

# Pano3DComposer: Feed-Forward Compositional 3D Scene Generation from Single Panoramic Image

## Supplementary Material

### Overview

This supplementary material provides additional details, including dataset descriptions, implementation details, ablation studies, qualitative results, failure cases, and limitations to complement the main paper. We also provide supplementary videos showcasing qualitative results and rendered 3D scenes, which further demonstrate the effectiveness of our method.

### 7. Datasets

Our experiments involve panorama-to-3D scene composition on synthetic benchmarks and real-world panoramas. Below we summarize the synthetic datasets used for training and quantitative evaluation. 3D-FRONT [8] is a professionally designed dataset comprising high-quality textured furniture models arranged in realistic room layouts. Structured3D [48] is a photo-realistic synthetic dataset featuring rendered images under diverse lighting and furniture configurations, accompanied by rich annotations (semantics, albedo, depth, normals, and layout) but does not release object meshes. For real-world in-the-wild panoramas used in qualitative evaluation, we collect images from public online sources and ensure they are only used for non-commercial research visualization.

### 8. More Implementation Details

**Fine-tune of SceneGen.** To adapt SceneGen [27] for equirectangular panoramic (ERP) inputs, we follow its official data preprocessing pipeline and extend it to handle the equirectangular panoramas rendered from the 3D-FRONT dataset. A representative example of the processed panoramic input, including the panorama, instance masks, and object crops, is shown in Fig. 7. Our model is initialized from the official SceneGen pretrained checkpoint. We fine-tune the model using a global batch size of 8 and an initial learning rate of  $1 \times 10^{-5}$  with AdamW optimizer. The model is trained for 7 days on a single NVIDIA RTX 4090 GPU under mixed-precision (BF16) training.

**Inference.** In our experiments, the input equirectangular panoramas are at a resolution of  $512 \times 1024$ . For evaluation on 3D-FRONT and Structured3D, we directly use the ground-truth instance segmentation annotations provided by each dataset, in the same way as SceneGen [27]. For real-world in-the-wild data, we manually obtain instance masks using the 2D foundation model SAM [17]. To develop a fully automated pipeline, one may

Table 4. Ablation of fine-tuning strategies. “-D”, “-D-F”, and “-D-F-G” indicate progressively freezing DINO, frame, and global attention modules.

Method	CD-S ↓	CD-O ↓	F-Score-S ↑	F-Score-O ↑	IoU-B ↑
Full	0.1883	0.1946	0.4992	0.4907	0.3855
-D	0.1236	0.1177	0.5565	0.5550	0.4360
-D-F	<b>0.0787</b>	<b>0.0765</b>	<b>0.6923</b>	<b>0.6926</b>	<b>0.5679</b>
-D-F-G	0.1120	0.1063	0.5788	0.5850	0.4818

integrate open-vocabulary recognition models (e.g., RAM [47], various visual language models (VLMs)) with detection/segmentation models capable of grounding (e.g., GroundingDINO [25], SAM [17], Grounded-SAM [31]) to identify, localize, and segment objects directly on ERP panoramas. In the Object-World Transformation Predictor, we render 4 multi-view images for each object. We uniformly sample four horizontal viewing directions at azimuth angles  $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ , and apply a fixed  $20^\circ$  downward pitch. All renderings use a resolution of  $518 \times 518$ . These rendered views provide appearance-conditioned geometric cues that significantly stabilize the relative pose estimation stage.



Figure 7. Example inputs of SceneGen.








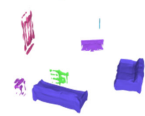






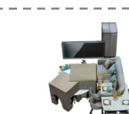
















3D-FRONT						
						
Structured3D						N/A
						N/A
Real World						N/A
		Failed				N/A
Input Panorama		DeepPanoContext	SceneGen	Pano3DComposer (Ours)	Pano3DComposer-C2F (Ours)	GT

Figure 8. Visualization of panorama-to-3D scene composition results without background.

## 9. More Experiments

### 9.1. Ablation of Trainable VGGT Modules.

We compare different fine-tuning strategies by freezing specific modules of VGGT [37] (Table 4). “Full” denotes full fine-tuning. “-D” freezes the DINO backbone; “-D-F” further freezes the frame attention layers; and “-D-F-G” also freezes the global attention layers. We find that keeping the global attention and camera/scale heads trainable (“-D-F”) yields the largest performance gains across all metrics.

### 9.2. More Visual Comparisons

Fig. 8 shows additional qualitative comparisons between our approach and baselines. To better illustrate the full-room generation capability beyond object synthesis, we provide more rendered videos in the supplementary attachment.

## 10. Failure Cases

When backgrounds exhibit complex geometry, clutter, or heavy occlusions, the inpainting network may fail to recover a clean room structure. This can lead to visible artifacts or incorrect structural completions. In addition, the Flash3D-based [35] monocular reconstruction is affected by

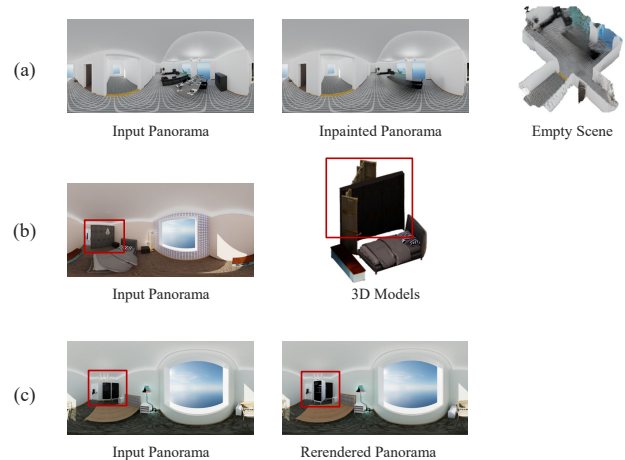


Figure 9. Failure cases. (a) Background inpainting and Flash3D-based monocular reconstruction failures. (b) Object generation failures. (c) Alignment failures.

the quality of depth estimation; inaccurate depth may lead to distorted backgrounds and other artifacts, as illustrated in Fig. 9 (a). Moreover, since the input panorama is constrained to a resolution of  $512 \times 1024$ , the extracted object crops often have relatively low resolution. As a result,

object generation models (e.g., TRELLIS [40]) may occasionally produce suboptimal outputs or even fail to generate plausible results (Fig. 9 (b)).

When the generated 3D object differs drastically from the observed object in the input panorama (in terms of geometry, silhouette, or texture), or when objects in the panorama appear at very low resolution, the alignment network may fail to reliably estimate the relative pose, resulting in misaligned insertions (Fig. 9 (c)).

## 11. Limitations

Our approach primarily targets indoor scenes. Very small items and highly articulated or multi-part objects can still exhibit residual misalignment. Highly glossy or transparent materials pose challenges for appearance modeling and silhouette consistency. Future work includes: (i) integrating physical awareness and multi-instance relation modeling, (ii) improving appearance and geometry prediction for transparent/specular objects, and (iii) scaling training data realism and diversity to further improve generalization.