Supplemental Materials for ComboVerse: Compositional 3D Assets Creation Using Spatially-Aware Diffusion Guidance

Yongwei Chen[1*†], Tengfei Wang[2*], Tong Wu[3],
Xingang Pan[1], Kui Jia[4], and Ziwei Liu[1]

[1] S-Lab, Nanyang Technological University
[2] Shanghai Artificial Intelligence Laboratory
[3] The Chinese University of Hong Kong
[4] The Chinese University of Hong Kong, Shenzhen
https://cyw-3d.github.io/ComboVerse/

A Quantitative Ablation Analysis

To evaluate the effectiveness of the proposed SSDS, we performed an ablation analysis and have shown qualitative results in Fig. 10 in the main paper. Beyond visual comparison, we also provide quantitative ablation analysis in table S1. Two CLIP models, including CLIP B/16 and ResNet50 backbones, are utilized to examine different guidance configurations in object combinations. To disentangle geometry and appearance quality, we evaluate multi-view CLIP similarities for both colored rendering and untextured geometry rendering. *Base* only imposes reconstruction loss in the reference view, lacking additional guidance in novel views. Applying either a standard *SDS* loss or *depth* loss from a depth prediction model as spatial guidance yielded sub-optimal CLIP scores for color and geometry. However, by strengthening the attention to spatial layout through the proposed *SSDS loss (full)*, the full model achieves the best result, confirming its enhanced spatial control over standard SDS. As discussed in Sec. 3.3, high noise intervals ([800, 900]) were selected for Stable Diffusion during SSDS due to their bigger impact on the spatial layout of a generated image. We also experiment with SSDS with different sample ranges of noise timesteps, *low noise range* ([100, 200]), and *uniform noise range* ([20, 980]) and observe a performance drop.

B Validation of improvement on leaking pattern.

Fig. S1 illustrates our improvements in addressing the leaking pattern issue. As shown in the image, the side view of the Wonder3D reconstruction exhibits leaking patterns in both geometry and texture, while our reconstruction method effectively mitigates this phenomenon. Leaking patterns are common in other

[*] Equal contribution.

[†] Work done when interning at Shanghai Artificial Intelligence Laboratory.

Table S1: Quantitative analysis for ablation study.

Guidance	CLIP Score↑			
	CLIP B/16		ResNet50	
	Color	Geometry	Color	Geometry
Base (without guidance)	86.62%	75.24%	80.35%	74.19%
Depth Loss	84.57%	78.42%	81.69%	75.83%
SDS	84.16%	78.25%	84.08%	74.66%
SSDS (uniform noise range)	85.33%	78.49%	85.55%	75.85%
SSDS (low noise range)	84.86%	79.03%	84.42%	75.44%
SSDS (full)	**89.01%**	**79.66%**	**86.60%**	**78.10%**

methods when reconstructing complex compositions, as evidenced in Fig. 2(c) and Fig. 6 of our paper. In contrast, our method significantly alleviates this problem by reconstructing each object separately.

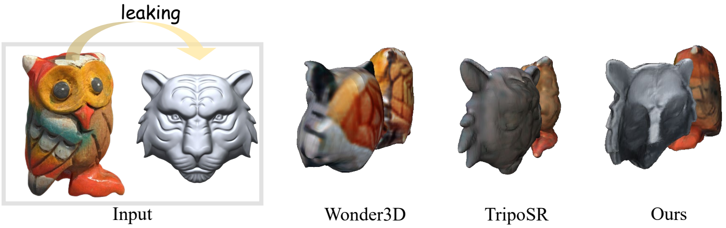


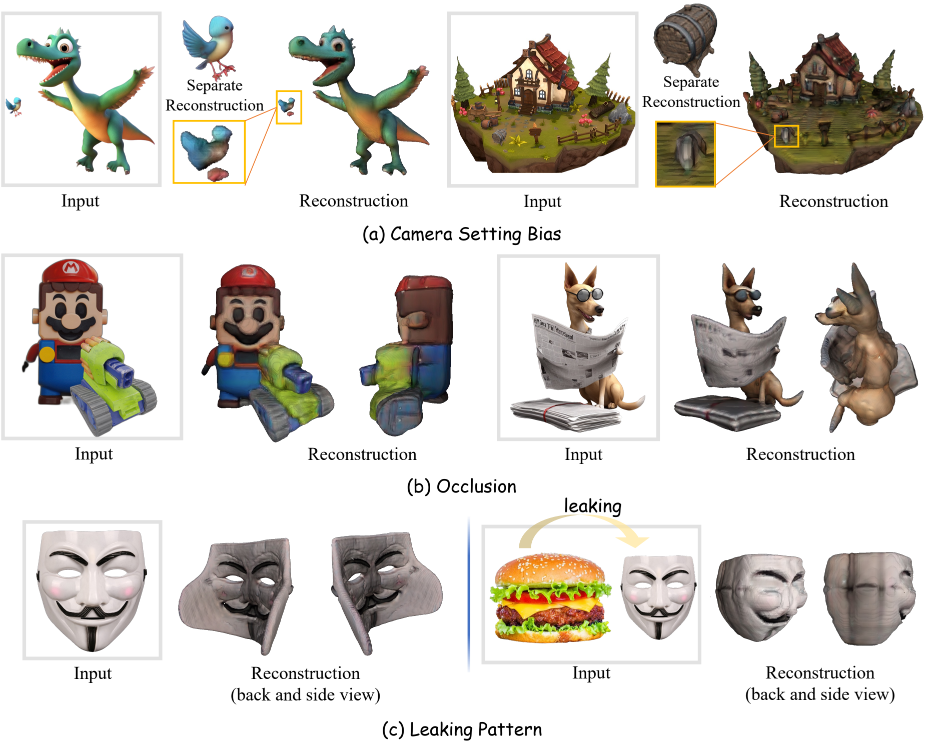
Fig. S1: Validation of improvement on leaking pattern.

C More Results of "Multi-Object Gap"

As discussed in Sec.3.1, the current feed-forward models, mainly trained on Objaverse, exhibit limitations in generalizing to multi-object scenarios. Fig. S2 uses TripoSR, a state-of-the-art method in image-to-3D reconstruction, as another case to demonstrate this limitation. Despite its advancements, TripoSR still exhibits three typical failure modes when tasked with generating multiple objects, stemming from inherent data and model biases. The detailed analysis was illustrated in Sec. 3.1.

D User Study

Besides numerical metrics, we also perform a user study to compare our method with others. We collect 990 replies from 22 human users. Participants are shown a reference image and a random pair of 3D models (ours and baselines) at once and are asked to select a more realistic one in terms of both geometry and texture



(a) Camera Setting Bias

(b) Occlusion

(c) Leaking Pattern

Fig. S2: "Multi-object gap" of models trained on Objaverse. (a) Camera Setting Bias. The reconstruction quality for small and non-centered objects will significantly downgrade compared to separate reconstruction. (b) Occlusion. The reconstruction results tend to blend when an object is occluded by another. (c) Leaking Pattern. The shape and texture of an object will be influenced by other objects in the input image.

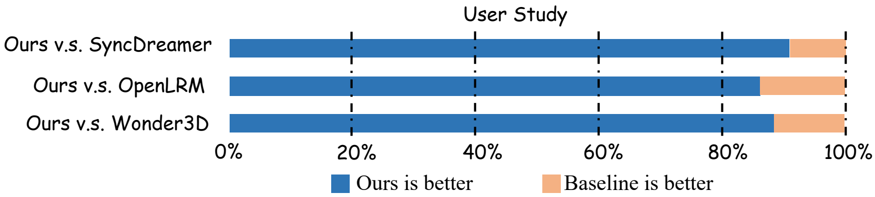
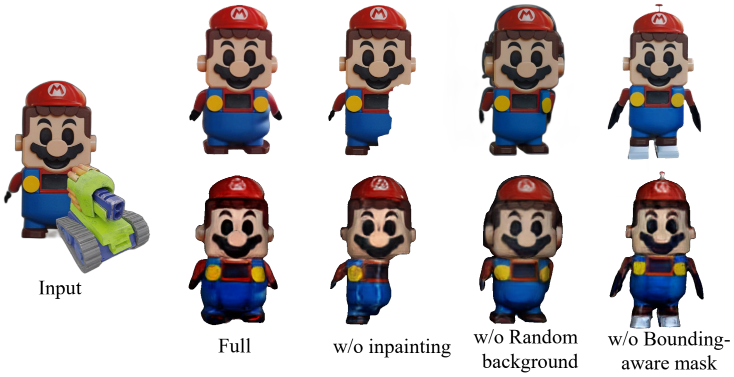


Fig. S3: User study. Our method consistently outperforms competitors in terms of human evaluation.

quality. All choices are given in a shuffled order without time limitation. Fig. S3 illustrates that our method outperforms previous approaches in terms of human preference.

E More analysis of the inpainting step

Besides ablation studies in Sec. 4.3, Fig. S4 further shows how the inpaint results affect the final 3D reconstruction. As shown in the third column, without the inpainting, the reconstruction results are incomplete due to occlusions. The fourth and fifth column shows that our proposed techniques significantly improve the inpainting effects, thereby ensuring higher quality 3D reconstruction.



Input

Full

w/o inpainting

w/o Random background

w/o Bounding-aware mask

Fig. S4: Validation of the inpainting step.