SceneMaker: Open-set 3D Scene Generation with Decoupled De-occlusion and Pose Estimation Model

Supplementary Material

A. De-occlusion Datasets

The curation pipeline of our de-occlusion datasets is shown in Figure 10. We first generate detailed captions with GPT [2] for each predefined class. Then, we generate corresponding images with FLUX [17]. Finally, we design three mask patterns for occlusion: object cutouts without backgrounds for object occlusion, right-angle cropping for image borders, and random brush strokes for user prompts, as shown in Figure 9. We also randomly resize the objects and the whole images to simulate patterns of small objects and low-resolution images.

B. Open-set Scene Datasets

Each scene is rendered using Blender's CYCLES engine from 20 viewpoints at a 512 resolution, with camera elevations randomly sampled between $[15, 60]$ degrees. Simultaneously, we uniformly sample 20K random points on the object's surface to serve as the input geometric information for our object generation module. We also augment the image backgrounds through random selection. To ensure physical plausibility, we place the lowest point of each object on the same plane and enforce that their bounding boxes do not intersect. We present samples from our dataset in Figure 11. We randomize the pitch angle of input meshes during training to better align the generated 3D objects.

Figure 9. **Occlusion patterns of the de-occlusion datasets.**

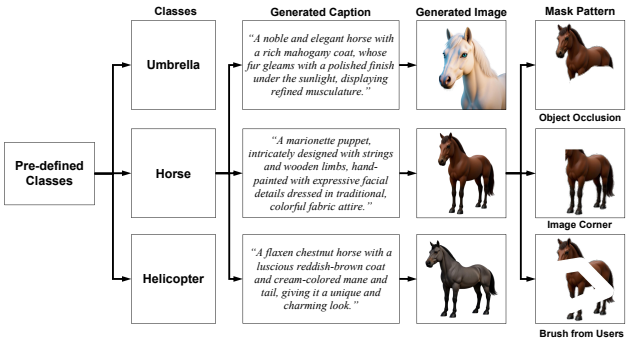

Figure 10. **Curation pipeline of the de-occlusion datasets.**

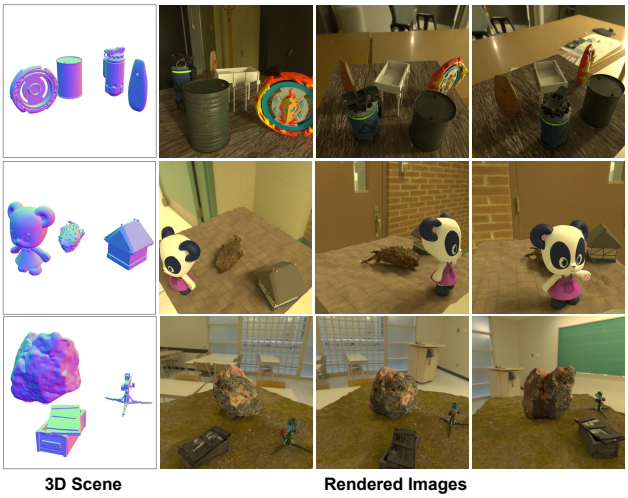

Figure 11. **Samples from the collected open-set dataset.**

C. More Ablation Studies

C.1. Component Contribution

Although some contributions of modules are discussed across Tables 1, 2, 3, and 5 in the main paper, we consider that a single component contribution table would enhance clarity. Specifically, we perform the ablation starting from a baseline (Amodal3R + pose model w/o GSA/LSA) and incrementally add the open-set data, the decoupled de-occlusion model, and the pose estimation attention mechanisms. As shown in Table 6, our results demonstrate that each contribution yields considerable and consistent performance gains. We will add this component contribution ablation to the final version.

Method	CD-S↓	FS-S↑	CD-O↓	FS-O↑	IoU-B↑
Baseline	0.1501	0.3429	0.3623	0.2171	0.6448
+ Open-set Data	0.0387	0.5247	0.3419	0.2704	0.6948
+ Decoupled De-occlusion	0.0363	0.5662	0.0707	0.5752	0.7020
+ Attention Mechanisms (Ours)	**0.0285**	**0.6125**	**0.0671**	**0.5948**	**0.7549**

Table 6. Ablation study of component contributions.

C.2. Controllable Object Generation

Benefiting from our decoupled de-occlusion model, compared with 3D-native methods [22, 53], our model further

enables the controllable generation of the occluded areas of objects through prompts. As shown in Figure 12, our model is able to control the color of the pot and items in the penguin's hand through prompts during de-occlusion.

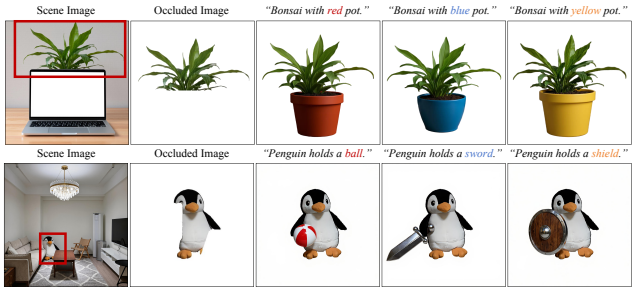

Figure 12. Text-controllable object de-occlusion.

C.3. Robustness to Perception Noise

To simulate segmentation noise, we randomly apply morphological dilation or erosion (1–2 iterations) to the masks using a 3×3 kernel. For depth noise, we inject Gaussian perturbations with a standard deviation set to 5% of the depth map's standard deviation. As shown in Table 7, our framework demonstrates exceptional robustness to both segmentation inaccuracies and depth noise.

Noise Type	CD-S↓	FS-S↑	CD-O↓	FS-O↑	IoU-B↑
None (SceneMaker)	**0.0285**	**0.6125**	**0.0671**	**0.5948**	**0.7549**
Segmentation Noise	0.0302	0.5794	0.0744	0.5870	0.7197
Depth Perturbation	0.0297	0.5857	0.0684	0.5730	0.6993

Table 7. Robustness to perception noise.

C.4. Computational Cost

We test inference on a single HGX A100 80GB GPU. Segmentation (1s) and depth (0.4s) are highly efficient. De-occlusion (20s/object) and 3D generation (10s/object) are the most time-consuming but can be parallelized. Including pose estimation (12s), the total scene generation takes about 40 seconds. The VRAM bottleneck is the de-occlusion model fine-tuned from Flux Kontext, requiring at least 35GB.

D. Limitations and Future Work

Although our framework effectively generalizes to arbitrary objects, the real-world arrangement of objects is often much more complex than what our datasets capture, particularly when force interactions are involved. Therefore, a key future research topic is how to construct or refine 3D scenes more accurately in a physically plausible manner, including handling interpenetration and force interactions. Meanwhile, existing methods can only control scene generation through images or simple captions, and further development is needed for more control signals and natural language interactions. Moreover, how to perform more in-depth understanding tasks and adapt embodied decision-making based on the generated high-quality 3D scenes is also an unsolved challenge.

E. More Results

We conduct qualitative comparisons on both indoor and open-set scenes in Figure 13 and Figure 14, including both synthetic and real-world captured images. Our method offers better generalization to open-set scenes. Moreover, it delivers more accurate poses and finer geometry under severe occlusion or for small objects.

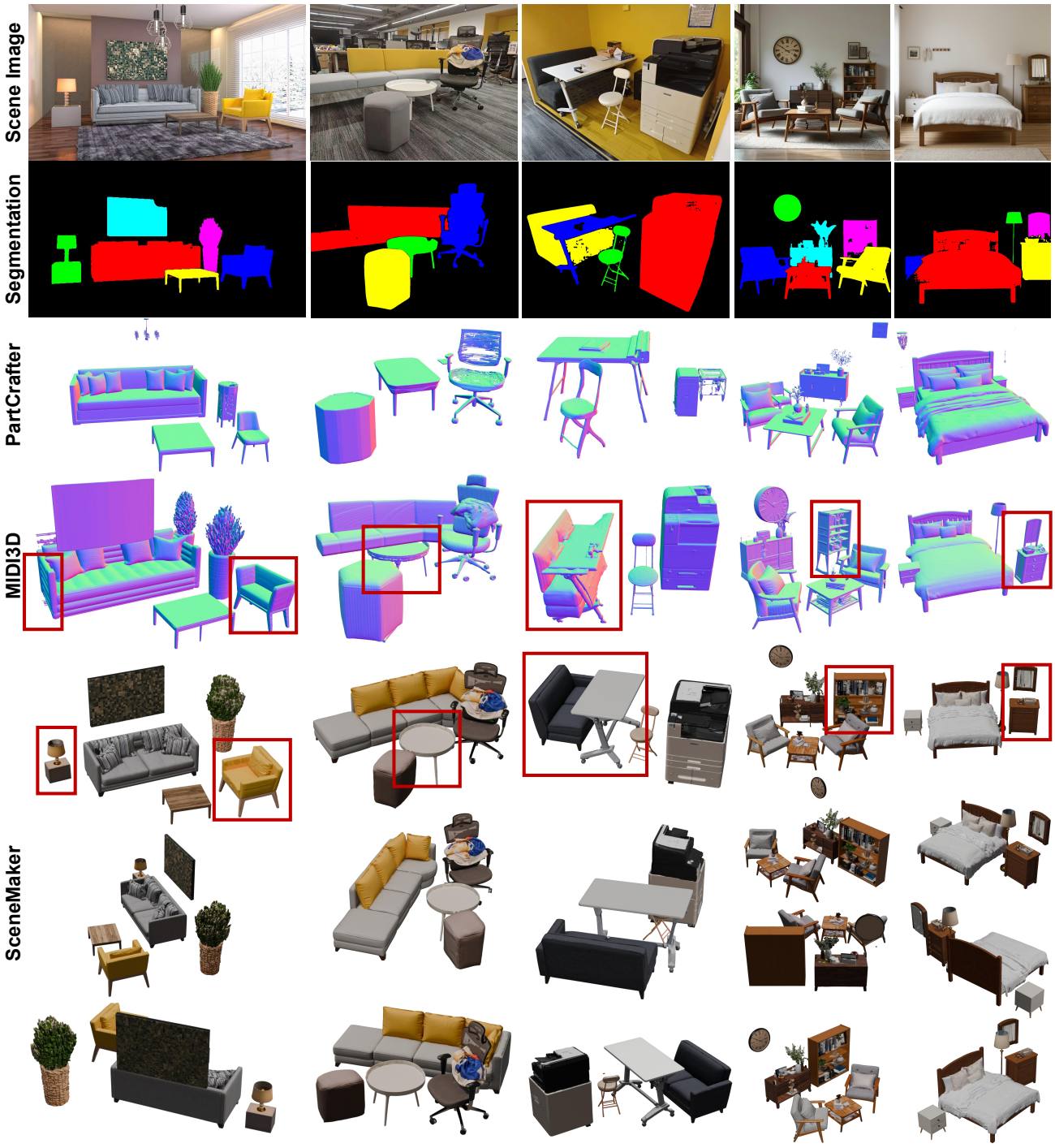

Figure 13. **Qualitative comparison with scene generation methods on indoor scenes.**

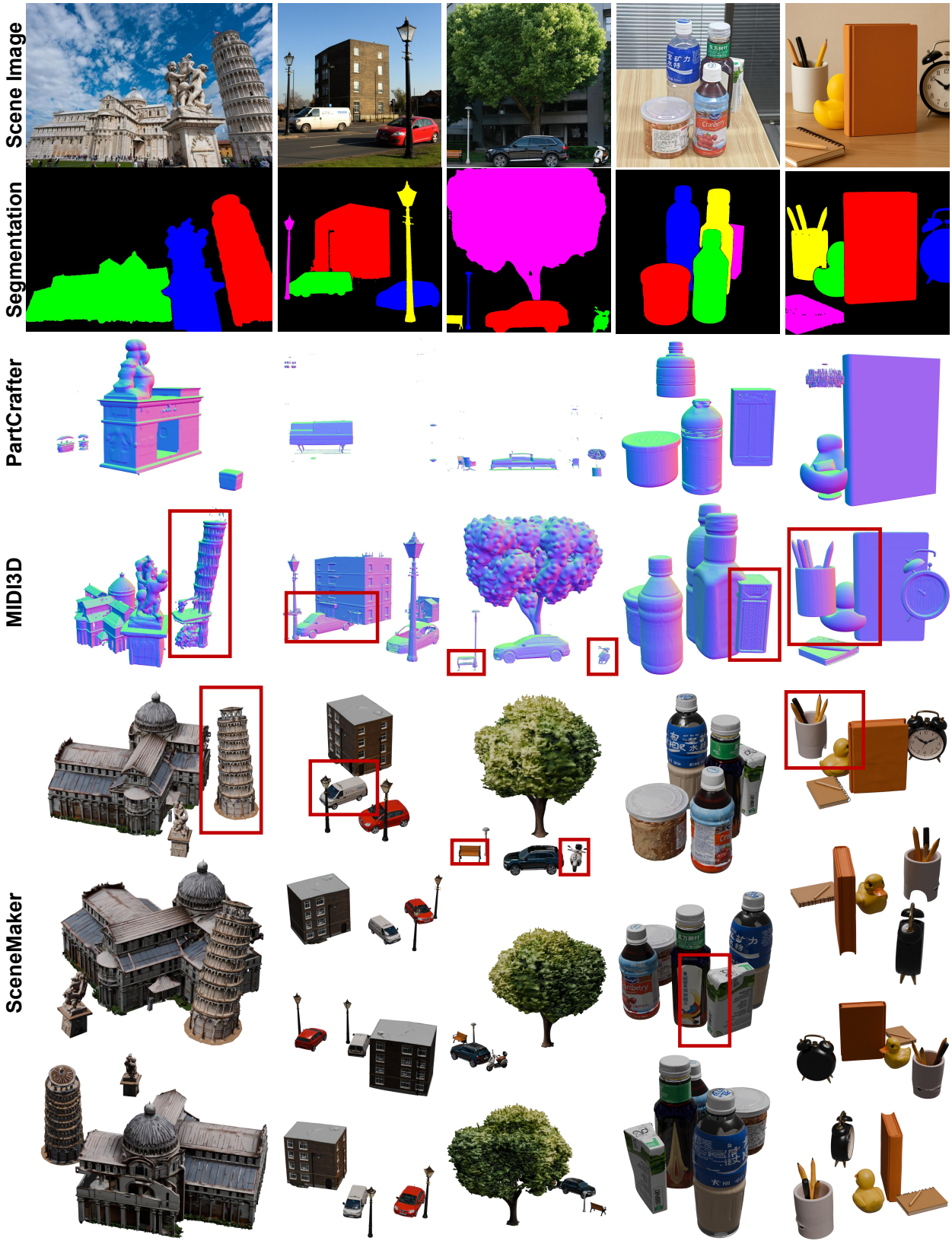

Figure 14. **Qualitative comparison with scene generation methods on open-set scenes.**