Open-Vocabulary Functional 3D Scene Graphs for Real-World Indoor Spaces

Supplementary Material

1. Details in the Method

1.1. Confidence-Aware Remote Reasoning

In Section 4.3 of the main paper, we introduce the methodology of confidence-aware remote relationship reasoning. Here, we provide two additional qualitative examples (see Figure 1) to demonstrate both the process and the effectiveness of the proposed strategy in handling complex spatial relationships with confidence-aware reasoning.

In the first example, the remote relationship between the light switch and the ceiling light is identified as the most confident. The Vision-Language Model (VLM) first evaluates the feasibility by focusing on the relative layouts, concluding that "the light switch is in the correct position to control the ceiling light fixture." The Large Language Model (LLM) then integrates all three feasibility descriptions into a global context, summarizing that "the light switch is more likely to control the ceiling light, as it is common for switches to control ceiling fixtures, and the relative positions are more plausible for a functional connection."

In the second example, the VLM directly observes that the kettle is plugged into the electric outlet and asserts that "the outlet is connected with the kettle." In contrast, it determines that the television is not connected to either the power outlet or the cord. Based on these observations, the LLM synthesizes the confidence results, summarizing that "the outlet and the kettle are in close proximity," and assigns high confidence to this relationship.

1.2. Other Implementation Details

We provide additional implementation details of our method, OpenFunGraph, in this section. In the node candidate description, the N_v we implement is 9, and we discard candidates that share fewer than N_v multi-views to ensure robustness in the multi-view summary. All experiments are conducted on a single NVIDIA 3090 GPU, primarily utilized for RAM++, GroundingDINO, and LLAVA inference. For a typical scene composed of 200 frames, the OpenFunGraph pipeline takes approximately 20 minutes to process. This includes 11 minutes for candidate detection (10 minutes) and fusion (1 minute), 8 minutes for LLAVA processes, and 1 minute for GPT-4 inference.

2. The Developed Annotation Tool

As introduced in Section 5 of the main manuscript, we developed an annotation tool for creating functional 3D scene graph annotations in the SceneFun3D and our proposed FunGraph3D datasets. The designs of the tool are showcased in Figure 2. Annotators can navigate the 3D scene and annotate instances of objects and interactive elements with a free-form label. They are also required to connect each interactive element to the corresponding object it controls and provide a description of their relationship. To assist in the annotation process, annotators can view iPad videos or egocentric human-scene interaction videos captured during data collection. These videos provide additional context, enabling annotators to identify objects, functional elements, and their relationships more accurately and efficiently.

	Overall Nodes		Overall Triplets	
Experiments	R@3	R@10	R@5	R@10
LLAVA v1.6 → v0	48.6	71.2	36.3	47.0
GPT v4 → v3.5	51.5	65.7	31.5	45.0
Ours	**73.0**	**82.8**	**60.4**	**70.3**
LLAVA v1.6 → v0	38.3	53.8	22.3	29.1
GPT v4 → v3.5	41.4	49.3	18.1	23.3
Ours	**55.5**	**65.8**	**29.8**	**45.0**

Tab. 1. Ablation study for the foundation models on SceneFun3D [1] (Top) and our FunGraph3D (Bottom).

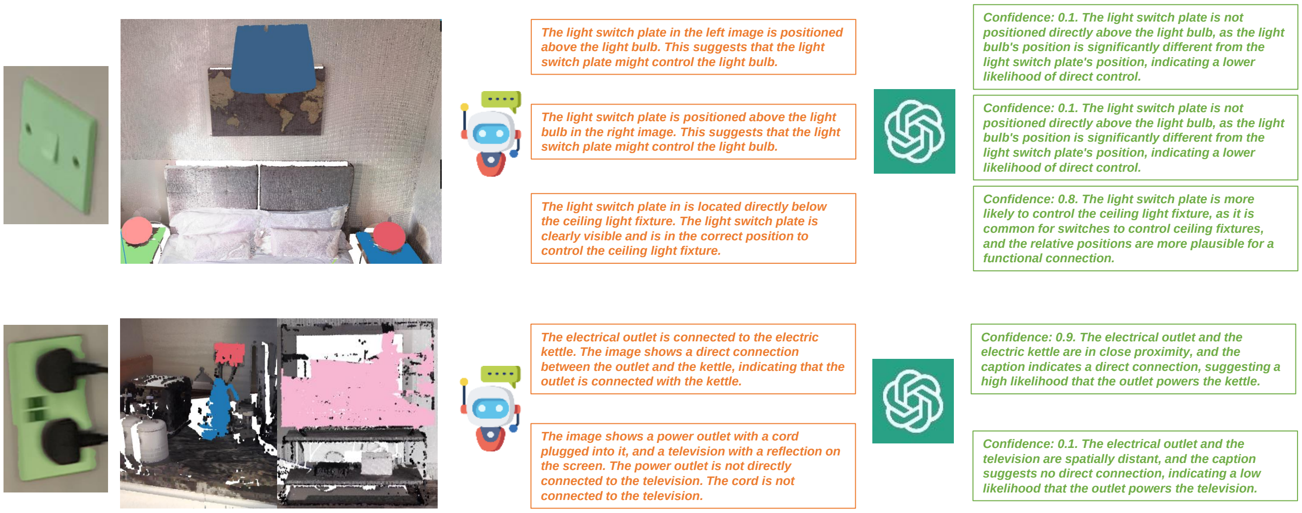
3. Additional Ablation Studies

We ablate different versions of the foundation models used, *i.e.*, the VLM and the LLM, to evaluate how their capabilities affect the understanding of indoor functional relationships, as shown in Table 1. A higher version of the VLM notably enhances node recognition, while improvements in the LLM substantially boost both triplet prediction and node recognition. This is achieved by strengthening the LLM's understanding of indoor functionalities and its responsiveness to prompts.

4. Limitations

The proposed pipeline relies on a 2D-based method due to limitations in 3D models' ability to accurately understand and localize small interactive elements, as illustrated by [1]. With advancements in 3D models, we aim to develop a fully 3D-based functional scene graph inference method in the future.

Additionally, due to resource constraints, the developed dataset is currently limited in scale, which restricts adaptive training for functional scene graph construction. We plan to significantly expand the dataset to address this limitation and support more robust and scalable training.



The light switch plate in the left image is positioned above the light bulb. This suggests that the light switch plate might control the light bulb.

The light switch plate is positioned above the light bulb in the right image. This suggests that the light switch plate might control the light bulb.

The light switch plate in is located directly below the ceiling light fixture. The light switch plate is clearly visible and is in the correct position to control the ceiling light fixture.

Confidence: 0.1. The light switch plate is not positioned directly above the light bulb, as the light bulb's position is significantly different from the light switch plate's position, indicating a lower likelihood of direct control.

Confidence: 0.1. The light switch plate is not positioned directly above the light bulb, as the light bulb's position is significantly different from the light switch plate's position, indicating a lower likelihood of direct control.

Confidence: 0.8. The light switch plate is more likely to control the ceiling light fixture, as it is common for switches to control ceiling fixtures, and the relative positions are more plausible for a functional connection.

The electrical outlet is connected to the electric kettle. The image shows a direct connection between the outlet and the kettle, indicating that the outlet is connected with the kettle.

The image shows a power outlet with a cord plugged into it, and a television with a reflection on the screen. The power outlet is not directly connected to the television. The cord is not connected to the television.

Confidence: 0.9. The electrical outlet and the electric kettle are in close proximity, and the caption indicates a direct connection, suggesting a high likelihood that the outlet powers the kettle.

Confidence: 0.1. The electrical outlet and the television are spatially distant, and the caption suggests no direct connection, indicating a low likelihood that the outlet powers the television.

Fig. 1. Illustration of the confidence-aware remote relationship reasoning.

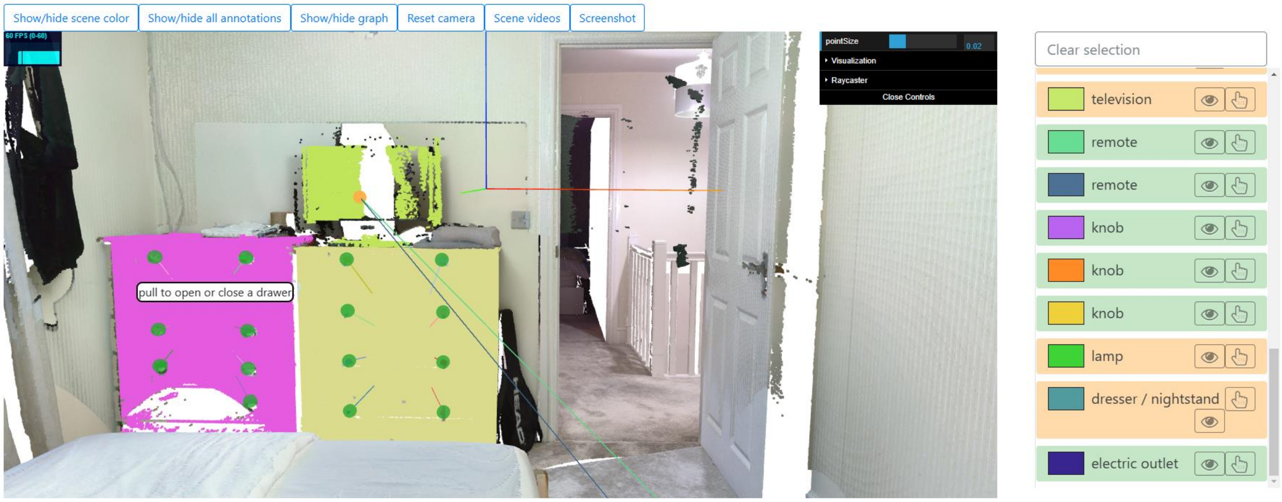


Fig. 2. Our developed annotation tool for functional 3D scene graphs.

References

[1] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1