Agent3D-Zero: An Agent for Zero-shot 3D Understanding

Supplementary Material

This supplementary material is made up of two sections. First, we visualize more examples to help intuitively present the question answering (QA) and 3D-assisted dialog results. Then, we provide more ablation study results to explore the impact of different 3D reconstructions qualities on the selection of camera poses and 3D Scene Understanding.

A Additional Qualitative Results

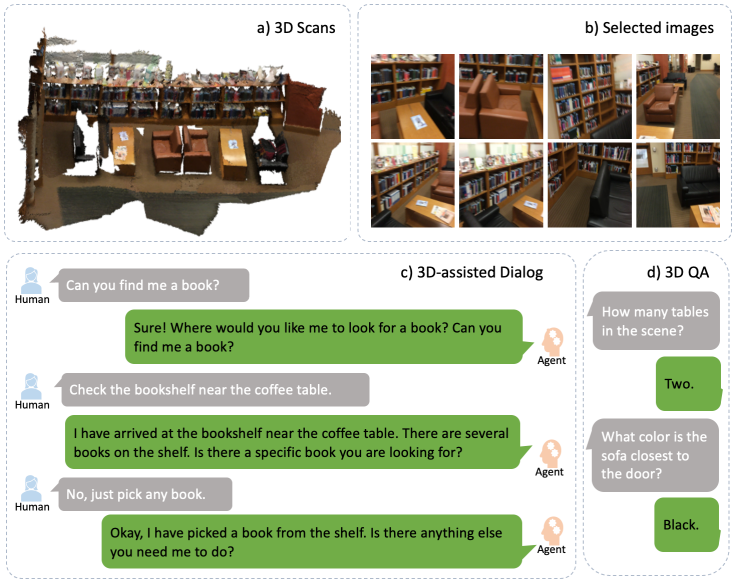

Fig. S1: Supplement of Qualitative results on QA and 3D-assisted dialog. The top presents the 3D scan of a reading room and part of the images selected from different viewpoints. We show examples of 3D-assisted dialog and 3D QA at the bottom.

In our main manuscript, we only visualize two tasks on one scene due to the limitation of pages. Here, we add more qualitative examples in Figure S1 to help intuitively present the reasoning results of our method. From the visualized results, it can be seen that our proposed method is able to answer the questions accurately according to the selected images of the specific scene. Meanwhile, the 3D-assisted dialog is presented at the bottom of the figure. These

examples vividly demonstrate Agent3D-Zero's adeptness at accurately identifying and describing detailed objects and their relationships within specific 3D environments. Through an intelligent analysis of the scene informed based on the selected images, Agent3D-Zero showcases its remarkable ability to analyze and summarization multiple objects' information and provide precise answers, underscoring its advanced 3D reasoning capabilities.

B Additional Ablation Study Results

To evaluate how the quality of 3D reconstructions affects the selection of camera poses, We add an experiment by using BEV images based on 3D reconstructions with 50% masking noise and then predict camera poses. We simulate one of the most common artifacts in 3D reconstruction by random masking. As shown in Table S1, the impact of 3D reconstruction quality on camera pose selection is insignificant.

Table S1: Effect of reconstruction quality.

	B-1	METEOR	ROUHE-L	CIDEr	EM
w/o noise	23.3	15.0	35.3	67.9	16.9
w/ noise	22.5	14.1	33.2	62.2	14.8

Additionally, we explore the impact of different 3D reconstructions qualities on 3D Scene Understanding. Discontinuities are a common phenomenon in the reconstruction process, and we apply random masking to simulate these low-quality reconstructions. We then apply the corresponding operations on images and evaluate the 3D QA task on the ScanQA dataset. Results presented in Table S2 indicate that adding noise leads to performance reduction in 3D scene understanding. Lower-quality 3D reconstructions result in lower-quality images. Lower-quality images provide less clear details, leading to less accurate 3D scene understanding.

Table S2: Effect of image quality.

noise	B-1	METEOR	ROUHE-L	CIDEr	EM
+0%	23.3	15.0	35.3	67.9	16.9
+1%	21.3	14.3	32.9	65.9	15.1
+5%	12.9	9.7	25.3	47.7	12.8
+10%	10.2	8.7	22.2	41.2	11.5
+25%	4.8	6.0	13.1	22.5	6.1