Supplementary Material for DreamComposer: Controllable 3D Object Generation via Multi-View Conditions

Yunhan Yang[1*] Yukun Huang[1*] Xiaoyang Wu[1] Yuan-Chen Guo[3,4]
Song-Hai Zhang[4] Hengshuang Zhao[1] Tong He[2] Xihui Liu[1†]

[1] The University of Hong Kong [2] Shanghai Artificial Intelligence Lab [3] VAST [4] Tsinghua University
[*] Equal Contribution Project Page: https://yhyang-myron.github.io/DreamComposer/

A. Implementation Details

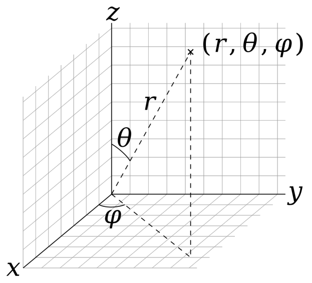

Figure 1. Spherical Coordinate System [8].

Camera Embedding. Following Zero-1-to-3 [4], we utilize a spherical coordinate system to represent camera locations and their relative transformations. As shown in Figure 1, during the training stage, camera locations of two images from disparate viewpoints are designated as (θ_1, ϕ_1, r_1) and (θ_2, ϕ_2, r_2), respectively. The *relative* transformation between these camera positions is expressed as $(\theta_2 - \theta_1, \phi_2 - \phi_1, r_2 - r_1)$. In both the training and inference stages, four parameters delineating the relative camera viewpoint $[\Delta\theta, \sin(\Delta\phi), \cos(\Delta\phi), \Delta r]$ are inputted into the cross-attention layers of DreamComposer's Target-Aware 3D Lifting Module and Target-View Feature Injection Module to provide camera view information.

Architecture and Hyperparameters. We design Target-Aware 3D Lifting Module based on the U-Net architecture from Stable Diffusion [7]. This model's architecture is specifically configured with a model dimension of 192 and includes two residual blocks at each resolution level. A distinctive feature of our approach is the integration of a cross-attention module, which facilitates the processing of relative camera embeddings.

For our experiments, we standardize the image dimensions at 256×256 pixels. Correspondingly, this establishes the latent space dimensionality at 32×32. Additionally, we configure the triplane dimensions at $32 \times 32 \times 3$, with the feature dimension of each triplane element being set to 32.

Training Details. We adopt a two-stage training strategy for DreamComposer. In the first stage, we focus on the 3D feature lifting module and pre-train it for 80k steps ($\sim$ 3 days) with 8 80G A800 GPUs using a total batch size of 576. The pre-trained 3D lifting module can be applied in conjunction with different pre-trained diffusion models for subsequent training. In the second stage, we jointly optimize the 3D lifting and feature injection module. This stage takes 30k steps ($\sim$ 2 days) with 8 80G A800 GPUs using a total batch size of 384.

B. Additional Ablation Analysis

B.1. Comparison with NVS from sparse views

To compare with novel view synthesis methods from sparse-view inputs, we choose ViewFormer [3] as competitor, which achieves significant results on the CO3D dataset [6]. ViewFormer is designed for novel view synthesis using sparse-view inputs, employing transformers to process multiple context views and a query pose. This approach allows for the synthesis of novel images within an advanced neural network architecture. For our evaluation, we utilize the ViewFormer model that has been comprehensively trained on the CO3D dataset [6], ensuring a fair comparison with its contemporary counterparts. The evaluation dataset setting is same as the one in Section 4.3. The quantitative results are shown in Table 1, and the qualitative results are shown in Figure 4. While ViewFormer shows proficiency in handling the CO3D dataset, it exhibits limitations in processing out-of-distribution data. DC-Zero-1-to-3 far surpasses novel view synthesis methods with sparse-view inputs in both qualitative and quantitative analysis.

Our method is capable of zero-shot learning and also demonstrates superior performance compared to other few-shot reconstruction methods when testing on their datasets. A qualitative comparison with PixelNeRF (PN), NerFormer (NF), SF (SparseFusion) is presented in Figure 2. The con-

(a) Elevation Degree - 0

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
ViewFormer	13.45	0.630	0.359
Zero-1-to-3	20.82	0.840	0.139
DC-Zero-1-to-3 (Ours)	**25.25**	**0.888**	**0.088**

(b) Elevation Degree - 15

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
ViewFormer	13.00	0.618	0.371
Zero-1-to-3	21.38	0.837	0.131
DC-Zero-1-to-3 (Ours)	**25.85**	**0.891**	**0.083**

(c) Elevation Degree - 30

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
ViewFormer	13.02	0.618	0.373
Zero-1-to-3	21.66	0.837	0.128
DC-Zero-1-to-3 (Ours)	**25.63**	**0.885**	**0.086**

Table 1. Quantitative comparisons of novel view synthesis on GSO dataset using four orthogonal angles' images as inputs. DC-Zero-1-to-3 far surpasses other methods on all metrics.

tents in CO3DV2 dataset is not at the center of images, which is not aligned to the setting of Zero-1-to-3 and SyncDreamer, so it is not proper to give quantitative results.

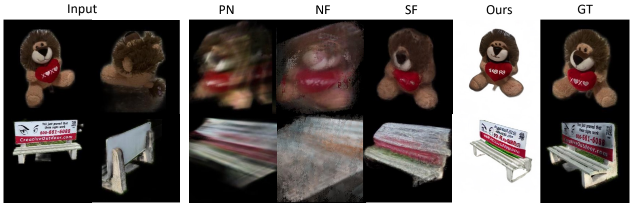

Figure 2. Qualitative comparison with PixelNeRF (PN), NerFormer (NF), SF (SparseFusion) in novel view synthesis.

B.2. Scalability for arbitrary inputs

We further explore our model's flexibility and scalability in managing arbitrary numbers of inputs. We evaluate the model's performance using the same set of 30 objects from Section 4.3, but with differing input counts. To ensure the robustness of the experiment, we strategically select input perspectives to encompass a broad area. Specifically, for 2 inputs, we use angles 0° and 180°; for 3 inputs, we use angles 0°, 90°, 180°; for 4 inputs, we use angles 0°, 90°, 180°, 270°. We show the quantitative results in Table 2. Subsequently, we assess additional datasets and present the qualitative outcomes in Figure 6.

B.3. Necessity of view-conditioning for 3D lifting.

Under different angle difference inputs, we visualize the triplane features from the target view. As shown in Figure 3, the projection from the target view has the highest quality. The experimental setup, as outlined in the first column, involves two inputs with the primary view presented at the top. The difference in views is computed in relation to this primary view. In the first row, the specified view difference is 20 degrees, hence, only the subsequent result at the corresponding 20 degrees is deemed valid. Similarly, in the second row, a view difference of 70 degrees is specified, making only the subsequent result at the corresponding 70 degrees valid.

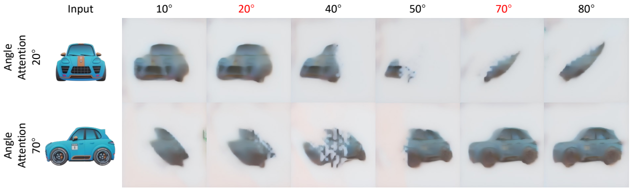

Figure 3. Latent space visualization with different angle attention.

C. Additional Results

C.1. DC-SyncDreamer Results

We present more qualitative comparisons of SyncDreamer [5] and our method on the GSO [2] dataset, as shown in Figure 5. We utilize an image as the input for SyncDreamer, as well as the main input view for our DC-SyncDreamer. We further generate the back-view image of the object with Zero-1-to-3 [4], serving as an additional condition-view for DC-SyncDreamer. We present additional qualitative results on Objaverse [1] dataset in Figure 7. Video demonstrations of these generated objects are included in the supplementary material. Leveraging the multi-view information about the object, DC-SyncDreamer is capable of generating controllable novel views and 3D objects.

D. Limitations

Although DreamComposer can leverage multi-view inputs to enhance zero-shot novel view synthesis, we empirically found that it is still unsatisfactory in preserving fine-grained textures from non-main view input images. It may be caused by the fact that we adopt multi-view conditioning on a low-resolution latent space, which is efficient but suffers from the loss of high-frequency details. In addition, angular deviations between multi-view input images may affect the generation quality.

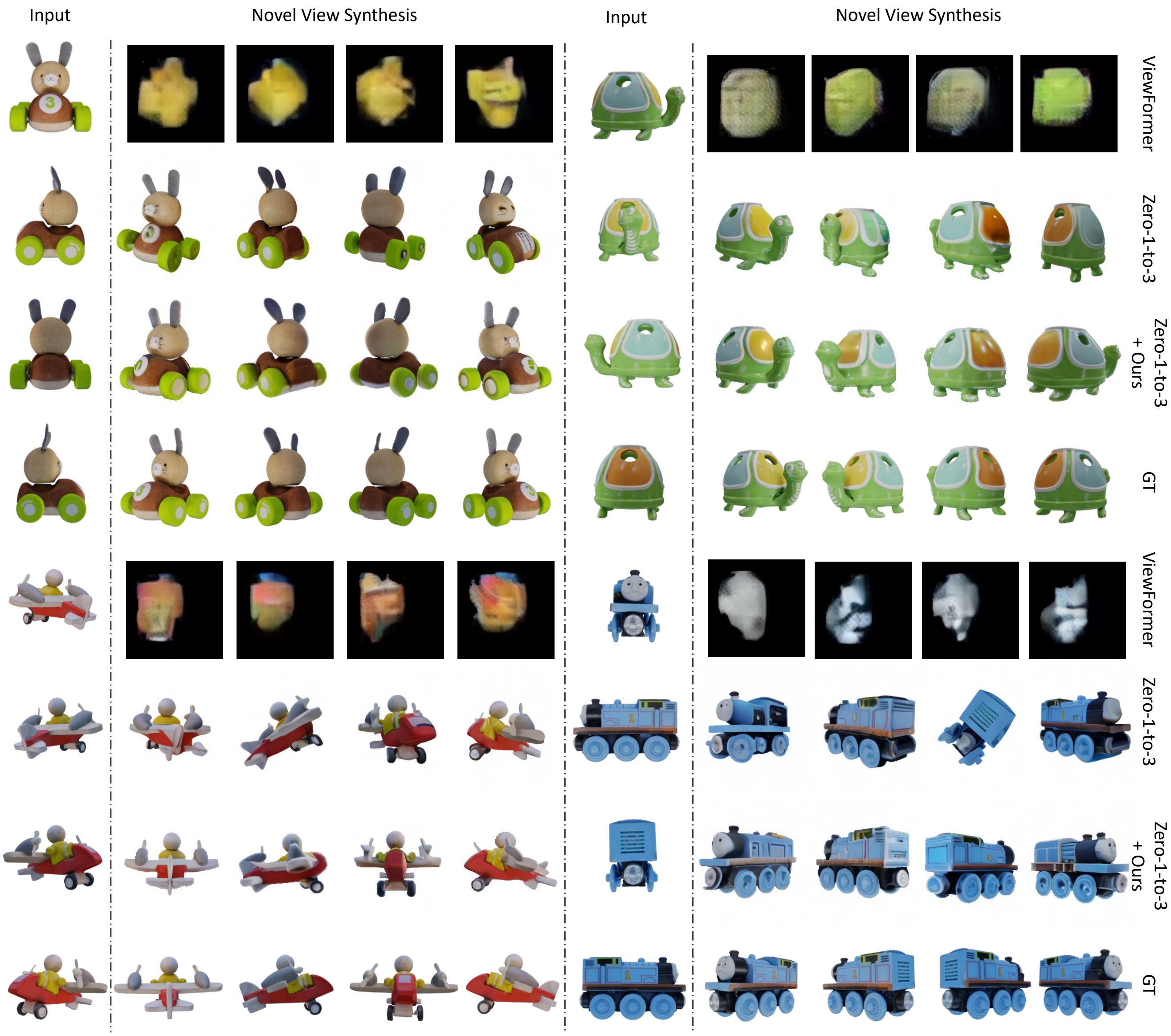

Figure 4. Qualitative comparisons with ViewFormer [3] and Zero-1-to-3 [4] of novel view synthesis on GSO dataset using four orthogonal angles' images as inputs. ViewFormer, despite its training on the CO3D dataset, demonstrates limitations in processing out-of-domain data. In contrast, by integrating multi-view information, our model exhibits the capability to produce controllable and superior-quality images from new perspectives of in-the-wild data.

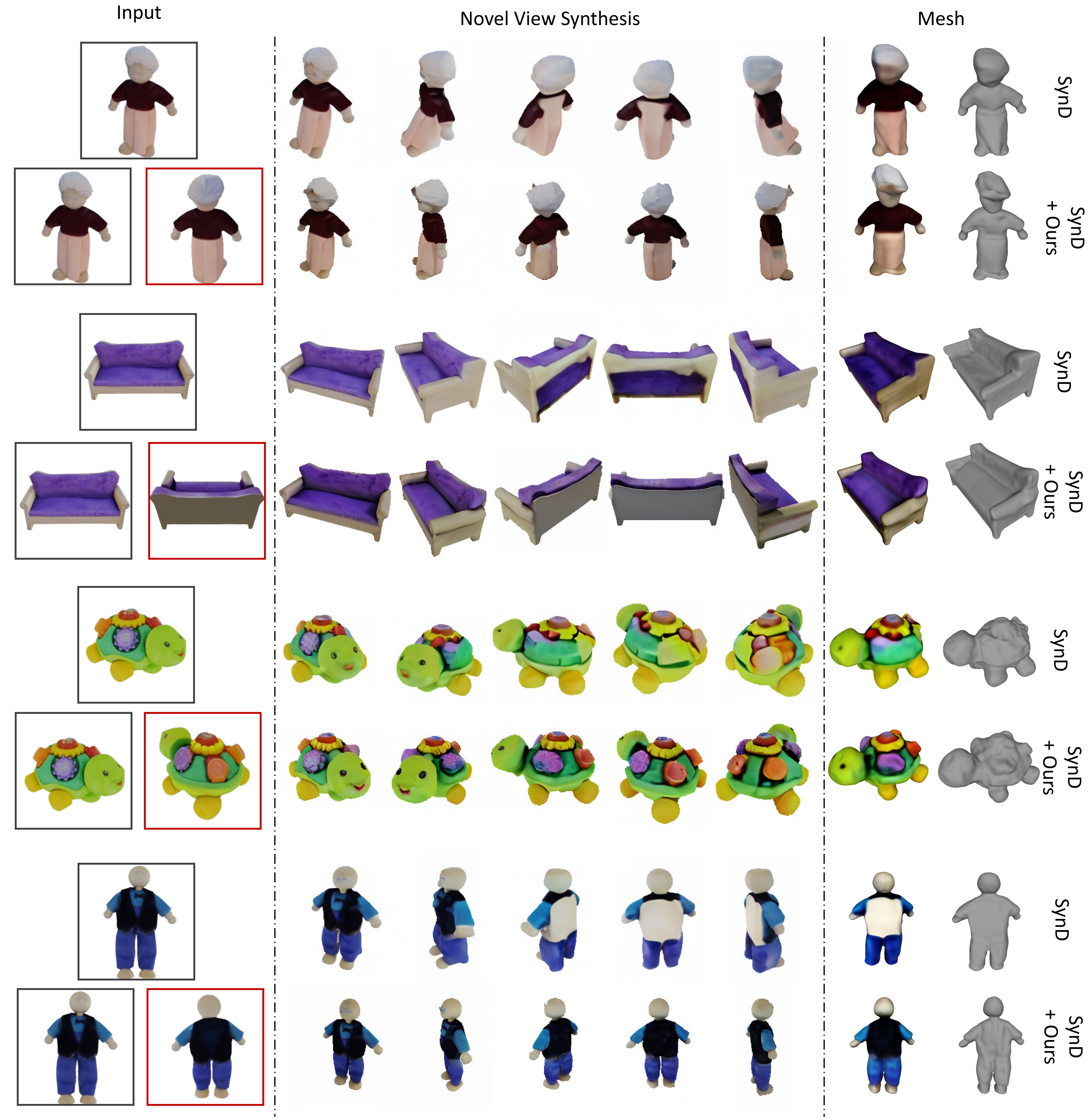

Figure 5. Qualitative comparisons with SyncDreamer (SyncD) [5] in controllable novel view synthesis and 3D reconstruction. The image in □ is the main input, and the other image in □ is the conditional input generated from Zero-1-to-3 [4]. With more information in multi-view images, DC-SyncDreamer is able to generate more accurate back textures and more controllable 3D shapes.

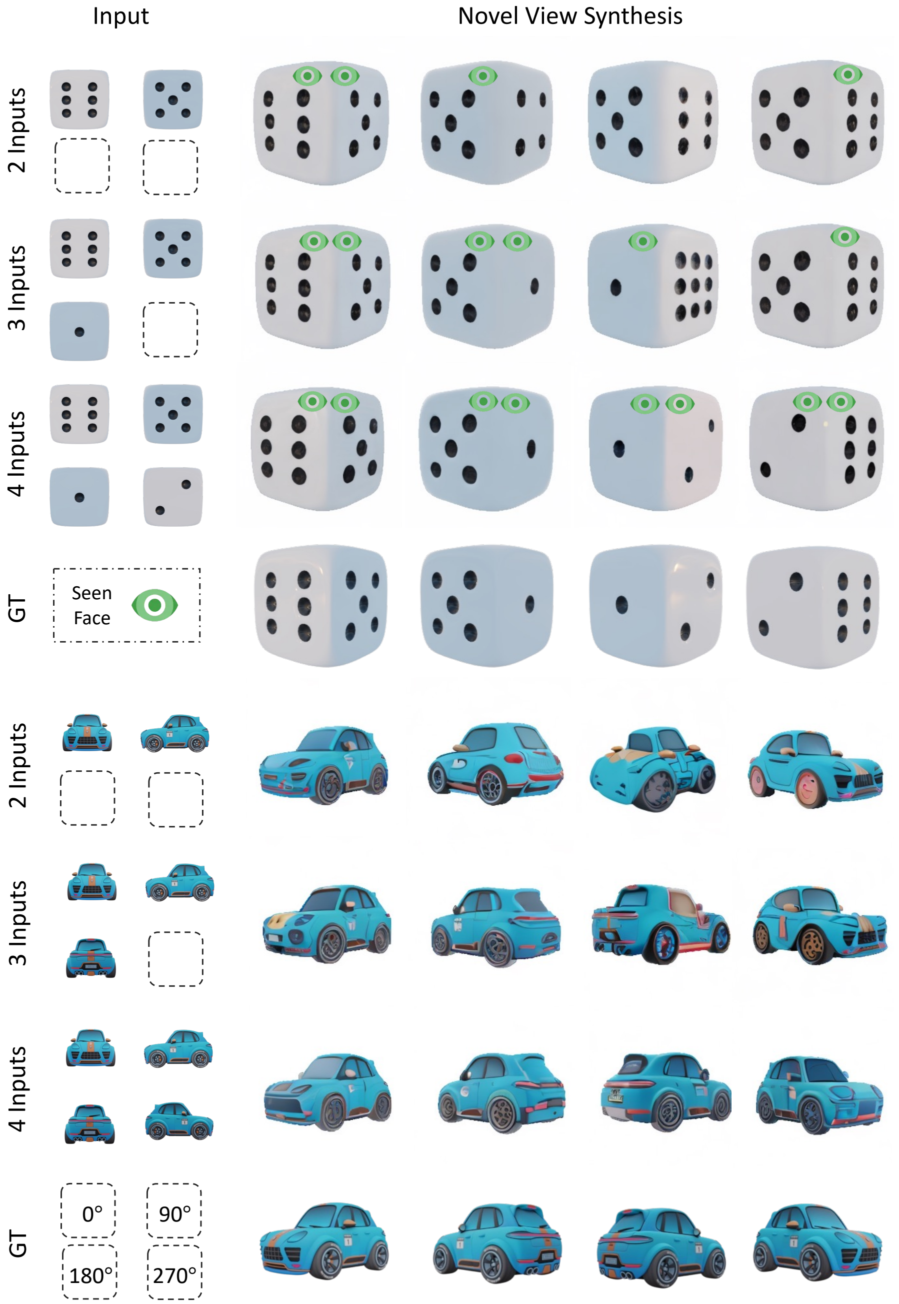

Figure 6. Ablation study to demonstrate the scalability of DreamComposer. Our model can handle various inputs and that its control over the results gets better as the amount of information from the inputs increases.

	Elevation Degree - 0			Elevation Degree - 15			Elevation Degree - 30		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
2 views	20.38	0.826	0.159	22.33	0.847	0.125	22.42	0.845	0.124
3 views	23.68	0.869	0.108	24.56	0.875	0.098	24.27	0.867	0.102
4 views	25.25	0.888	0.088	25.85	0.891	0.083	25.63	0.885	0.086
5 views	26.10	0.897	0.081	26.62	0.899	0.078	26.52	0.895	0.079
6 views	**26.99**	**0.907**	**0.074**	**27.39**	**0.907**	**0.072**	**27.26**	**0.903**	**0.073**

Table 2. Quantitative comparisons on the GSO dataset with different number of inputs. As the number of input images increases, the generation of new perspectives becomes more controllable.

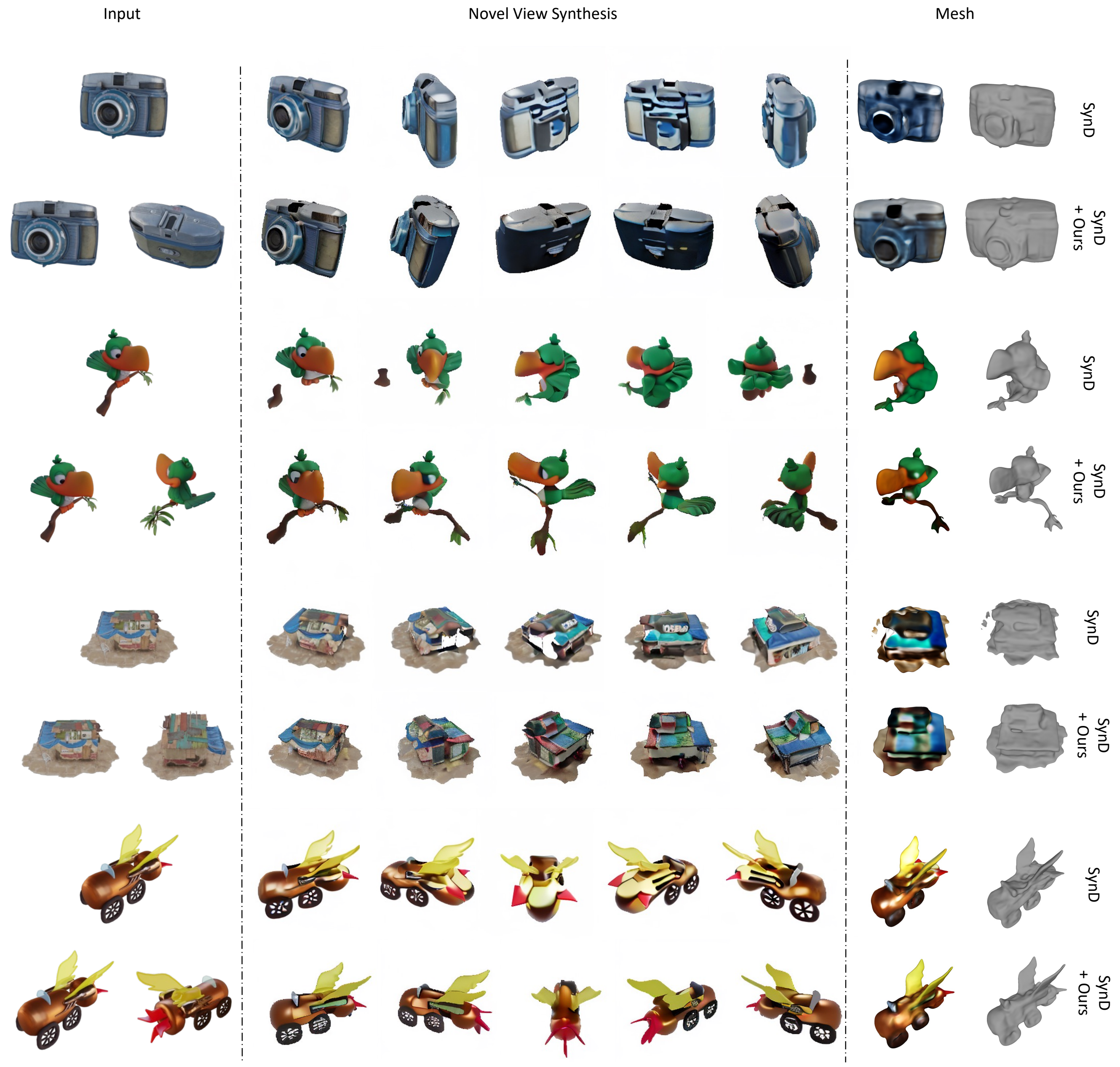

Figure 7. Qualitative comparisons with SyncDreamer (SyncD) on Objaverse dataset.

References

[1] 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, 2022. 2

[2] 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 *ICRA*, 2022. 2

[3] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. Viewformer: Nerf-free neural rendering from few images using transformers. In *European Conference on Computer Vision*, pages 198–216. Springer, 2022. 1, 3

[4] 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, 2023. 1, 2, 3, 4

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

[6] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[8] Wikipedia. Spherical coordinate system — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Spherical%20coordinate%20system&oldid=1142703172, 2023. [Online; accessed 14-March-2023]. 1