

ControlDreamer: Blending Geometry and Style in Text-to-3D

- Supplementary Material

A Additional Experiments

A.1 Qualitative results

In Figs. A.1 and A.2, we present a variety of qualitative results used for evaluating directional CLIP similarities and conducting human evaluations.

A.2 Ablation Studies

A.2.1 Exploring 3D Representations in Style Stages

Fig. A.3 demonstrates the approach of refining 3D models using NeRF instead of DMTet during the style stage. Volume rendering with NeRF often results in significant alterations to the overall geometry, particularly when geometry and style texts are semantically different, leading to instability in depth-aware score distillation. Conversely, using DMTet, coupled with our timestep annealing approach, is more effective. It ensures appropriate alterations in the overall texture while maintaining the geometry aligned with the source mesh, leading to more seamless style generation. For NeRF-based 3D stylization, we sample diffusion timesteps from a range of $[0.02, 0.5]$ to minimize significant geometry changes. MV-ControlNet's tendency to overlook camera conditions often leads to multiple artifacts, further substantiating the advantage of using DMTet in the style stage.

A.2.2 Advantages of MV-ControlNet in multi-view image generation

Fig. A.4 illustrates the advantages of the proposed MV-ControlNet in 2D image generation. In panel (A), conventional methods exhibit significant semantic variations due to random seed and camera parameter changes, often deviating from the original identity of the source image. In contrast, our proposed method demonstrates stronger alignment with the depth condition, ensuring consistent, object-centric generation quality across different camera parameters and random seeds.

In panel (B), when handling text with complex concepts, the baseline method displays stereotypical and geometry-biased generations that compromise image quality, as discussed earlier in Fig. 1 of the main text. Conversely, our method consistently produces high-quality

Geometry Stage

Style Stage



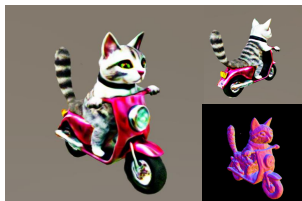
Nike Jordan shoe



Leather shoe



Flower shoe



Cat riding a scooter



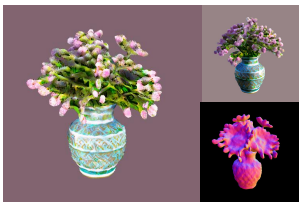
Tiger



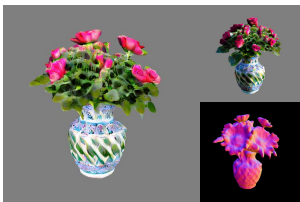
Panda



Sunflowers in a vase



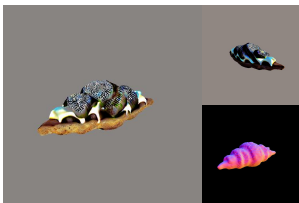
Cherry blossoms



Roses



Delicious croissant



Oreo



Red velvet



Astronaut riding a horse



Knight



Gandalf riding a donkey

Figure A.1: Our two-stage 3D generation pipeline first involves generating a coarse-grained geometry, followed by the generation of a fine-grained stylized 3D model using a style prompt.

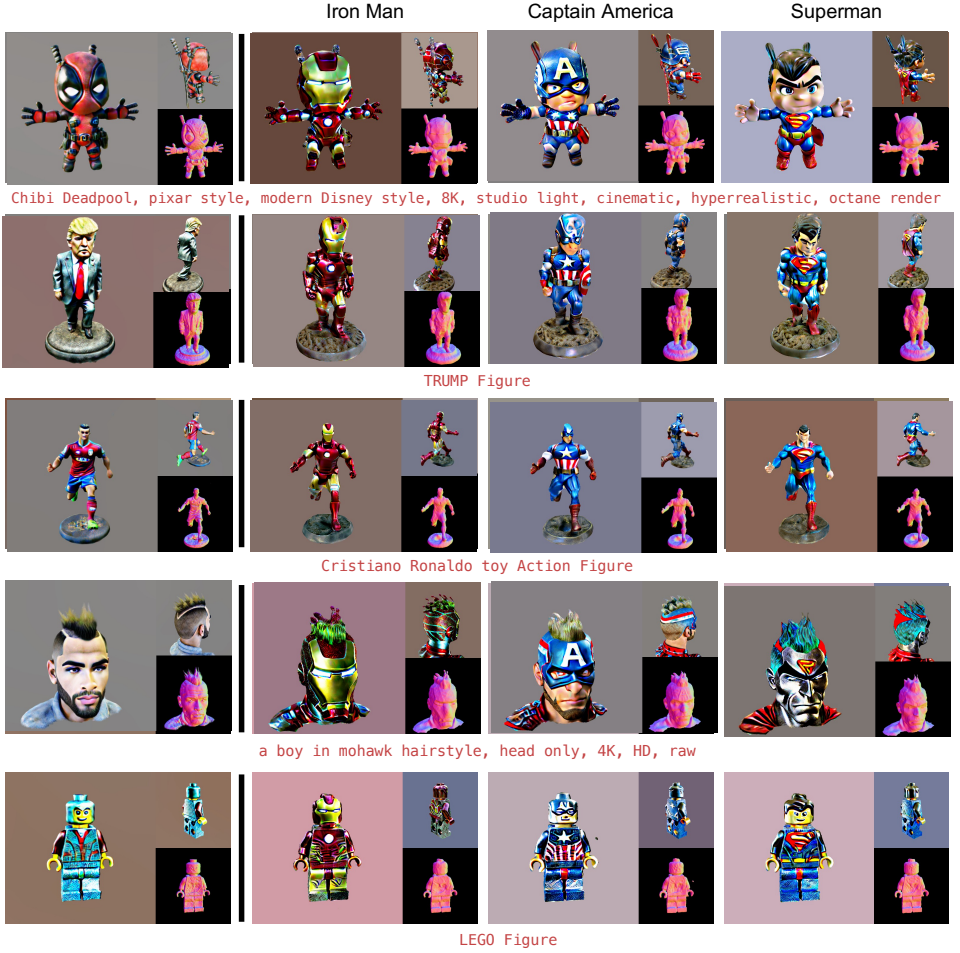


Figure A.2: Additional qualitative results. We generate a variety of styles on a range of source 3D models, demonstrating the versatility of our method.

images that align closely with the provided text, preserving key identities from the source, such as the backpack of an astronaut.

In summary, MV-ControlNet is firmly grounded in depth conditions and adept at incorporating complex textual concepts into the generation of multi-view images.

A.3 Effectiveness of Prompt Engineering for 3D Generation

When using user-defined text prompts that were not included in the training set of the MV-Dream model, we observe the generation of artifacts in the learned source geometry, as illustrated in Fig. A.6. However, we find that careful prompt engineering can effectively mitigate these artifacts. An example of such refined text in our experiments is ‘Hulk, muscular, green, 4K, photorealistic’.

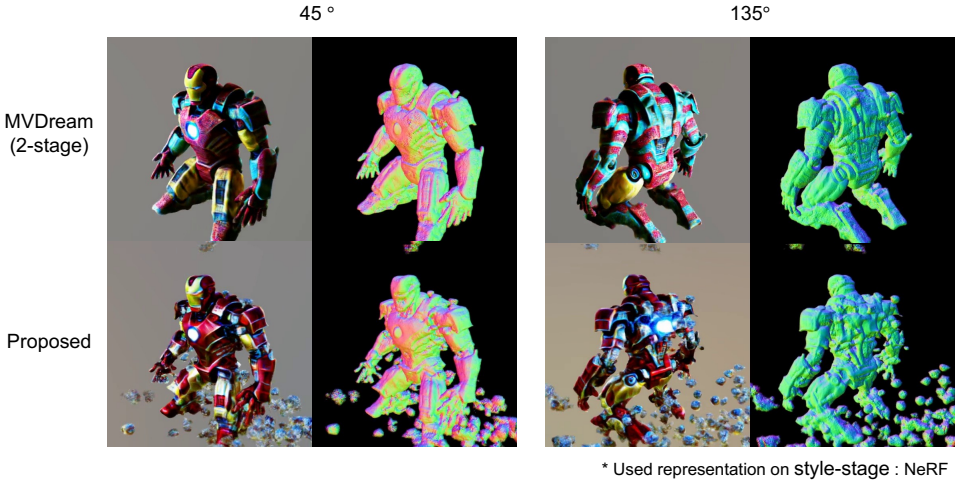


Figure A.3: We visualize the results for both the two-stage MVDream and our ControlDreamer method in the style stage using NeRF representation. Notably, with NeRF as the representation, the geometry undergoes significant changes. This substantial alteration in geometry and depth information during NeRF training leads to instability in our depth-aware score distillation process, resulting in the generation of gravel-like artifacts in our results.

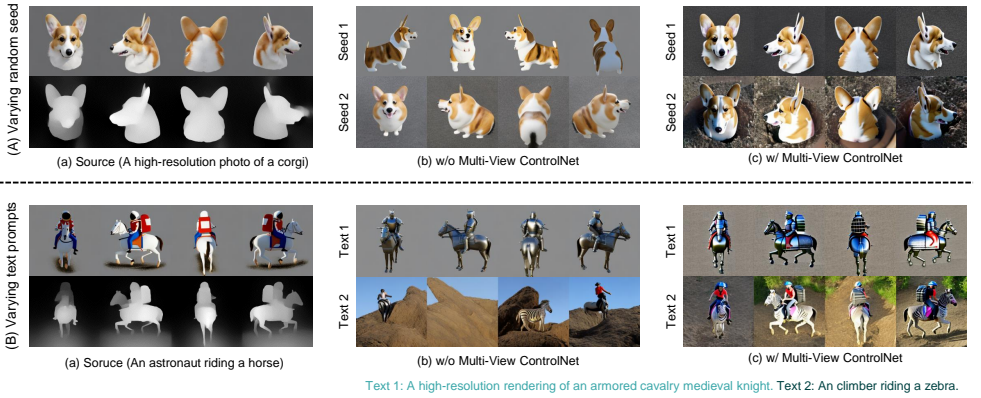


Figure A.4: (A) The first column displays source multi-view images along with their corresponding depths. We present a comparison of images generated by MVDream in the second column and MV-ControlNet in the third column, originating from two distinct initial noises. MV-ControlNet's depth-aware generation consistently preserves the source's geometry and effectively manages variations in initial noise. (B) Our model excels at generating unique images, such as a knight equipped with a backpack, and adeptly handles imaginative prompts like 'climber zebra'. It consistently produces high-quality images, outperforming MVDream, which often struggles with such complex tasks.

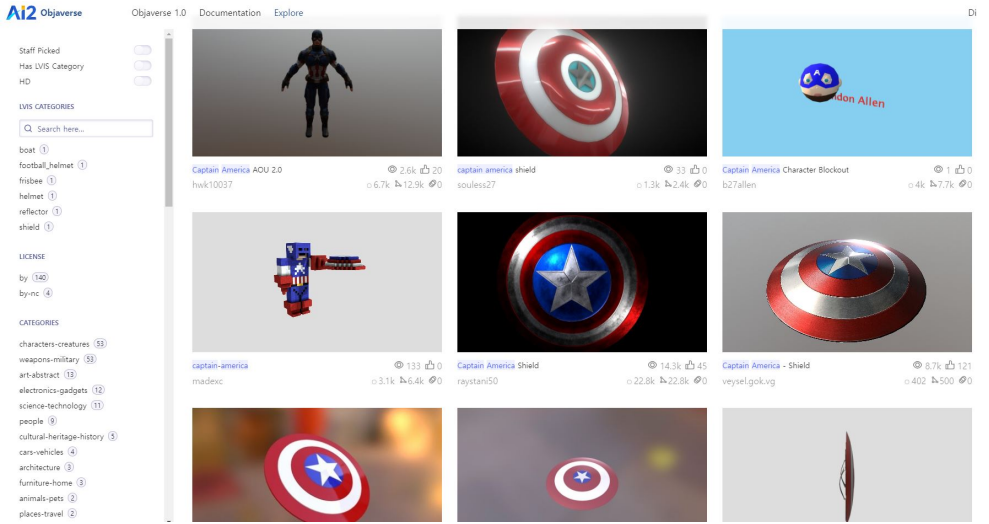


Figure A.5: Intriguingly, a search for ‘Captain America’ on <https://objaverse.allenai.org/explore> primarily yields content related to shields.

A.4 Geometry bias in Objaverse dataset

We include further results on the geometry bias. Fig. A.5 provides a straightforward visualization of text-based 3D model retrieval on the Objaverse dataset, showcasing the frequency of certain geometries. Notably, as we highlight in the main text, 3D models predominantly associated with a shield are frequently generated from the prompt ‘Captain America’.

A.5 Enhanced Stylization in the Image-to-3D Framework

Our methodology demonstrates exceptional stylization capabilities, independent of the initialized geometry. Building on this strength, we have extended our experiments to include the image-to-3D framework as described in ImageDream [2], to further evaluate its editing capabilities. On 3D models generated using ImageDream, we applied our MV-ControlNet to modify the style. The results, depicted in Fig. B.1, show the adaptability of our approach to the image-to-3D generation.

A.6 Multi-view image editing

Prompt-to-Prompt (P2P) [2] is an image editing method that aligns the geometries of source and target images by injecting attention maps into diffusion models. For our qualitative comparison in the main text, injection ratios were set at 0.4 for cross-attention and 0.8 for self-attention to ensure strong geometric alignment. However, P2P is somewhat constrained by the requirement for equal token lengths in both source and target texts, which often results in less aligned text-guided editing outcomes. Conversely, our proposed MV-ControlNet, which utilizes depth conditions to maintain source geometry, is not restricted by text length, allowing for more creative and diverse prompts in the style stage.

In Fig. A.7, we showcase editing results of multi-view images using various prompts.

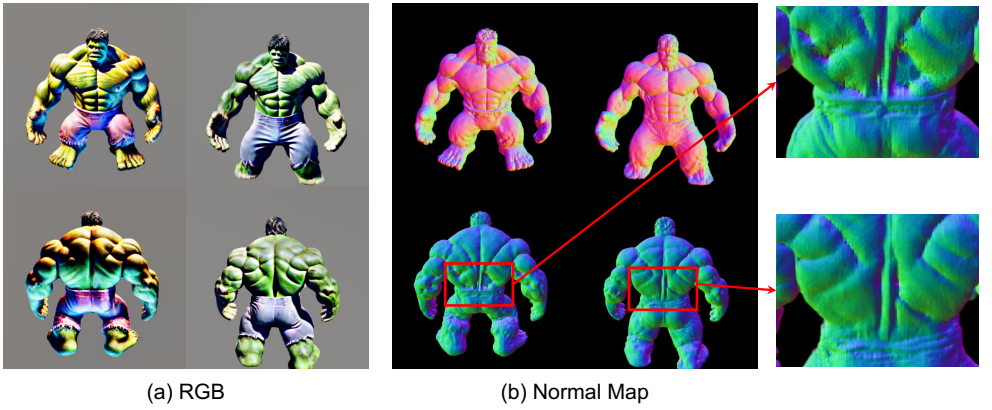


Figure A.6: The comparison showcases 3D models generated from user-provided texts versus those recommended by ChatGPT. Notably, models from user texts may exhibit artifacts, particularly in normal maps, whereas models from ChatGPT’s refined texts generally result in higher-quality 3D models.

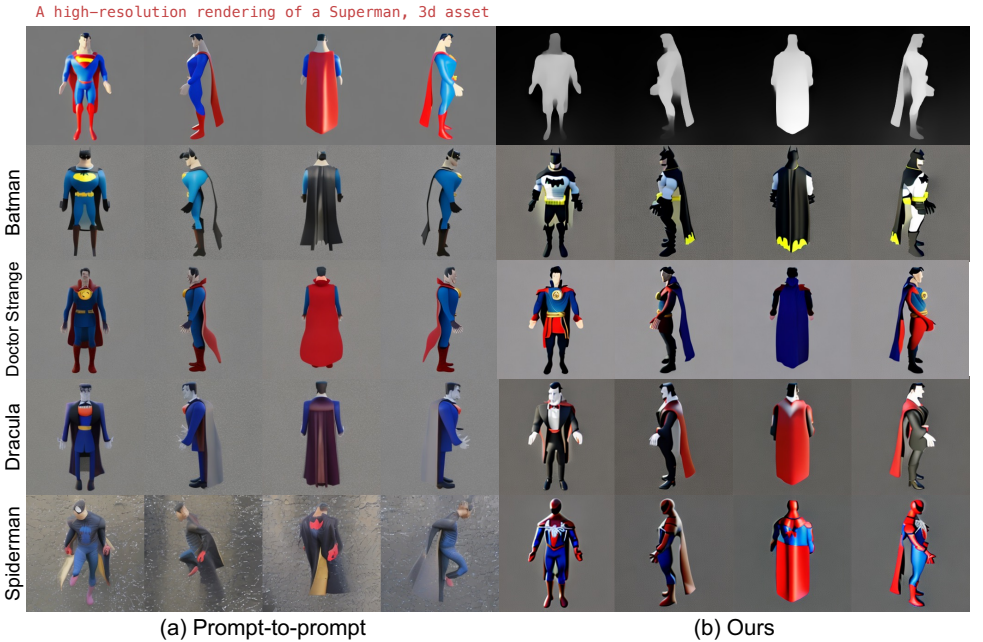


Figure A.7: The comparative results on multi-view image editings. Our method consistently generates more photorealistic results than P2P [9].

Using ‘Superman’ as a geometry prompt, we added a cloak to the initial geometry. The P2P method, which relies on cross-attention control, struggles with photorealistic images for complex scenarios like Spiderman wearing a cloak. Conversely, our MV-ControlNet incorporates depth conditions, achieving better style alignment and consistently producing high-

Table B.1: Comprehensive descriptions of the utilized geometry and style prompts.

Domain	Geometry texts	Style texts
Animals	A bald eagle carved out of wood A high-resolution photo of a frog, 3d asset A cute shark, plush toy, ultra realistic, 4K, HD A cute shark, plush toy, ultra realistic, 4K, HD A high-resolution rendering of a Shiba Inu, 3d asset A high-resolution photo of a British Shorthair, 3d asset	A highly-detailed rendering of a magpie, ultra hd, realistic, vivid colors A high-resolution photo of a toad, 3d asset A highly-detailed photo of a tuna, 4k, HD A highly-detailed rendering of a killer whale, 4k, HD A white tiger, super detailed, best quality, 4K, HD A highly-detailed photo of a lioness, 4K, HD
Character	A high-resolution rendering of a Hulk, 3d asset A high-resolution rendering of a Hulk, 3d asset A high-resolution rendering of a Hulk, 3d asset A high-resolution rendering of a Hulk, 3d asset A high-resolution rendering of a Hulk, 3d asset	A highly-detailed 3d rendering of a Superman A high-resolution rendering of an Iron Man, 3d asset A highly-detailed photo of a Spider-Man, 4K, HD Renaissance sculpture of Michelangelo's David, Masterpiece A highly-detailed 3d rendering of Thanos wearing the infinity gauntlet A highly-detailed 3d rendering of a Captain America
Foods	A delicious croissant, realistic, 4K, HD A delicious croissant, realistic, 4K, HD A delicious croissant, realistic, 4K, HD wedding cake, 4k wedding cake, 4k wedding cake, 4k	A delicious red velvet cake, realistic, 4K, HD A delicious pecan pie, realistic, 4K, HD A delicious oreo cake, realistic, 4K, HD stack of macarons, 4k strawberry cake, 4k cheeseburger, 4k
General	a cat riding a scooter like a human a cat riding a scooter like a human a cat riding a scooter like a human An astronaut riding a horse An astronaut riding a horse An astronaut riding a horse	panda riding a scooter like a human, 4k tiger riding a scooter like a human, 4k corgi riding a scooter like a human, 4k A high-resolution rendering of an armored cavalry medieval knight, 3d asset Gandalf riding a donkey, fairy tale style, 4K, HD A beautiful portrait of a princess riding an unicorn, fantasy, HD
Objects	A cute shark, plush toy, ultra realistic, 4K, HD Battletech Zeus with a sword!, tabletop, miniature, battletech, miniatures, wargames, 3d asset Battletech Zeus with a sword!, tabletop, miniature, battletech, miniatures, wargames, 3d asset Nike Jordan shoe, 4k Nike Jordan shoe, 4k 3d rendering of a mug of hot chocolate with whipped cream	A DSLR photo of a submarine, 4k, HD Optimus Prime fighting, super detailed, best quality, 4K, HD A high-resolution rendering of a Gundam, 3d asset leather shoe, 4k shoe with flower printed on it, 4k A highly-detailed photo of a mug with Starbucks logo, 4K, HD

quality, text-aligned images.

A.7 Failure cases

In Fig. B.2, we present examples of our failure cases. We empirically observe that when the style prompt significantly deviates from the initial geometry, our method struggles to generate visually appealing 3D models. This limitation likely arises from the pre-trained MVDream, which has insufficient capabilities to seamlessly integrate two distinct concepts.

B Experiment Details

B.1 Baselines

We utilize the threestudio [1] implementation of Magic3D [2], Fantasia3D [3], and ProlificDreamer [4] as our baselines. Similar to ControlDreamer, all baselines involve a refinement process on textured meshes derived from the pre-trained MVDream [5] using the DM Tet [6] algorithm. Specifically, Fantasia3D employs a two-stage process, focusing on appearance modeling in the second stage to modify the textured mesh. For ProlificDreamer, we engage in a textured mesh fine-tuning stage for stylization. Additionally, although ProlificDreamer does not utilize a separate shading model, we apply shading to ensure fair comparisons with other baselines. Our paired geometry and style prompts are detailed in Table B.1.

B.2 Generated images used for MV-ControlNet training

In Fig. B.3 and Fig. B.4, we present additional examples used in training our MV-ControlNet. Both figures illustrate samples created through random sampling, with the camera views for each sample also randomized. These figures clearly show that samples from a filtered 100K text corpus feature a wide range of 3D assets, such as objects, animals, and characters. We used these high-quality multi-view images to train MV-ControlNet.

B.3 Human evaluations

Our human evaluation spans paired prompts from five domains, with each domain involving six prompts across five methodologies: Fantasia3D, Magic3D, ProlificDreamer, MVDream, and our approach. The evaluations assess editability and text alignment, as detailed in the table below.

Human Evaluation Template

Your task is to evaluate the quality of 4.0-second-long food video clips. Each clip showcases the results of editing 3D models using text prompts.
Please choose the most favorable video based on the following criteria:

- 1) **Overall editing quantity:** Has the edited video completely changed in visual appearance compared to the Source?
- 2) **Textual alignment:** How well does the text description align with the edited video?

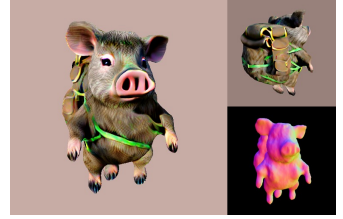
References

- [1] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, October 2023.
- [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.
- [3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023.
- [4] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.
- [5] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.

Image Prompt



Pig to Wild Boar



A pig wearing a back pack

Image Prompt



Crab to Coconut crab toy



A crab, low poly

Figure B.1: Visualization of our stylization process within the image-to-3D pipeline. The source image prompts chosen for this demonstration were the subjects ‘pig’ and ‘crab’. Across these distinct subjects, our method maintained robust editing performance, showcasing its consistency in 3D stylization.

- [6] 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.
- [7] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.
- [8] 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, 2023.

A 3D rendering of Winnie the Pooh, photorealistic, 8K, HD



A highly-detailed cartoon of a Doraemon, 4K, HD



Figure B.2: We visualize failure cases of ControlDreamer. In these examples, we use the Hulk geometry from Fig. 1 of the main manuscript. We observe that when the stylization text significantly deviates from the source geometry, it results in the generation of visually unappealing 3D models.

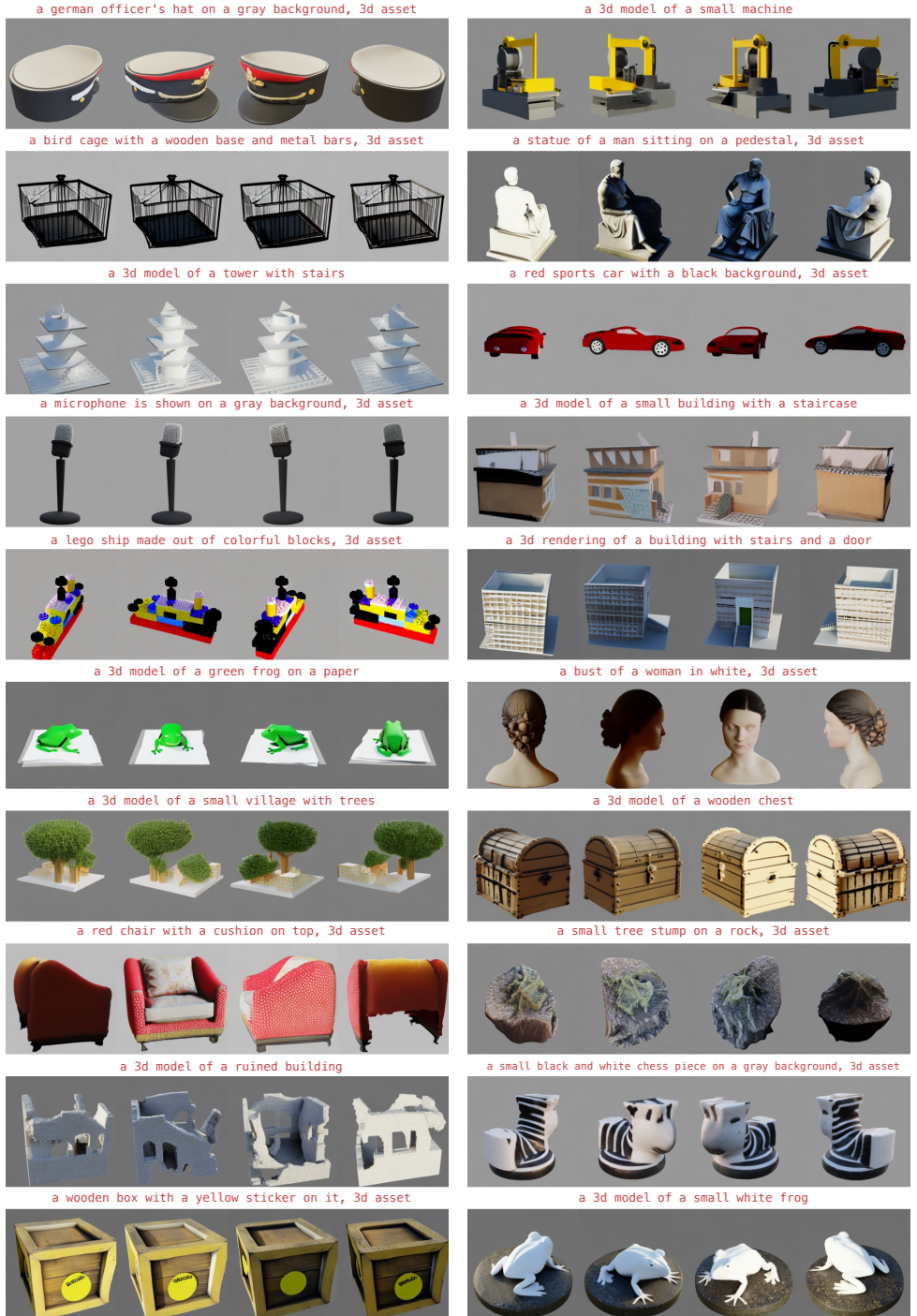


Figure B.3: Examples from our generated dataset used for training MV-ControlNet, utilizing the refined text corpus. Each image was created employing a randomized camera setup.



Figure B.4: Additional examples of our generated dataset. A randomized camera is used for the generation of each image.