Tool-Augmented VLM Agents for Zero-Shot 3D Visual Grounding

Cuong Huynh, Maxim Popov, Denis Gridusov, and Sergey Kolyubin

ITMO University

Introduction

3D Visual Grounding (3DVG) localizes objects in 3D scenes based on natural language descriptions (e.g., *"pick up the white porcelain sink next to the counter"*). This is essential for embodied AI.

Challenges in Existing Zero-Shot Methods:

- Unreliable logical reasoning due to early anchor-target matching errors.
- Non-selective visual analysis leading to token overhead.
- Untransparent exploitation of 3D geometric relationships.

Our Solution: *AgentGrounder*, a zero-shot tool-driven agent that leverages an Object Lookup Table (OLT) and selectively interleaves geometry parsing with on-demand visual rendering to resolve ambiguities robustly.

Pipeline

1. Offline Object Lookup Table (OLT): A 3D segmentation network (Mask3D) extracts instance-level masks and bounding boxes. This creates an OLT with IDs, semantics, centers, and box sizes.

2. Tool-Driven VLM Agent: An LVLM agent uses tools to execute the query without global dense evaluations.

- `write_plan`: Explicit step-by-step reasoning.
- `get_object_info`: Retrieves object boxes directly from the OLT.
- `calculate_distance`: Checks 3D proximity/geometry constraints.
- `see_image`: Triggers on-demand rendering.

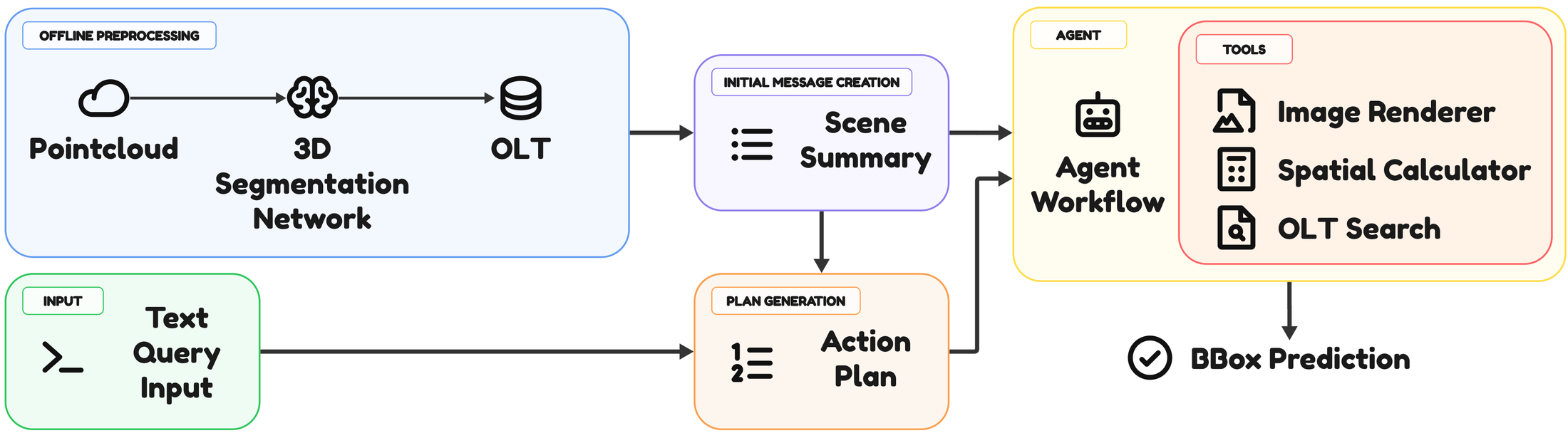

Figure 1. Overview of ***AgentGrounder*** Pipeline

Quantitative Results

Nr3D Benchmark

AgentGrounder achieves an overall accuracy of **52.4%**, yielding a massive **6.3%** improvement over SeeGround. The gains are consistent across difficulty and viewpoint splits, notably a **+7.1%** boost on *Hard* queries.

Table 1. Performance results on *Nr3D* validation set. Queries are labeled as *"Easy"* (only one distractor) or *"Hard"* (with multiple distractors), and as *"View-Dependent"* or *"View-Independent"* based on viewpoint requirements for grounding.

Method	Easy	Hard	Dep.	Indep.	Overall
Supervision: Fully Supervised					
MiKASA	69.7	59.4	65.4	64.0	64.4
ViL3DRel	70.2	57.4	62.0	64.5	64.4
SceneVerse	72.5	57.8	56.9	67.9	64.9
Supervision: Zero-Shot					
ZSVG3D	46.5	31.7	36.8	40.0	39.0
VLM-Grounder	55.2	39.5	45.8	49.4	48.0
SeeGround	54.5	38.3	42.3	48.2	46.1
AgentGrounder (Ours)	**59.6**	**45.4**	**47.9**	**54.5**	**52.4**

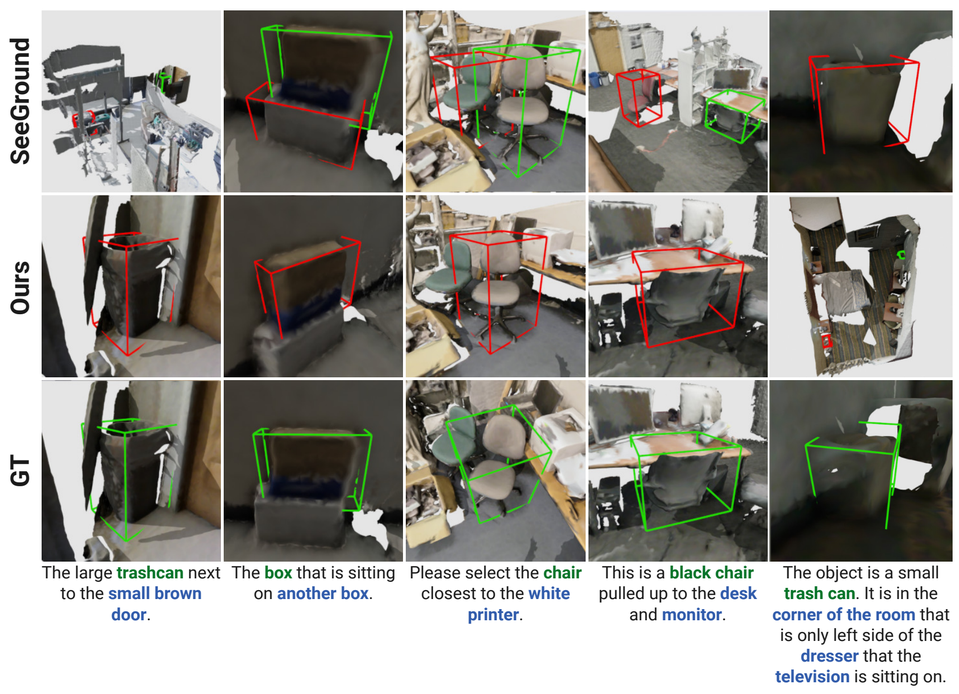

Figure 2. Results comparison between SeeGround and AgentGrounder

Agent Thought Process Example

Given *"Select the long table in the middle of the room"*:

1. **Query Analysis:** Identifies "table", "middle", and "long".
2. **Retrieve Candidates:** Filters tables → matches IDs 2 and 17.
3. **Visual Inspection:** Uses `see_image` to render image of object IDs 2 and 17.

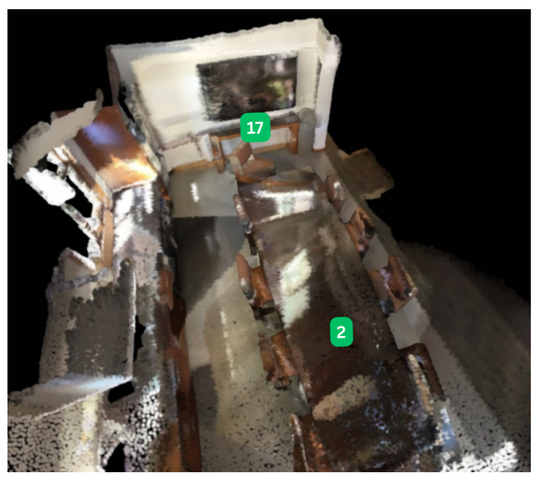

Figure 3. Image rendered by `see_image` tool.

4. **Reasoning:** Computes length via OLT, validates center position via rendered output → Selects ID 2.

Conclusion

We proposed ***AgentGrounder***, a zero-shot tool-augmented LVLM agent that performs 3D visual grounding directly on colored point clouds. By interleaving logic reasoning with deterministic 3D geometric constraints and on-demand spatial rendering, our framework bypasses cascading mismatches. We demonstrated consistent SotA performance on ScanRefer and Nr3D datasets, affirming the generalizability of our framework for complex reasoning in diverse 3D environments.

`https://github.com/be2rlab/AgentGrounder`

Scan the QR code to access the repository.