A Appendix: SweetDreamer: Aligning Geometric Priors in 2D Diffusion for Consistent Text-to-3D

A.1 Ablation Study

We also validate the effectiveness of key algorithmic design choices using the NeRF-based pipeline. Regarding the geometric representation, we have also derived variants of our method using two types of geometric representations, namely depth and normal maps, and show that, while these two geometric representations can also be incorporated into our pipeline for aligning the geometric priors in 2D diffusion, they indeed fail in producing 3D results as high-quality as CCM.

On one hand, a depth map is a single-channel map storing the per-pixel distance between the surface to the camera. Compared to CCM, the depth map alone does not contain essential information about the relative orientation of the 3D object and the camera, so the network can solely rely on the camera information for learning 3D-aware priors, consequently leading to degraded robustness of the learned 3D priors. In stark contrast, the CCM alone has actually encoded the information of the orientation to some extent. This is manifested by the distinct color coding results of CCMs from different viewpoints. As a result, the variant using depth maps is more prone to multi-view inconsistent issues, compared to our method using CCM. On the other hand, a normal map stores the local directions of fragments on the 3D surface, we use normal maps obtained from the canonical coordinate space for fair comparisons. Our experimental results show that the text-to-3D optimization process has difficulty in generating a 3D world of which the derivative follows faithfully generated normal maps, and may fail to generate meaningful 3D results. Those successfully generated 3D results tend to be slightly smoother regarding the geometric and appearance details of the surface.

Last, we have also conducted a study where the camera information injection is removed from our method. The results demonstrate that removing camera parameters leads to a decrease of the consistency in the generated results, due to the lack of critical camera information. Please see Figure 5 for the visual results and refer to Section A.6 for more.

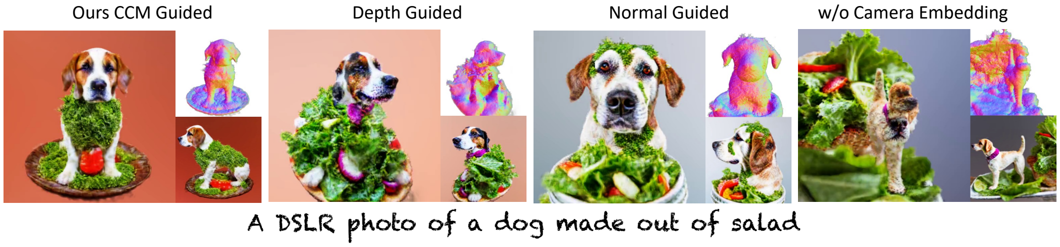


Figure 5: Ablation study. Replacing CCM with depth maps results in inconsistent geometry and the use of normal maps as guidance produces smooth geometry. Additionally, removing the camera embeddings led to noisy and inconsistent results.

A.2 More Implementation Details

Dataset Notably, while there is no explicit specification of the coordinate system in Objaverse, most 3D objects in Objaverse are uploaded by artists who usually adhere to a convention regarding the orientation when creating 3D assets. Furthermore, for some special categories, such as characters, we filter out misoriented data simply by the ratio of its axis-aligned bounding box. After these filterings, we found that approximately 80% (based on statistical random samplings) of the remaining data are orientated canonically, which is sufficient for our purpose as evidenced by extensive results in our paper (please also refer to the appendix for more discussions on the noise in the dataset).

Fine-tuning We use Diffusers (von Platen et al., 2022) to finetune the Stable Diffusion v2.1 using our rendered CCMs at a resolution of 64 x 64. During the fine-tuning process, we remove the encoder of the VAE and directly concatenate the CCM with corresponding alpha channels as the latent to the UNet, with the background color of the CCM set to random. We use the default parameters as in Diffusers, including setting the learning rate to 1e-5 with the constant scheduler, and a batch size of 96 per GPU with 4 gradient accumulation steps. The entire fine-tuning process takes approximately 2 days using 8 V100 GPUs for 100k steps.

Ours (DMTet-based) We integrate our Aligned Geometric Priors in the official repository of Fantasica3D (Chen et al., 2023) as described in Section 3.2. We follow the same parameters as in the original paper. We also disentangle the learning of geometry and appearance. It takes about 12 and 8 minutes to generate a fine geometry and its corresponding Physically-Based Rendering (PBR) materials, respectively, for each object. For the time step range of SDS loss, We adopt a uniform sampling strategy of annealing from [0.5, 0.98] to [0.05, 0.5]. The whole process takes about 0.5 hours to generate each object using 4 V100 GPUs.

Ours (NeRF-based full) We implement it in the threestudio (Guo et al., 2023), which implemented a diverse set of state-of-the-art text-to-3D generation pipelines. Specifically, we use Instant-NGP (Müller et al., 2022) as the 3D representation to optimize, which uses a multi-resolution hash-grid to predict the RGB and the density of the sampled ray points. The sampled camera views follow the same protocol as the render dataset to fine-tune the UNet. We use DeepFolyd at the coarse stage with 64 x 64 resolution and then switch to Stable Diffusion with 512 x 512 for detailed optimization. In addition, we also use time annealing, negative prompts, and CFG rescaling tricks from open source implementation for improved performance. For SDS, the maximum time step is decreased from 0.98 to 0.5 linearly and the maximum time step is kept to 0.02. We use a rescale factor of 0.7 for the CFG rescale. The whole process takes about 1 hour to generate each object with 10, 000 steps using 2 V100 GPUs.

A.3 More comparison results using prompts from MVDream

Note that, since MVDream's official implementation is unavailable by the time of our submission, we use the same prompts as listed on their website for side-by-side comparisons. We present the visual comparisons in Figure 6. Although the concurrent work, MVDream, can also resolve the multi-view inconsistency problem, we observe that it is prone to overfit the limited 3D data, consequently resulting in a compromise of the generalizability in the original powerful 2D diffusion model. Specifically, as shown in the results, MVDream misses the "backpack" in its generated result presented with the prompt "an image of a pig carrying a backpack". Additionally, since they use synthetic multi-view renderings for fine-tuning their multi-view diffusion model, the appearance of the generated results lacks the desired level of photorealism.

A.4 Generalizability

Our method can effectively address the notorious multi-view inconsistency problem, and equally importantly, retains to the maximum extent the generalizability of the foundation text-to-image model in terms of the highly varied appearance and geometric details. We would like to highlight that the pre-trained text-to-image diffusion model is powerful, and this preservation of its generalizability is particularly attractive, as it can lead to more diverse and highly realistic 3D generation results. We have achieved this by only aligning the geometric priors in 2D diffusion using the coarse geometric information (CCM) of well-defined geometries, without compromising the original diffusion priors in the text-to-3D pipeline regarding detailed appearances and geometries. This is in contrast to other models that hinge on all appearance and geometric details in the 3D dataset for fine-tuning the diffusion priors, which is at the risk of compromising the integrity of the original geometric priors learned in the pre-trained text-to-image foundation model, leading to the degradation of the generalizability in terms of highly diverse and photo-realistic 3D objects. While we have validated this through the comparison results against MVDream (Figure 6), we have also prepared more extensive qualitative results to further showcase such ability of our method to generalize smoothly to 3D worlds with highly realistic and varied appearance and geometric details, that are unseen in the 3D dataset. See Figure 7 for more visual results.

More specifically, the generalization encompasses both geometric and appearance detils. In our work, we only perform fine-tuning using the coarse geometric structures, which enables us to preserve the highly realistic appearance details of the original image priors learned from a large real-

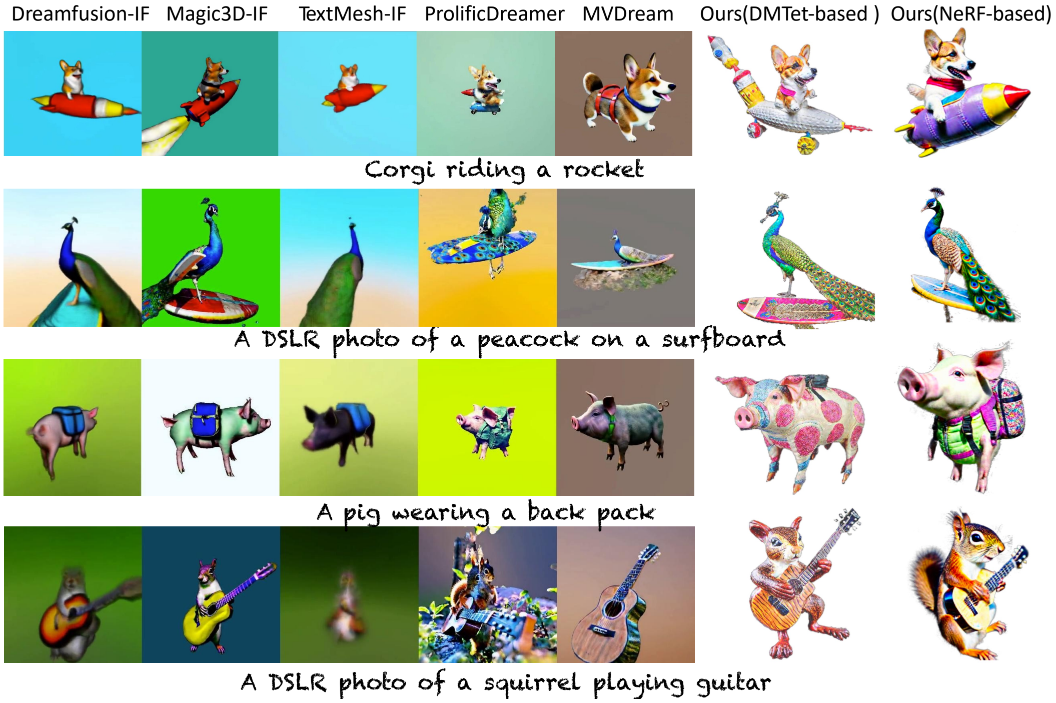


Figure 6: Side-by-side visual comparisons using prompts from MVDream. Note that some key concepts in the prompts are missing in MVDream results, such as the rocket, backpack, and squirrel missing in their results.

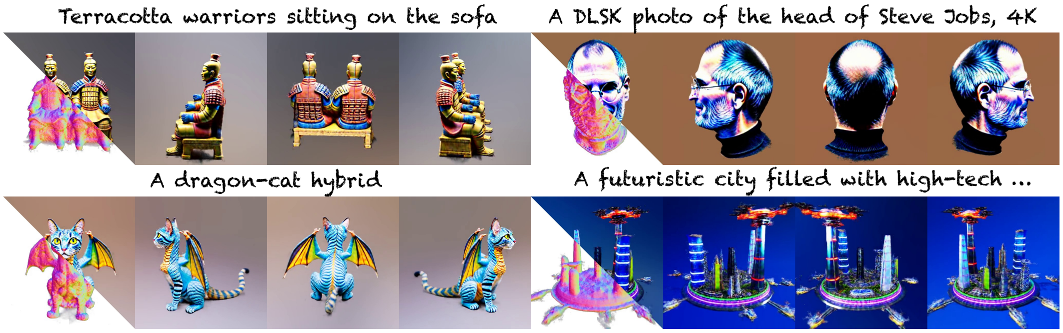

Terracotta warriors sitting on the sofa
A DLSK photo of the head of Steve Jobs, 4K
A dragon-cat hybrid
A futuristic city filled with high-tech ...

Figure 7: Our method can produce a wide range of 3D results spanning from weird characters to fancy architectural models. It is worth noting that none of these 3D models exists in the training data. Our approach can generalize very well to these highly diverse and high-quality results.

world dataset. Interestingly, we found that in our result of using the prompt ”corgi riding on a rocket”, the rough geometry obtained with only AGP priors is not yet completely accurate. However, after optimizing with SDS of image priors, we are able to fix the slightly incorrect geometry and obtain more detailed geometric and appearance details, as shown in Figure 8.

A.5 The Effect of Different Settings of Dataset

We used a set of rules to filter the original dataset and obtained approximately 270k objects, of which approximately 80% have a consistent orientation (estimated based on statistical random samplings).

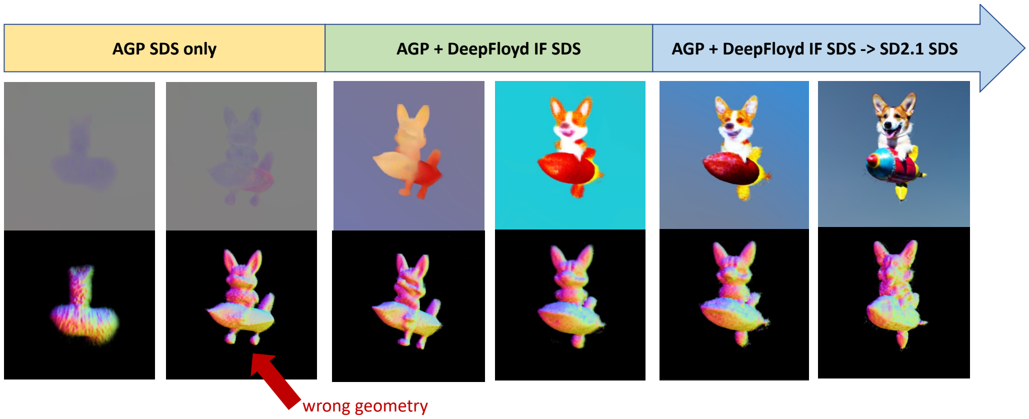


Figure 8: Visualization of each stage of our NeRF-based method of prompt ”Corgi riding a rocket”. Using only AGP SDS as supervision may result in incorrect geometry. However, the image priors can help eliminate these errors in the later stages, thereby improving the generalization performance.

In this section, we aim to investigate the impact of a different-sized dataset on the performance of our proposed method.

We calculated the distance between the front-view images and their corresponding captions' blip feature and sorted them accordingly. We then manually screened the top 20k data, resulting in a dataset of 18,488 instances with a uniform orientation and high correspondence with their descriptions.

From the experimental results obtained by training on such a smaller dataset, we are surprised to find that even with a smaller dataset of only 20k instances for fine-tuning, our text-to-3D pipeline still exhibits strong generalization and is capable of generating diverse 3D objects. As shown in Figure 9, the majority of the generated results are satisfactory, with some lacking intricate details. We hypothesize that this may be attributed to the inadequate acquisition of 3D prior knowledge from a rather limited dataset, which consequently leads to the occurrence of coarse geometries. This observation necessitates the need for further exploration and a more thorough study of the impact of the 3D dataset.

A.6 More Details of Ablation Study

We conducted ablation experiments using a manually curated small dataset (20k data) for each object. We rendered CCM, normal, and depth maps for each object and used the same parameters to fine-tune the Stable Diffusion model. We then tested using the same prompt and seed. Extensive visual results are presented in Figure 10.

A.7 Quantitative Evaluation of the Appearance of Generated Objects

We also present quantitative results using appearance-related metrics. However, it is worth noting that the image diffusion models used in Dreamfusion and Magic3D are not open-sourced, and the performance of the versions used here may differ from the original papers. For fairness, we used the same image generation model, DeepFloyd IF (IF, 2023) for all methods, and NeRF as the 3D representation.

To assess the appearance quality, we rely on DreamFusion's R-Precision metric. When provided with rendered images, R-Precision gauges the top-N accuracy in retrieving the correct caption from a pool of distractions, utilizing CLIP scores derived from averaging the similarity between each of four distinct rendered images and a caption. We employ the CLIP ViT-B/32 and ViT-B/16 models to respectively compute the top-1 and top-5 R-Precision, using the identical set of 80 prompts for evaluation as presented in the paper. Please note that since other methods may not have access to

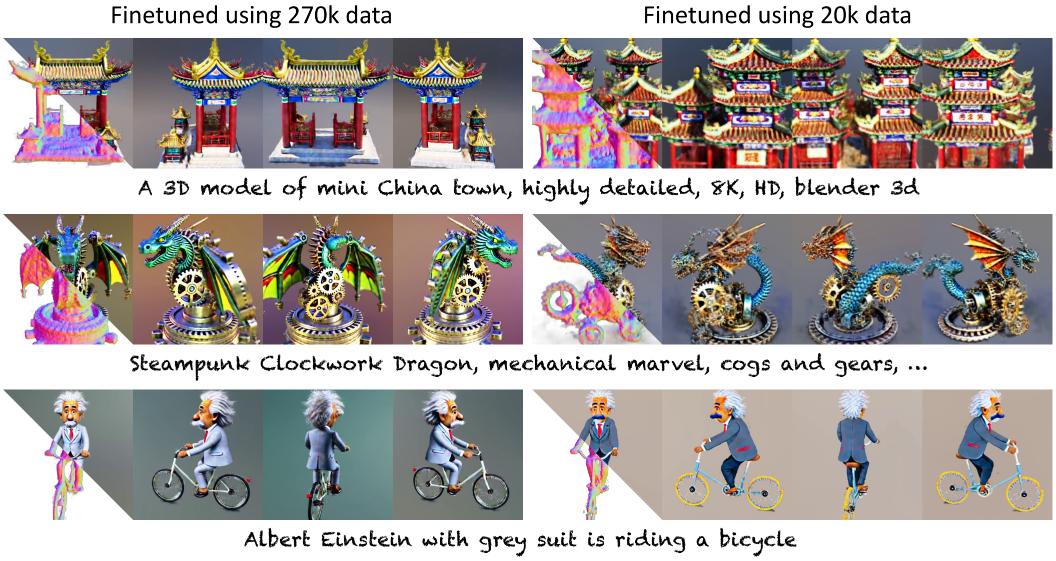

Finetuned using 270k data
Finetuned using 20k data
A 3D model of mini China town, highly detailed, 8K, HD, blender 3d
Steampunk Clockwork Dragon, mechanical marvel, cogs and gears, ...
Albert Einstein with grey suit is riding a bicycle

Figure 9: Comparison of finetuning results with different amounts of data. Despite reducing the training data to 1/10 of the original amount, the finetuned models still exhibit strong generalization capabilities (right). Note that, the model fine-tuned on fewer data tends to produce results of slightly fewer details and probably a bit more bulky geometries.

Method	CLIP R-Precision(%)			
	CLIP B/32		CLIP B/16	
	R@1	R@5	R@1	R@5
Magic3D	60.1	71.5	62.7	77.7
TextMesh	51.7	65.1	55.1	78.4
SJC	40.2	51.2	52.5	62.5
DreamFusion	59.7	70.2	61.6	74.3
Ours (NeRF-based)	**77.5**	**84.9**	**88.7**	**92.3**

Table 2: Quantitative results demonstrating the coherence of visual appearance with their corresponding prompts, as assessed by CLIP retrieval models.

the original image diffusion model, the performance of reproduced code may vary. Therefore, the specific numerical results can only serve as a reference.

The results highlighted in Table 2 underscore the significant advantages of our proposed approach in terms of appearance. This superiority primarily stems from the fine-tuning process which only learns from coarse geometries and leaves the appearance model of the text-to-3D pipeline untouched.

A.8 More details about the user study

In contrast to the quantitative evaluation above, the human users involved in this study are from various loosely related backgrounds in CS and EE. The interactive interface used in the user study can be found in Figure 11.

A.9 More Text-to-3D results

We present more text-to-3D synthesis results obtained with our methods (Figure 12, Figure 13, Figure 14, and Figure 15).

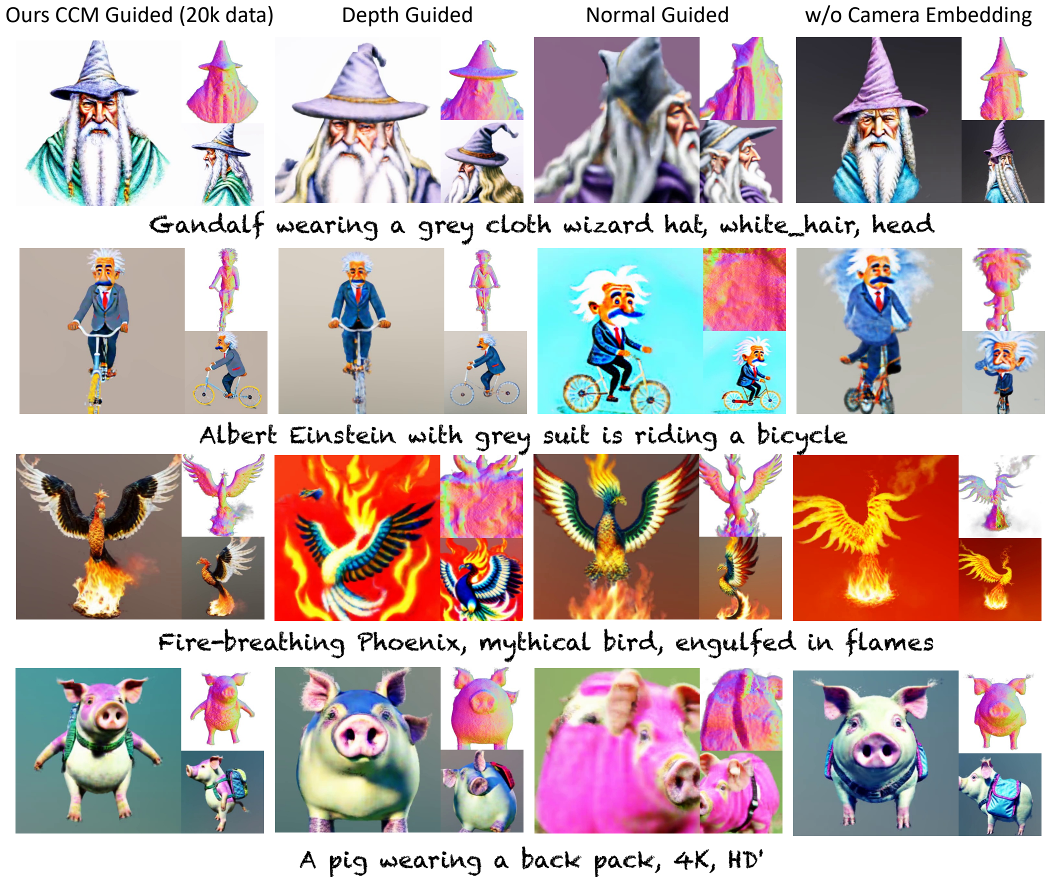
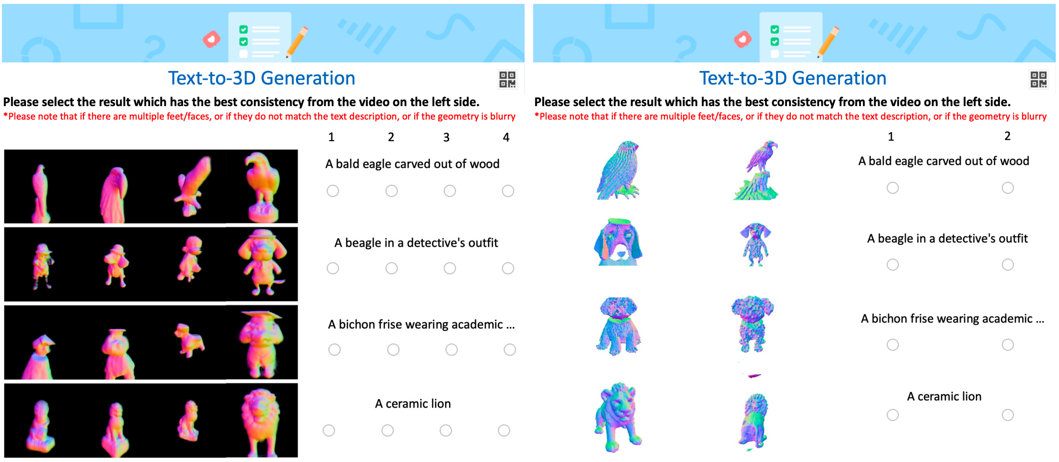

Ours CCM Guided (20k data)
Depth Guided
Normal Guided
w/o Camera Embedding
Gandalf wearing a grey cloth wizard hat, white_hair, head
Albert Einstein with grey suit is riding a bicycle
Fire-breathing Phoenix, mythical bird, engulfed in flames
A pig wearing a back pack, 4K, HD'

Figure 10: Comparing ablate settings with different guidance methods. Our proposed CCM proves superior to depth and normal guidance, delivering consistent and convincing geometry. Substituting CCM with depth maps leads to round (e.g. the pig) and inconsistent objects (e.g. the Gandalf and phoenix), while normal maps as guidance make the optimization unstable (e.g. the Gandalf, Einstein, and pig), leads to smoother (e.g. the phoenix) or noisy (e.g. the chihuahua) geometry. It is essential to note that removing the camera embedding may result in inconsistent results, particularly on complex structures (e.g. the Einstein).



Text-to-3D Generation

Please select the result which has the best consistency from the video on the left side.

*Please note that if there are multiple feet/faces, or if they do not match the text description, or if the geometry is blurry

	1	2	3	4
A bald eagle carved out of wood	○	○	○	○
A beagle in a detective's outfit	○	○	○	○
A bichon frise wearing academic ...	○	○	○	○
A ceramic lion	○	○	○	○

Text-to-3D Generation

Please select the result which has the best consistency from the video on the left side.

*Please note that if there are multiple feet/faces, or if they do not match the text description, or if the geometry is blurry

	1	2
A bald eagle carved out of wood	○	○
A beagle in a detective's outfit	○	○
A bichon frise wearing academic ...	○	○
A ceramic lion	○	○

Figure 11: User study interface. We have designed two survey forms separately for the NeRF-based method and the DMTet-based method and provide users with multiple rendered videos of generated results. The user needs to select the one they consider to be of the best quality from all results.

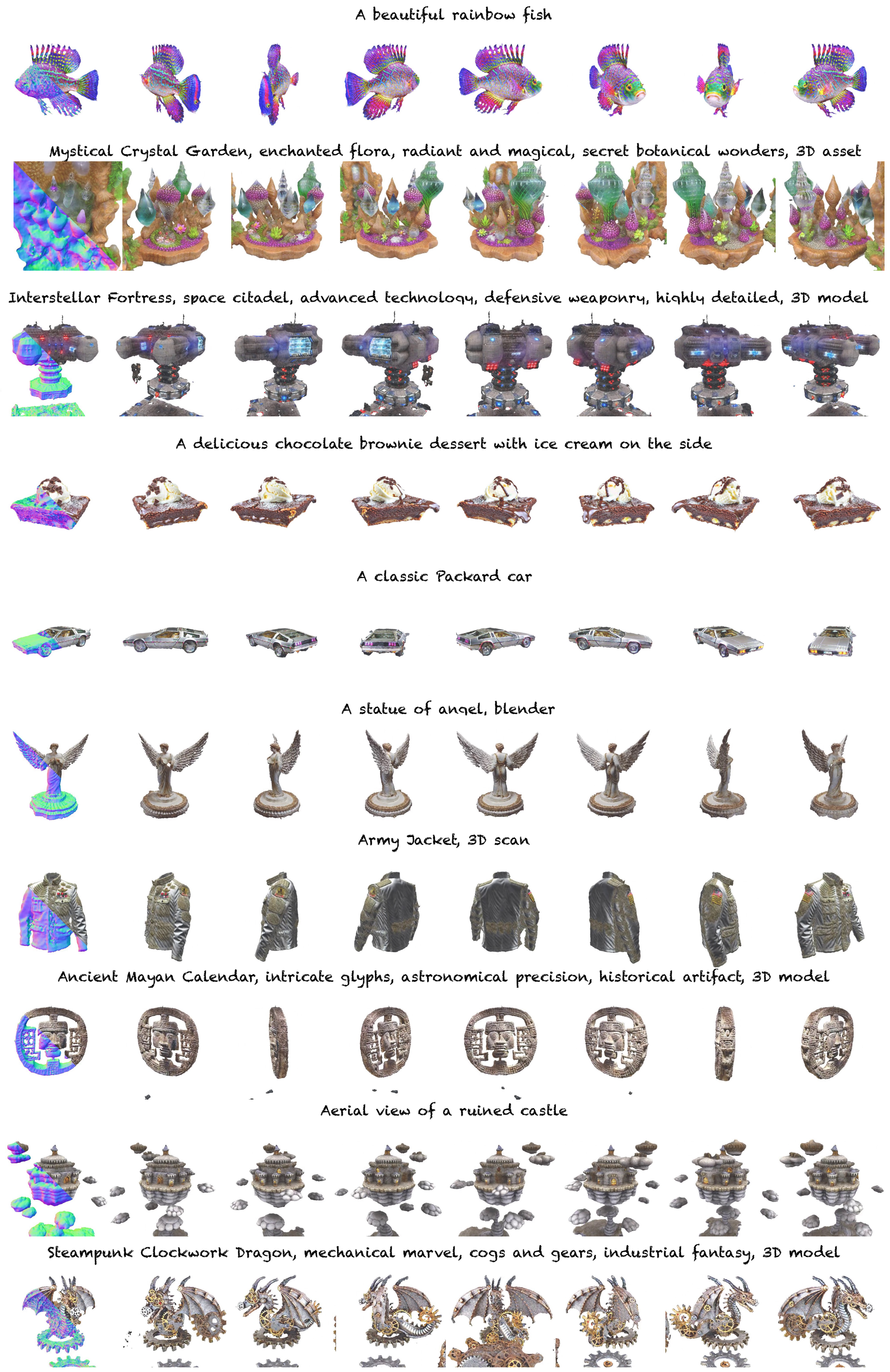

A beautiful rainbow fish
Mystical Crystal Garden, enchanted flora, radiant and magical, secret botanical wonders, 3D asset
Interstellar Fortress, space citadel, advanced technology, defensive weaponry, highly detailed, 3D model
A delicious chocolate brownie dessert with ice cream on the side
A classic Packard car
A statue of angel, blender
Army Jacket, 3D scan
Ancient Mayan Calendar, intricate glyphs, astronomical precision, historical artifact, 3D model
Aerial view of a ruined castle
Steampunk Clockwork Dragon, mechanical marvel, cogs and gears, industrial fantasy, 3D model

Figure 12: More generated results using our proposed DMTet-based model.

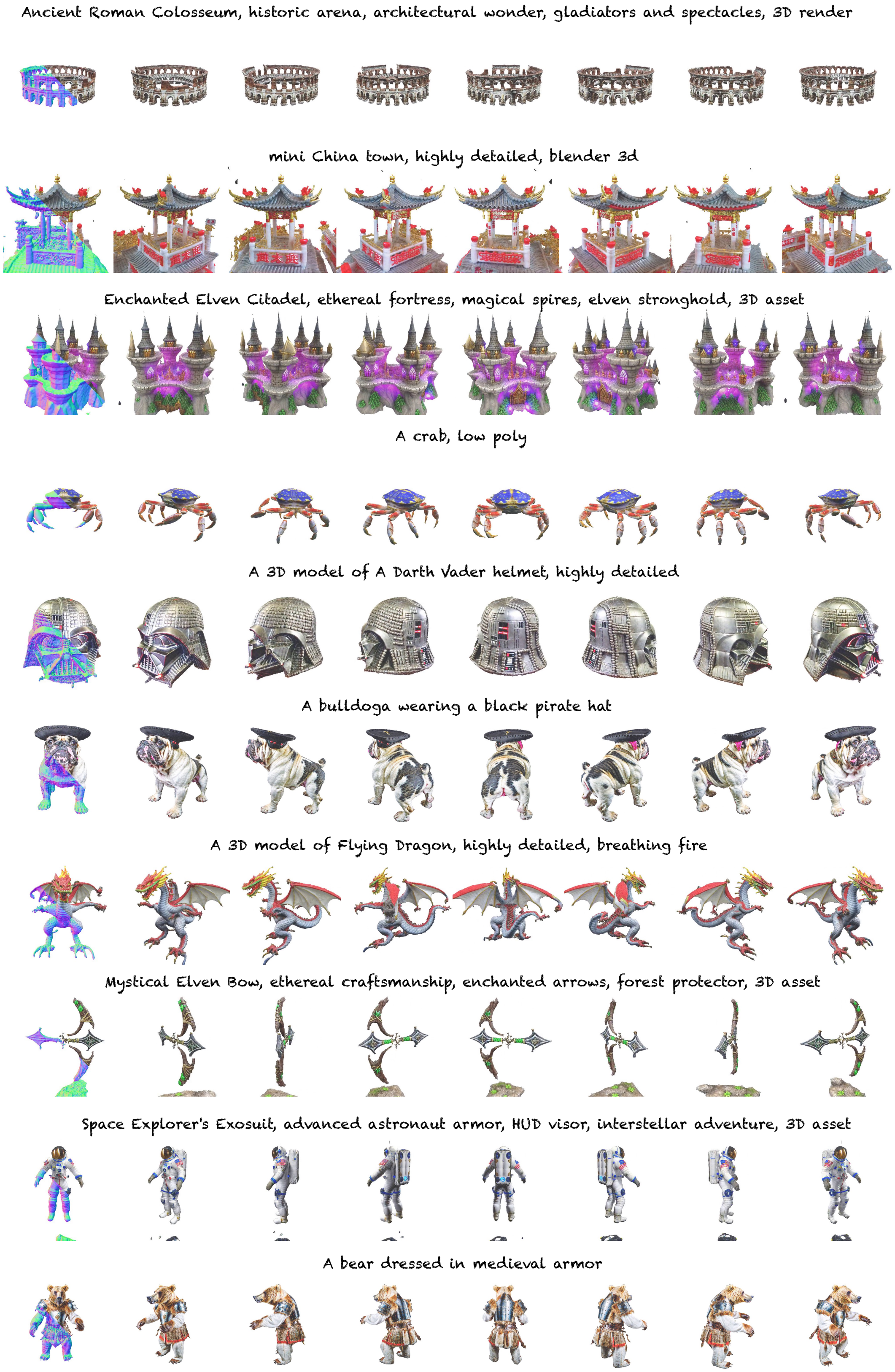

Ancient Roman Colosseum, historic arena, architectural wonder, gladiators and spectacles, 3D render
mini China town, highly detailed, blender 3d
Enchanted Elven Citadel, ethereal fortress, magical spires, elven stronghold, 3D asset
A crab, low poly
A 3D model of A Darth Vader helmet, highly detailed
A bulldoga wearing a black pirate hat
A 3D model of Flying Dragon, highly detailed, breathing fire
Mystical Elven Bow, ethereal craftsmanship, enchanted arrows, forest protector, 3D asset
Space Explorer's Exosuit, advanced astronaut armor, HUD visor, interstellar adventure, 3D asset
A bear dressed in medieval armor

Figure 13: More generated results using our proposed DMTet-based model.

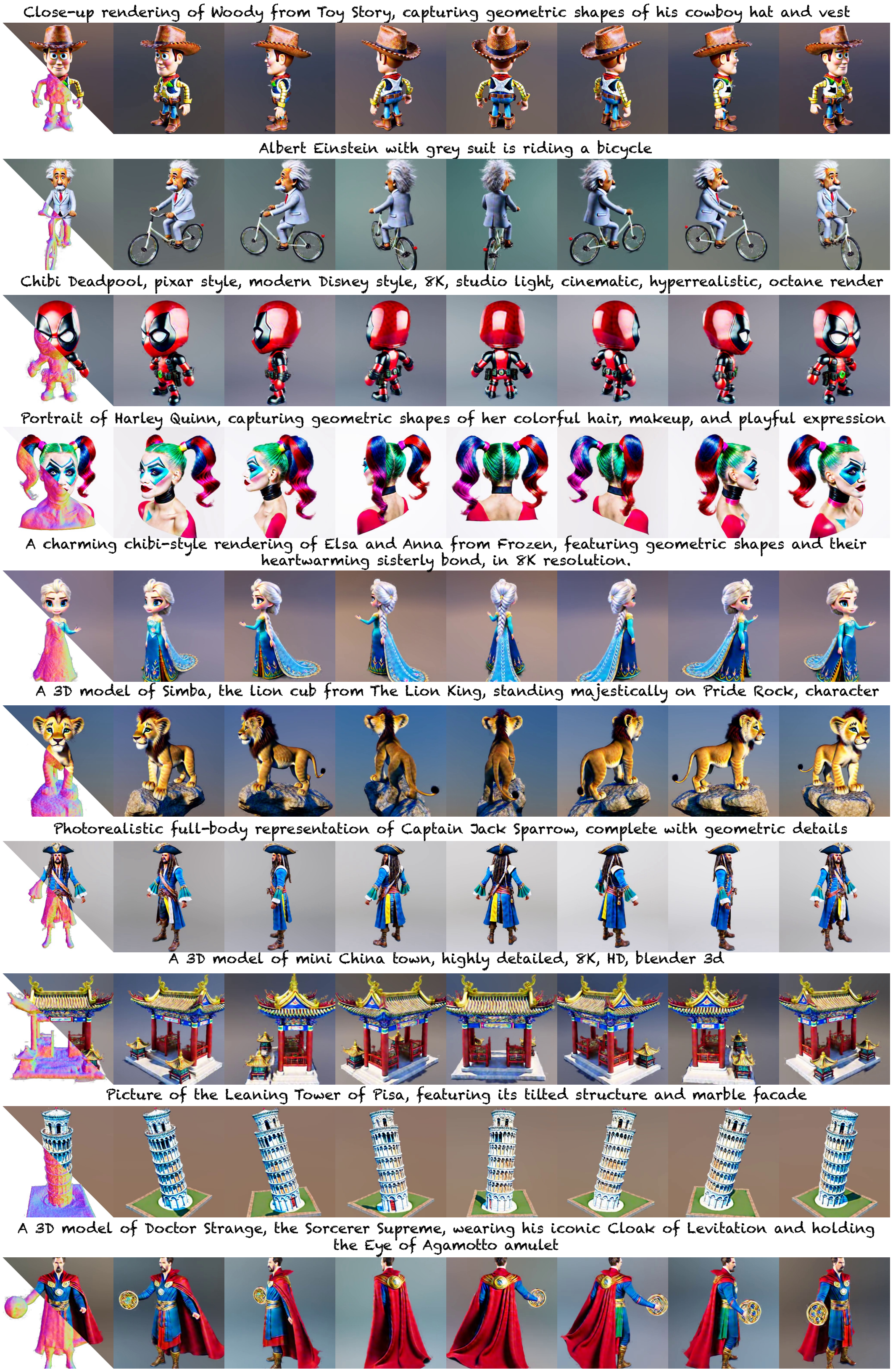

Close-up rendering of Woody from Toy Story, capturing geometric shapes of his cowboy hat and vest
Albert Einstein with grey suit is riding a bicycle
Chibi Deadpool, pixar style, modern Disney style, 8K, studio light, cinematic, hyperrealistic, octane render
Portrait of Harley Quinn, capturing geometric shapes of her colorful hair, makeup, and playful expression
A charming chibi-style rendering of Elsa and Anna from Frozen, featuring geometric shapes and their heartwarming sisterly bond, in 8K resolution.
A 3D model of Simba, the lion cub from The Lion King, standing majestically on Pride Rock, character
Photorealistic full-body representation of Captain Jack Sparrow, complete with geometric details
A 3D model of mini China town, highly detailed, 8K, HD, blender 3d
Picture of the Leaning Tower of Pisa, featuring its tilted structure and marble facade
A 3D model of Doctor Strange, the Sorcerer Supreme, wearing his iconic Cloak of Levitation and holding the Eye of Agamotto amulet

Figure 14: More generated results using our proposed NeRF-based model.

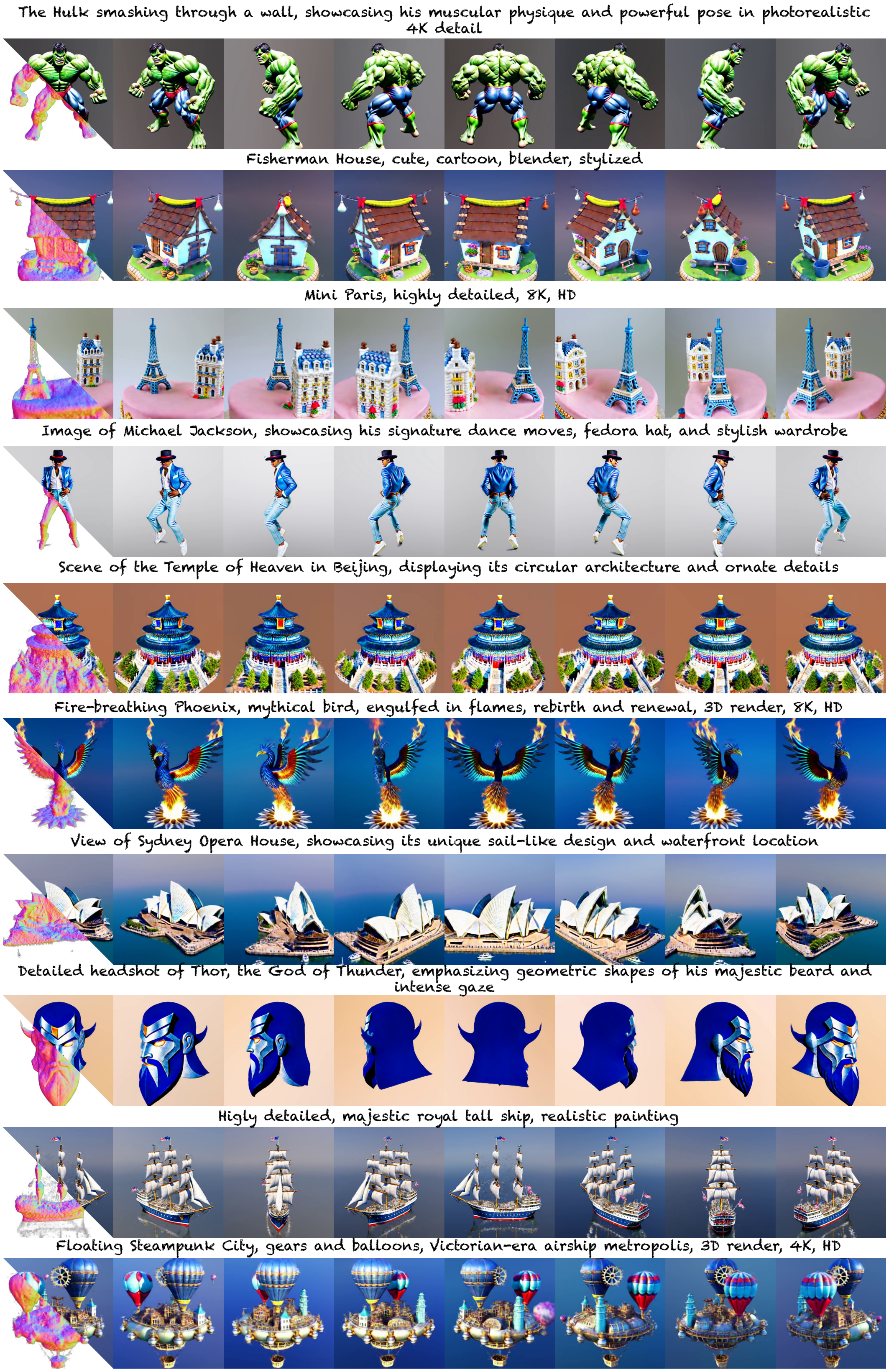

The Hulk smashing through a wall, showcasing his muscular physique and powerful pose in photorealistic 4K detail
Fisherman House, cute, cartoon, blender, stylized
Mini Paris, highly detailed, 8K, HD
Image of Michael Jackson, showcasing his signature dance moves, fedora hat, and stylish wardrobe
Scene of the Temple of Heaven in Beijing, displaying its circular architecture and ornate details
Fire-breathing Phoenix, mythical bird, engulfed in flames, rebirth and renewal, 3D render, 8K, HD
View of Sydney Opera House, showcasing its unique sail-like design and waterfront location
Detailed headshot of Thor, the God of Thunder, emphasizing geometric shapes of his majestic beard and intense gaze
Higly detailed, majestic royal tall ship, realistic painting
Floating Steampunk City, gears and balloons, Victorian-era airship metropolis, 3D render, 4K, HD

Figure 15: More generated results using our proposed NeRF-based model.