Synthesizing Consistent Novel Views via 3D Epipolar Attention without Re-Training
Supplementary Material

Botao Ye[1,2] Sifei Liu[3] Xueting Li[3] Marc Pollefeys[1,4] Ming-Hsuan Yang[5]
[1]ETH Zurich [2]ETH AI Center [3]NVIDIA [4]Microsoft [5]UC Merced

A. Epipolar Line Calculation

Here we provide detailed proof that the final epipolar line $\boldsymbol{l}_i$ is independent of the unknown focal length f.

Given the rotation matrix $\mathbf{R}$ and translation vector $\boldsymbol{t}$ between the two cameras, and the camera intrinsic parameters $\mathbf{K} = \begin{bmatrix} f & 0 & a \\ 0 & f & b \\ 0 & 0 & 1 \end{bmatrix}$, the epipolar line $\mathbf{l}_i$ in the reference image corresponding to a point $\boldsymbol{p}_i$ in the target image can be calculated as:

$$\boldsymbol{l}_i = \mathbf{E}\tilde{\boldsymbol{p}}_i = \mathbf{R}[\boldsymbol{t}]_\times \tilde{\boldsymbol{p}}_i, \tag{1}$$

where $\mathbf{E}$ is the essential matrix, $[\boldsymbol{t}]_\times$ is the skew-symmetric matrix representation of the translation vector $\boldsymbol{t}$, and $\tilde{\boldsymbol{p}}_i = \mathbf{K}^{-1}\boldsymbol{p}_i$ is the point $\boldsymbol{p}_i$ in the normalized image coordinates.

Now, expressing $\tilde{\boldsymbol{p}}_i$ in terms of $\boldsymbol{p}_i$ and $\mathbf{K}$:

$$\begin{aligned} \tilde{\boldsymbol{p}}_i &= \mathbf{K}^{-1}\boldsymbol{p}_i \\ &= \begin{bmatrix} 1/f & 0 & -a/f \\ 0 & 1/f & -b/f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} x/f - a/f \\ y/f - b/f \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} (x-a)/f \\ (y-b)/f \\ 1 \end{bmatrix}. \end{aligned} \tag{2}$$

Substituting this into the equation for $\boldsymbol{l}_i$:

$$\boldsymbol{l}_i = \mathbf{R}[\boldsymbol{t}]_\times \begin{bmatrix} (x-a)/f \\ (y-b)/f \\ 1 \end{bmatrix}. \tag{3}$$

Here, the coordinates $(x-a)/f$ and $(y-b)/f$ are simply scaled versions of the original image coordinates x and y, and this scaling does not affect the linearity of the equation. Therefore, the final expression for $\boldsymbol{l}_i$ does not explicitly depend on f.

(a) No clear correspondence (b) Occlusion

Figure B.1. When the occlusion occurs, or there is no clear geometric or semantic corresponding, epipolar attention tends to give multiple semantically similar points close similarity scores.

B. Property of the Epipolar Attention

To better understand our epipolar attention mechanism, we performed a visual analysis of the attentional weights in various cases. In Fig. B.1, two pairs of images show that our epipolar attention tends to give multiple semantically similar points close similarity scores when a point is occluded or when there is a lack of explicit geometric or semantic correspondence between the two points in the target and reference images. This behavior suggests that our method employs a broader range of contextual features, a favorable approach without explicit correspondences.

C. Different Features for Similarity Calculation

As discussed in Section 4.2 of our main paper, the similarity score derived from the output feature $\boldsymbol{F}$ of the attention block does not align well with our intended application, as it produces a relatively uniform similarity map. Instead, using the query $\boldsymbol{Q}$ from the target branch and the key $\boldsymbol{K}$ from the reference branch within the multi-head self-attention block provides a more accurate correspondence. This is illustrated in Figure C.1.

D. Results on More Datasets

We conduct experiments on the Objaverse dataset [1]. Specifically, we randomly sample 100 objects from the Objaverse test set, utilizing the camera setting of 16-views

Table D.1. Comparison of multi-view consistency, image quality, and input consistency on Objaverse test set. The camera setting is the same as SyncDreamer [3]. The results show that our method has similar consistency scores to SyncDreamer, but higher quality scores and input consistency scores.

	Multi-view Consistency			Quality Score			Input Consistency
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	LPIPS↓
Zero123	19.271	0.769	0.324	19.533	0.808	0.162	0.265
SyncDreamer	**23.827**	**0.849**	**0.257**	19.198	0.824	0.175	0.259
Ours	23.341	0.830	0.263	**21.147**	**0.830**	**0.144**	**0.235**

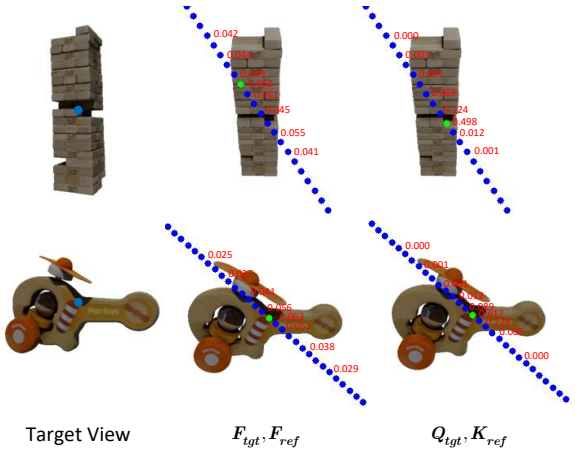

Figure C.1. Similarity scores using different features. Similarity scores computed using queries and key features in the self-attention block are sharper and more accurate than those computed using the output features of the attention block.

with a fixed camera pose, which aligns with SyncDreamer's setup for fair comparison. The results are presented in Tab. D.1 and share the same conclusion with the exprimences on GSO [2] dataset. Specifically, compared with our baseline model (Zero123), our method significantly improves the multi-view consistency, image quality, and input consistency on the Objaverse dataset. Compared with SyncDreamer, we achieve similar multi-view consistency but better image quality and input consistency. These results demonstrate the efficacy of our approach across different datasets.

E. More Ablation Studies

E.1. Number of Context Views

The quantity of context views, denoted as M, may influence the consistency of synthesized multi-view images. Ablation studies are conducted to examine the impact of varying numbers of context views, and the results are presented in Tab. E.1. It is evident that in the absence of context views (our baseline), the consistency is poor. As the number of context views increases, the consistency improves. However, as the context number is continuously increased, the consistency score decreases. This decline may be due to significant relative camera pose transformations, resulting in smaller overlapping regions between two views. Retrieving information from these views may adversely affect performance.

Table E.1. Ablation study on the effect of the number of context views used.

	PSNR↑	SSIM↑	LPIPS↓
0 (Baseline)	16.556	0.682	0.378
1	20.630	0.767	0.308
2 (Ours)	**21.151**	**0.780**	**0.302**
3	20.937	0.772	0.311
4	20.678	0.770	0.306
5	20.450	0.773	0.305

Table E.2. Ablation study on using different features for matching.

	PSNR↑	SSIM↑	LPIPS↓
Baseline	16.556	0.682	0.378
Output Features	20.045	0.771	0.327
Query, Key	**21.151**	**0.780**	**0.302**

E.2. Effect of Using Different Features

In Fig. 4 of our main paper, we visually compare the similarity scores obtained using different features, *i.e.*, employing query key features within the self-attention blocks and output features of the self-attention layers. Here, a quantitative comparison is conducted to demonstrate the impact of employing distinct features. The results in Tab. E.2 illustrate that utilizing query key features shows better consistency performance than using the output features from the self-attention layers, as they better locate the corresponding features.

E.3. Effectiveness on Different Overlap Ratios

In Section 5 of our main paper, we present three different view sampling methods used in our experiments. These

Table E.3. The effectiveness of our method when the target view has different overlap ratios with the input view. Our method consistently demonstrates improvements over the baseline across various overlap ratios, even when no overlap exists.

Overlap Ratio	0.7	0.4	0.1	0(no overlap)
baseline	17.089	15.296	14.354	13.350
ours	**17.214**	**15.678**	**14.603**	**13.448**

methods ensure that each view sufficiently overlaps with its neighboring views, facilitating the transmission of overlapping information. Here, we vary the overlapping ratio between the target and input views during the single-view synthesis process to examine the impact of different overlapping ratios. The results in Tab. E.3 show that our method consistently demonstrates improvements over the baseline across various overlap ratios. Notably, even in scenarios where there is no overlap between the reference and target views, our method obtains performance gains over the baseline. This can be attributed to our approach of utilizing the DDIM inverted noise from the reference view as the initial noise for the target view, thereby incorporating additional information from the reference view.

E.4. Other Hyperparameters

In regards to the feature fusion weight α, the step T, and the U-Net layer L after which we inject our epipolar attention layer, we conduct preliminary tests with various values on a few numbers of objects, ultimately selecting those that yield more visually appealing results. We do not attempt to determine the optimal values across the entire test set, as this approach is impractical. Furthermore, it is acknowledged that different objects may necessitate distinct hyperparameter values for better performance.

F. Application in Image-to-3D Task

To further validate the effectiveness of our method on downstream applications, we apply our method to the image-to-3D task and compare the results with our baseline Zero123. Specifically, given a single image, we use the output noise of our method and Zero123 to distill the NeRF [4] training process. We follow the method proposed in DreamFusion [5]; please refer to this paper for more details. The results in Fig. H.2 show that our method generates 3D objects with better geometric and texture details, especially the parts that are not visible in the input view.

G. Limitations

Utilizing our epipolar attention to locate and retrieve corresponding information in the reference views enhances the consistency between generated multi-view images compared to the baseline model. Nevertheless, our method cannot ensure absolute consistency in the generated images due to the inherent probabilistic nature of the diffusion model, which remains unchanged. Employing multiple model runs and selecting superior results may further enhance consistency.

Here we further discuss failure cases in more detail. 1) Illustrated in the first set of images in Fig. G.1, our method encounters situations where severe inconsistencies exist in the baseline model, impeding its ability to well rectify these inconsistencies even when reference information is injected during the image generation process. In real-world applications, tuning the feature fusing weight α for a specific object may acquire better consistency results. 2) Illustrated in the second set of images in Fig. G.1, despite the substantial improvement in consistency achieved by our method in the generated multi-view images, our approach may encounter challenges maintaining absolute consistency, particularly when dealing with objects exhibiting complex textures. This limitation could stem from the inadequacy of the baseline model. Notably, our experiments demonstrate that even when a zero camera translation is provided to the model, it struggles to accurately reconstruct the input image in the presence of complex textures.

Besides, our auto-regressive generation pipeline naturally increases inference time. On a single NVIDIA A100, Zero123 generates a single image in 3 seconds, while our method takes 5 seconds. For 16 views, Zero123 takes 14 seconds due to batch processing, whereas our auto-regressive generation takes 55 seconds. However, considering the alternative of unaffordable re-training whenever a stronger baseline model becomes available, the runtime increase of our method is acceptable, as it significantly improves consistency and enables the generation of arbitrary views.

H. More Visualization Results

More Reconstruction Results. We present additional 3D reconstruction results in Fig. H.1. These results illustrate that by increasing the consistency in the generated multi-view images, directly training 3D models using these images yields plausible 3D mesh representations.

More Qualitative Comparisons of Synthesized Multi-View Images. The results in Fig. H.3 and Fig. H.4 further provide comparisons of the multi-view images synthesized by the baseline model and our method. In these two figures, the images positioned on the left-hand side represent the input image. In each group of images, the images in the first row depict results generated by the baseline model (Zero123), while those in the second row display results obtained from our approach. The comparisons show that our method improves the consistency of generated multi-view images on different datasets.

The results in Fig. H.5 provide additional comparisons

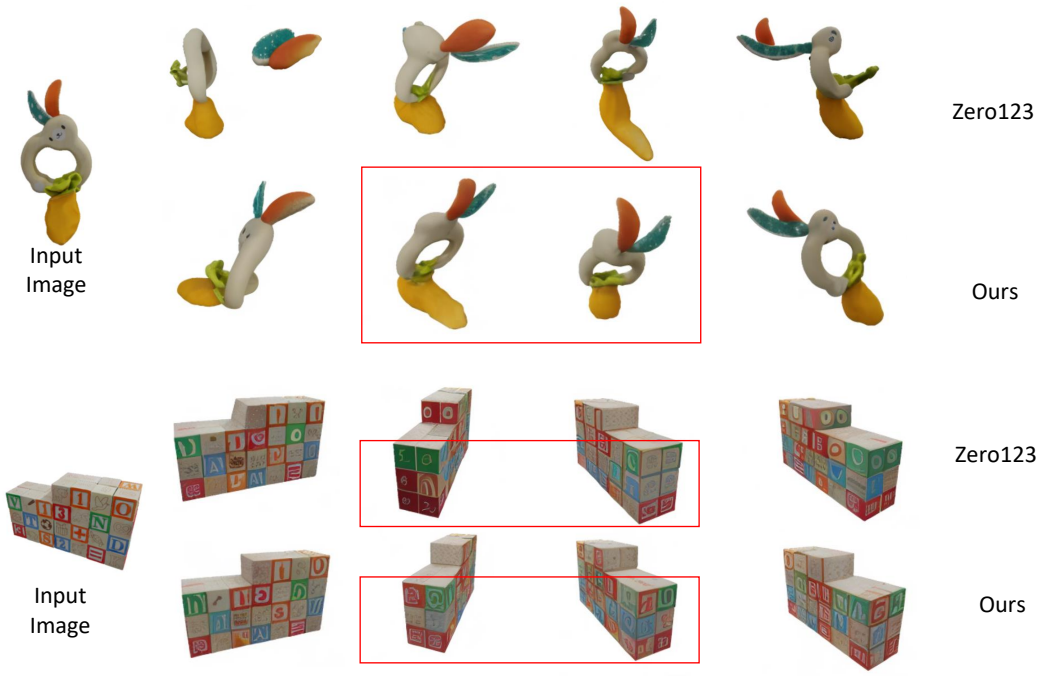

Figure G.1. Failure cases. We provide an in-depth analysis of failure cases arising when the baseline model exhibits severe inconsistencies or when dealing with objects with complex textures.

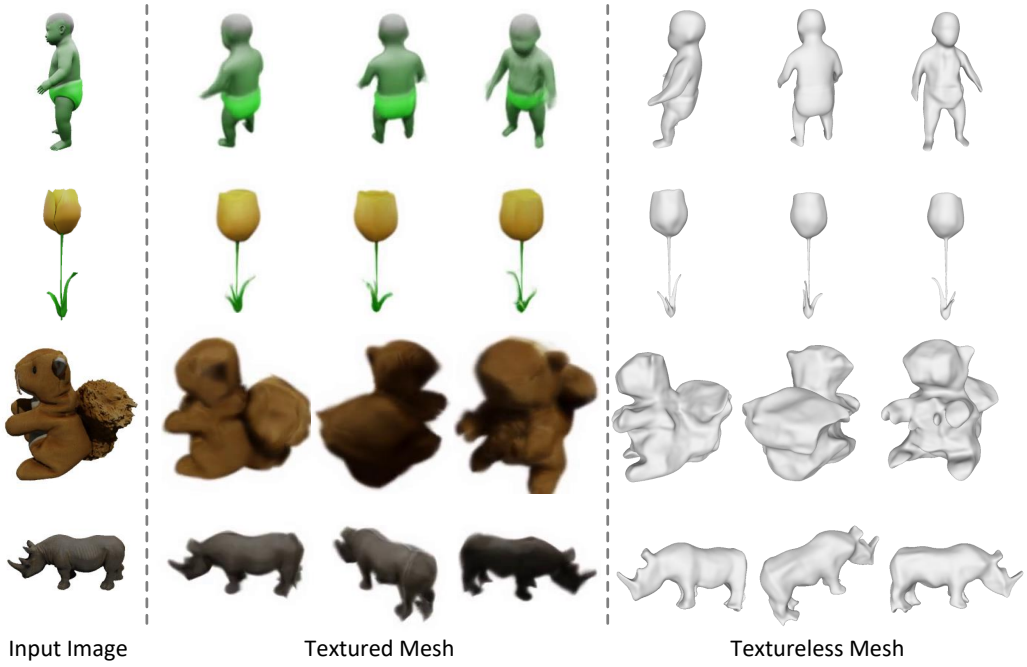

Figure H.1. More 3D reconstruction results.

between Zero123, SyncDreamer, and our method, demonstrating that our method significantly improves multi-view consistency compared to Zero123, while also exhibiting better image quality compared to SyncDreamer.

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. In *CVPR*, pages 13142–13153, 2023. 1

[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*, pages 2553–2560, 2022. 2

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

[4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 3

[5] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3

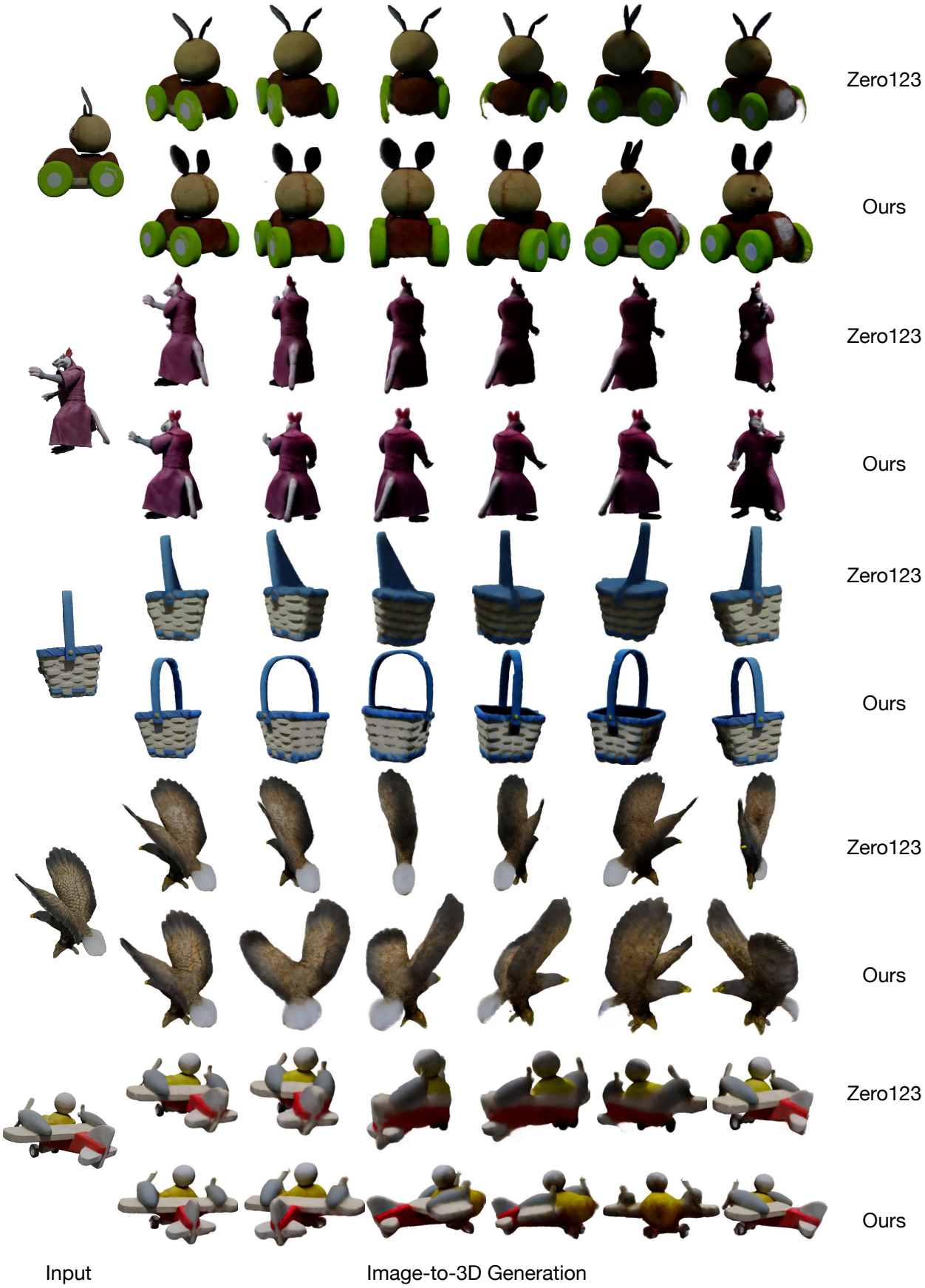

Figure H.2. Image-to-3D generation results. In each group of images, the images in the first row depict results generated by the baseline model (Zero123), while those in the second row display results obtained from our approach. The results show that our method generates better 3D objects, especially the parts of the object not seen in the input view.

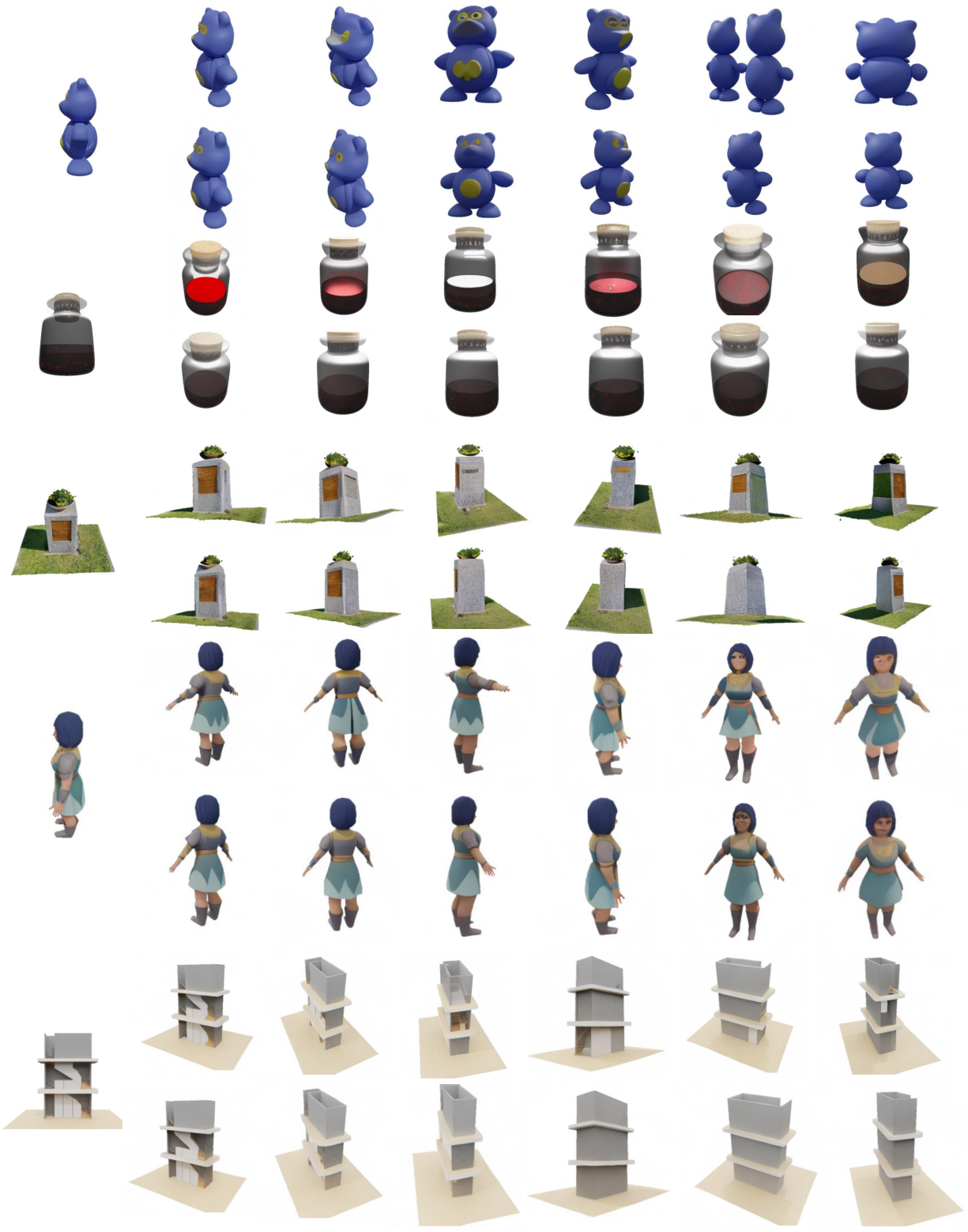

Figure H.3. Qualitative comparison with the baseline for generating a sequence of novel view images on the Objaverse dataset. The images positioned on the left-hand side represent the input image. In each group of images, the images in the first row depict results generated by the baseline model (Zero123), while those in the second row display results obtained from our approach. The comparison demonstrates that our method can generate multi-view images with higher consistency.

Figure H.4. More Qualitative comparison with the baseline for generating a sequence of novel view images on the GSO dataset. The image placement aligns with Fig. H.3.

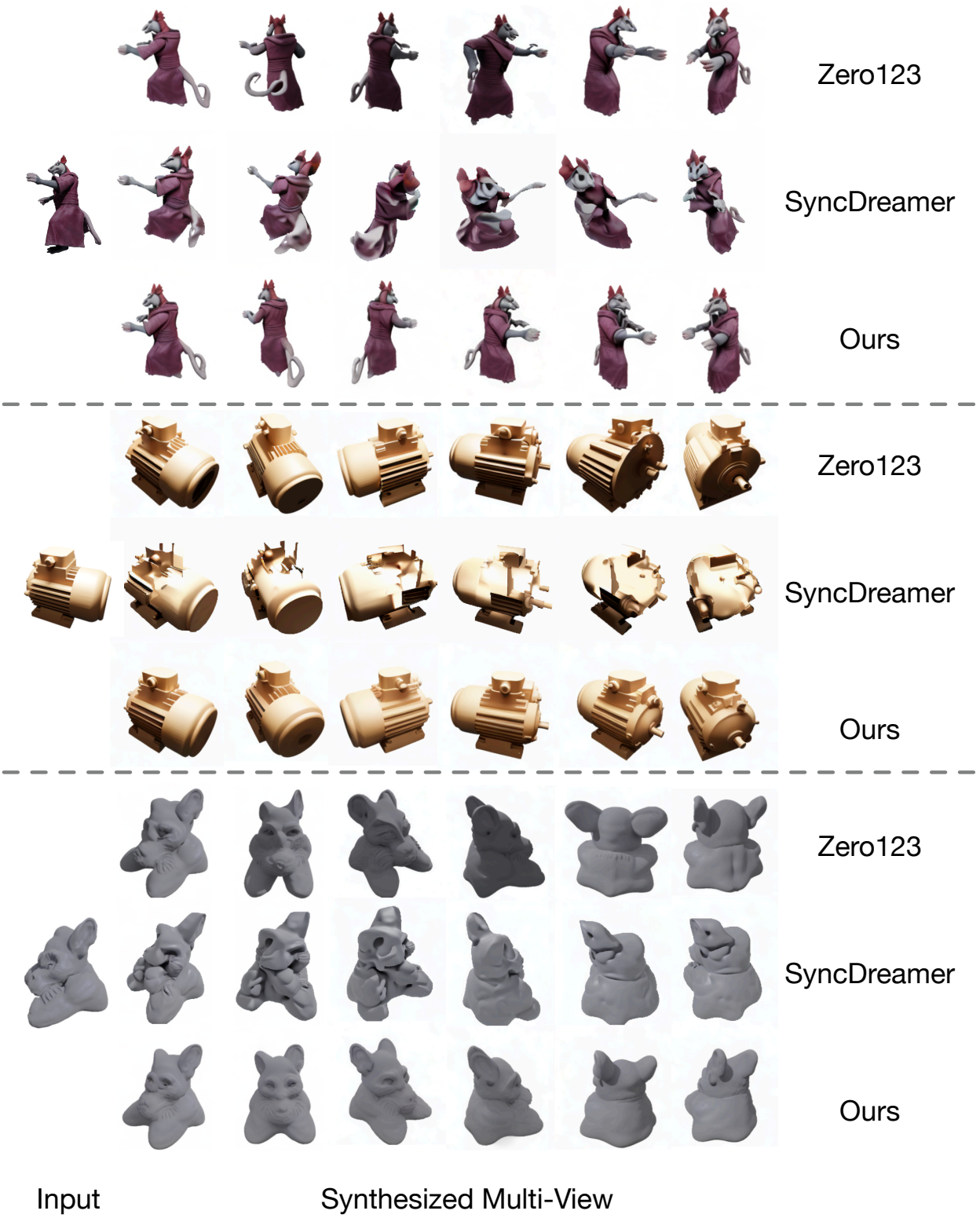

Figure H.5. More Qualitative comparison with Zero123 and SyncDreamer. The results show that our method significantly improves multi-view consistency compared to Zero123, while also exhibiting better image quality compared to SyncDreamer.