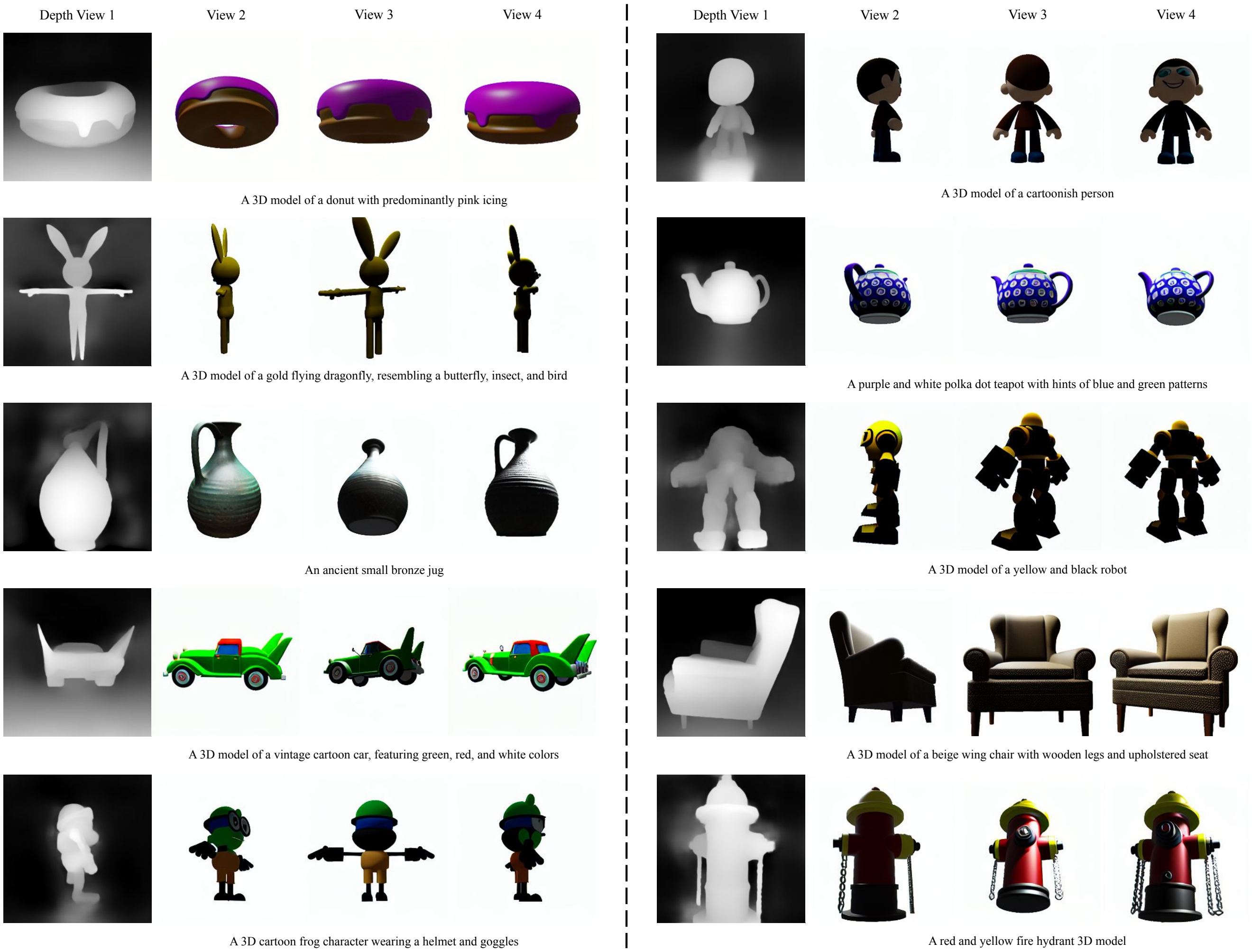
3D model of Japanese Samurai armor

Mandalorian Baby Yoda 3D model wearing a yellow coat

Green monster toy with big eyes from Cut the Rope 3D model

An intricate 3D model of the head of a punk-style boy, characterized by a rebellious and edgy aesthetic

Figure 1. Visual results of the perspective branch of Depth Guided Branched Diffusion (DGBD) pipeline on different prompts on a validation set from Objaverse.



A 3D model of a donut with predominantly pink icing

A 3D model of a cartoonish person

A 3D model of a gold flying dragonfly, resembling a butterfly, insect, and bird

A purple and white polka dot teapot with hints of blue and green patterns

An ancient small bronze jug

A 3D model of a yellow and black robot

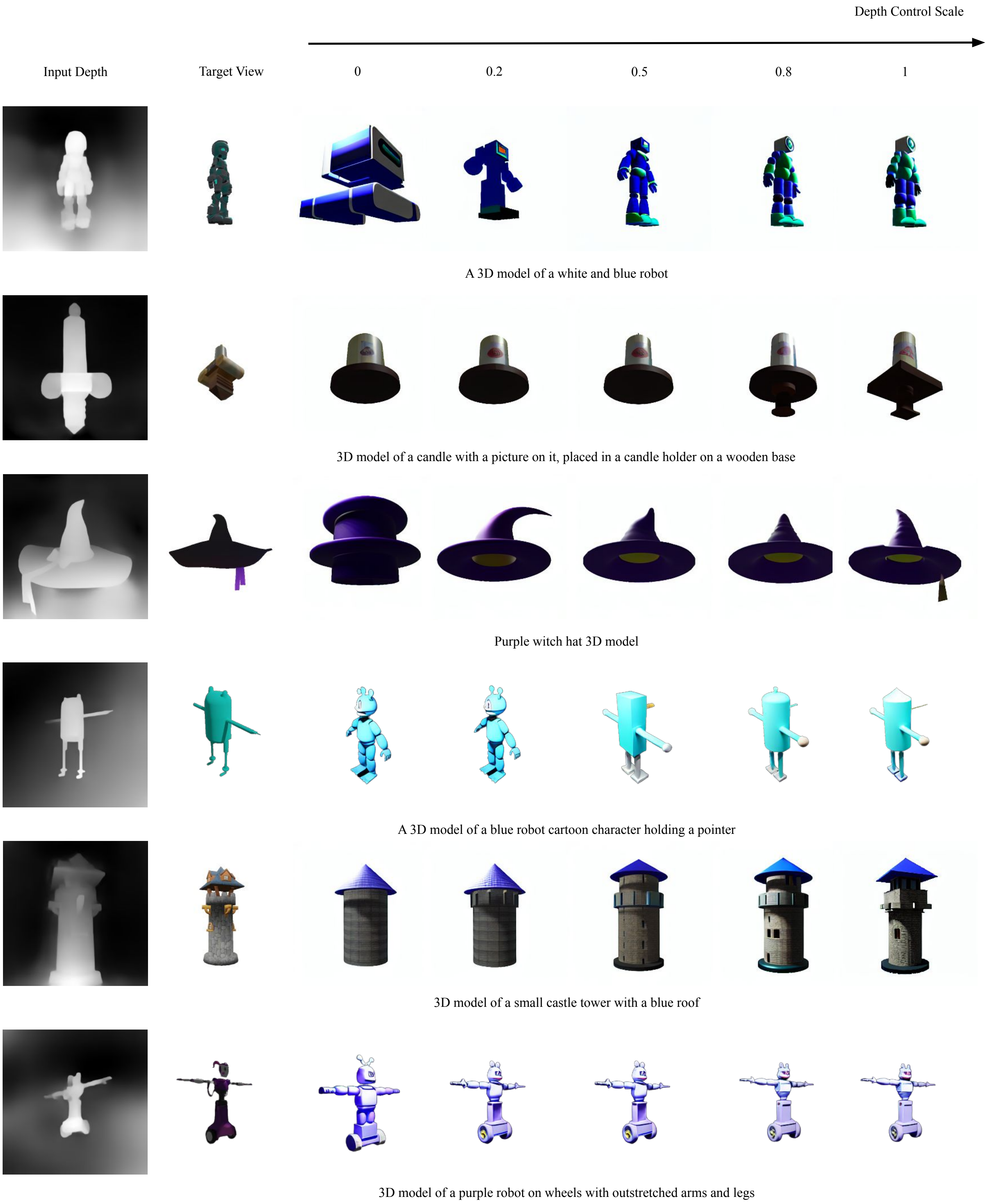
A 3D model of a vintage cartoon car, featuring green, red, and white colors

A 3D model of a beige wing chair with wooden legs and upholstered seat

A 3D cartoon frog character wearing a helmet and goggles

A red and yellow fire hydrant 3D model

Figure 2. Results of Depth Guided Branched Diffusion (DGBD) framework on a validation set from Objaverse.



A 3D model of a white and blue robot

3D model of a candle with a picture on it, placed in a candle holder on a wooden base

Purple witch hat 3D model

A 3D model of a blue robot cartoon character holding a pointer

3D model of a small castle tower with a blue roof

3D model of a purple robot on wheels with outstretched arms and legs

Figure 3. Results of the influence of depth control scale in Depth Guided Branched Diffusion (DGBD) method on a validation set from Objaverse.