Unreal3DSpace: A Unified and Controllable Framework for Comprehensive Evaluation of Spatial Understanding

Supplementary Material

A. Limitations and Future Work

One limitation of Unreal3DSpace is the scarcity of high-quality 3D scene assets. While 3D object assets are readily available, *e.g.*, from Objaverse, full-scale scene assets remain rare, particularly those featuring high-fidelity materials and realistic spatial layouts. A further limitation is the reliance on Unreal Engine's Chaos Physics. Although optimized for game performance, it is often not realistic enough for scientific research. We will incorporate external physics engines, such as MuJoCo, in future iterations.

Future work. We plan to investigate more scalable scene generation approaches, specifically by leveraging procedural generation for urban environments and adopting LLM-assisted synthesis pipelines. Furthermore, beyond serving as an evaluation framework, we plan to explore Unreal3DSpace as a training data engine. This would facilitate the generation of both in-distribution and valuable out-of-distribution data, featuring comprehensive 2D and 3D groundtruths, diverse occlusion levels, and varying lighting conditions.

B. Unreal3DSpace Data

B.1. Dataset Statistics

Our Unreal3DSpace is built on 24 high-quality 3D scenes obtained from the Unreal Engine Fab Marketplace. These scenes span a diverse range of domains featuring different object arrangements and background contexts, including urban cities, suburban neighborhoods, industrial facilities, and nature areas. Please refer to Table 1 for the full list of 3D scenes.

B.2. Qualitative Examples

We present the following qualitative results: (1) All 3D assets used in our Unreal3DSpace (Figure 8), (2) Simulation of different weather and lighting conditions (Figure 9), and (3) Qualitative examples from four scenes (Figure 10).

C. Quantitative Results

We report quantitative comparisons between visual foundation models for depth estimation in Table 3 and segmentation in Table 4.

D. Implementation Details

Baseline agentic model. The baseline agentic model used in Section is implemented with neural symbolic reasoner [38, 48] and executed with visual foundation models. In particular, the neural symbolic reasoner first parse the question into multiple executable programs. Each program involves either 3D parsing with visual foundation models or 3D computation from the intermediate results. As illustrated by the example in Figure 6, the agentic model starts by grounding the objects mentioned in the question. Then segmentation, depth estimation, and pose estimation models are used to extract 3D information about the objects. Lastly neural symbolic programs aggregate the intermediate results and predict spatial relationships.

Category	**List of 3D Scenes**
City (12)	City City Streets - Modular Pack Havana Street Hong Kong Street Japanese Street Miami Beach Nordic Harbour - Modular City Building Kit NYC East Village Russian Winter Town Street New York Tokyo Street Victorian Street
Rural or suburb (4)	Barnyard Megapack Medieval Village Megapack Slavic Village Tropical island
Nature (4)	City Park Environment Collection Mountain - Environment Set Northern Island Landscape Pack 4x4 km Zen Garden / 42 Assets
Industry (4)	Chemical Plant & Refinery Environment Factory Industrial Alley Old Factory

Table 1. **List of 3D scenes used in our Unreal3DSpace dataset.**

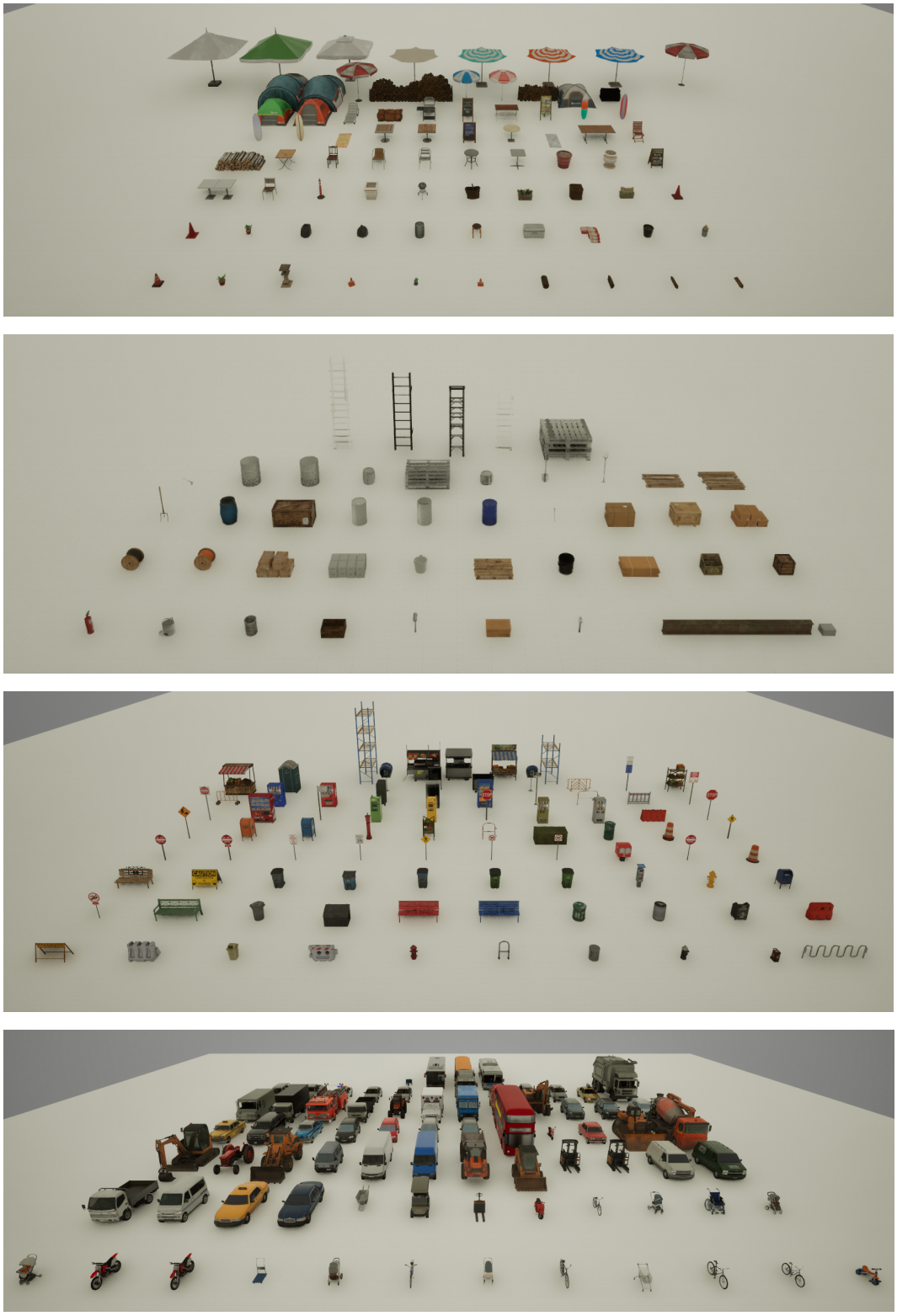

Figure 8. **Overview of 3D assets used in Unreal3DSpace.**

Figure 9. **Qualitative results on different weather and lighting conditions:** Clear day, dawn, fog level 1, fog level 2, and raining.

Figure 10. **Qualitative examples from Unreal3DSpace.**

Model	Overall	Height	Location	Orientation	Multi-Obj Reasoning
Cambrian-1-8B [36]	47.1	45.4	60.0	42.2	43.2
Qwen2.5VL-7B [4]	51.3	50.8	68.2	43.9	46.2
SpatialReasoner [20]	50.6	54.9	65.3	44.6	45.2
GPT-5	51.0	52.8	65.9	43.5	46.7

Table 2. **Spatial reasoning performance on Unreal3DSpace.**

Model	AbsRel ↓	δ_1 ↑
Depth Anything	2.172	0.630
DepthAnything v2	0.656	0.655

Table 3. **Depth estimation results on Unreal3DSpace.**

Model	IoU ↑
SAM	0.619
SAM 2	0.621

Table 4. **Segmentation results on Unreal3DSpace.**