

From Natural Language to Outdoor Autonomy via Structured Reasoning in Hierarchical Scene Graphs

Harel Biggie¹, Jacob Arkin^{1*}, Jaeyoun Choi^{1*}, Andrew Fishberg^{1*}, Nathan Hughes^{1*}, Travis Manderson^{1*}, Aryan Naveen^{1*}, Mason Peterson^{1*}, Aaron Ray^{1*}, Akila Saravanan^{1*}, Yasmin Veys^{1*}, Ruihan (Multy) Xu^{1*}, Claire Bonial^{2*}, Carlos Nieto-Granda^{2*}, Long Quang^{2*}, Ethan Stump^{2*}, Chuchu Fan¹, Luca Carlone¹, Jonathan How¹, and Nicholas Roy¹

¹Massachusetts Institute of Technology, Cambridge, MA, USA ²Army Research Laboratory, USA

*Authors listed alphabetically by institution, excluding the first and senior authors.

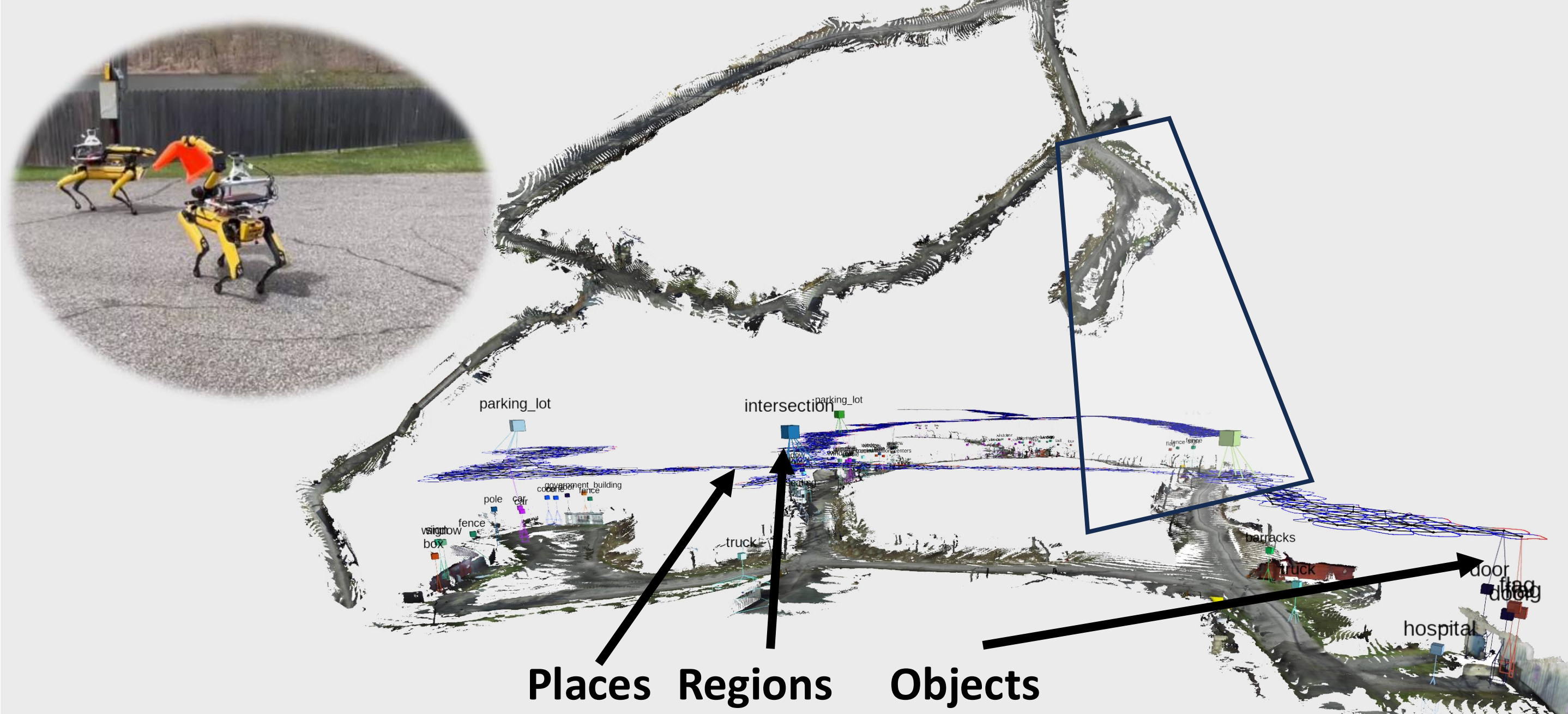


video.adt4.net
Mission Videos



Motivation

Directing robots through natural-language instructions in large, unstructured outdoor environments remains challenging. VLM's operating directly on raw perception struggle in such settings due to the scale and geometric complexity of outdoor scenes.

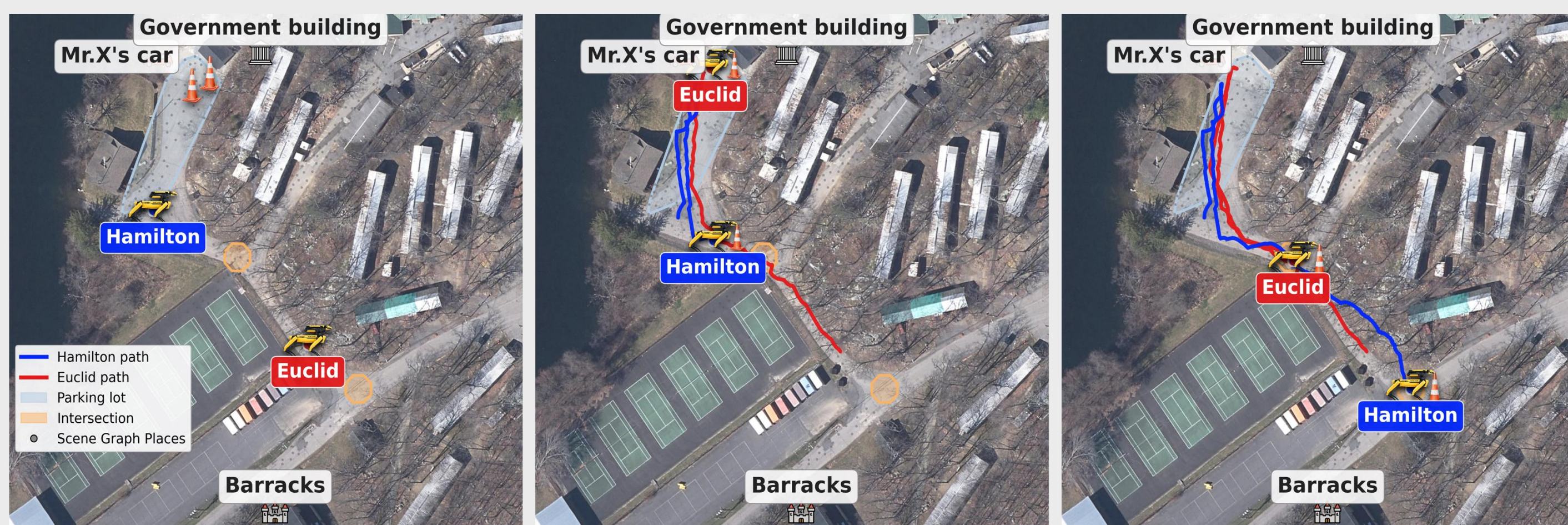


We address this limitation to VLMs by coupling hierarchical 3D scene graphs [1] with a structured language-grounding pipeline based on graph databases [3]. Our approach leverages sparse segment matching [2] for global localization in GPS-denied environments, and we evaluate the resulting system across 14 trials in a camp-like environment in New York state.

Mission Example

Two Spots (Euclid and Hamilton) are given the following language instruction: “Mr. X's car is in the parking lot by the government building. Tell Hamilton to use a cone to block the intersection nearest the barracks. Tell Euclid to use the other cone to block the other intersection.” The language grounding pipeline produces the following goals, and a task and motion planner solves for the plan:

Hamilton's Goal: (object-in-place O102 t4987)
Euclid's Goal: (object-in-place O103 t3917)



Starting Positions

Robots Pickup Cones

Robots Block Intersections

The two robots navigate through the places layer ($t < X$) of the graph (yellow), pick up cones ($O < X$), and drop them off at the intersections.

Language-Grounding Performance

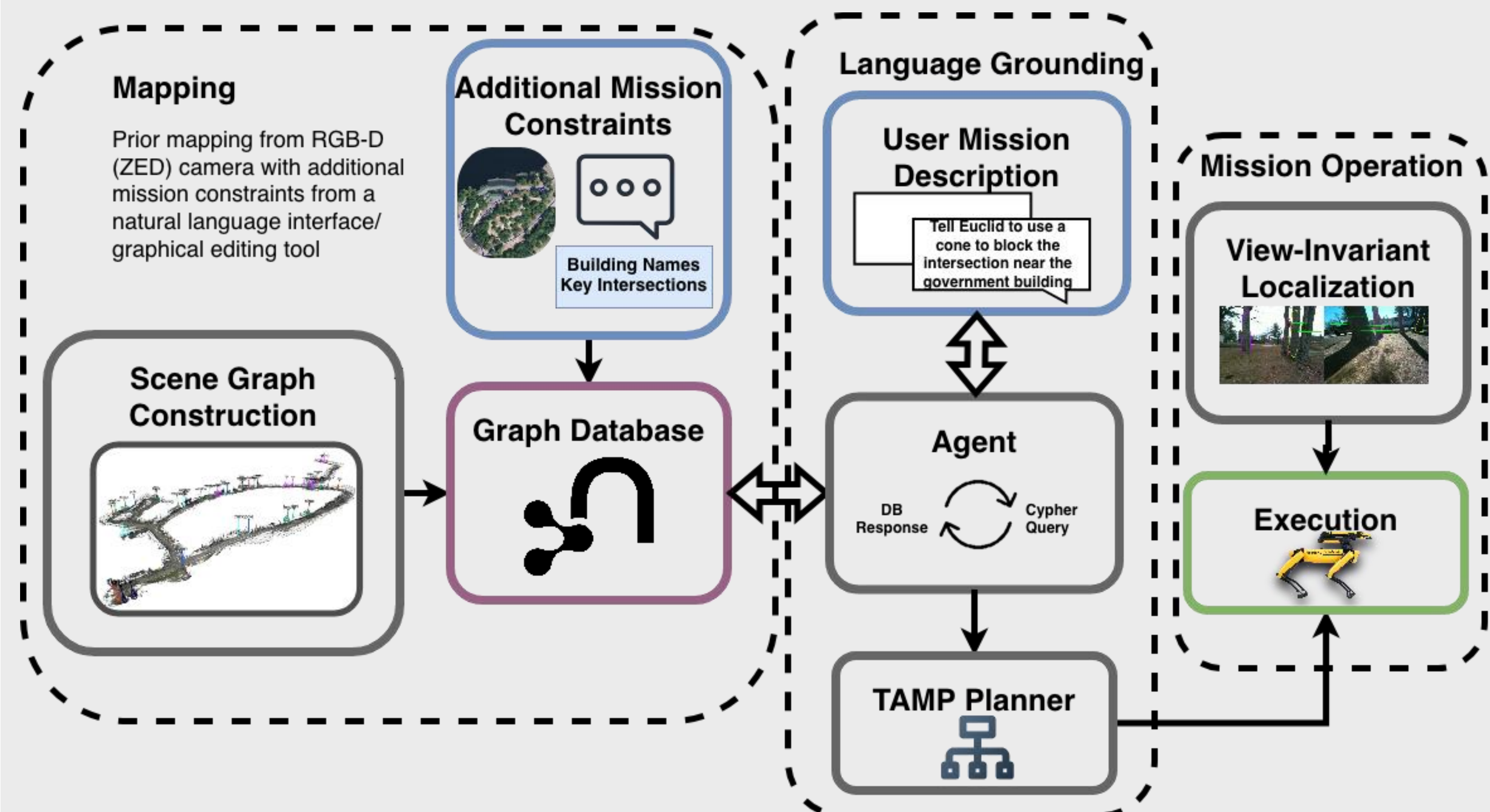


Our Cypher-based database improves language-grounding for both PDDL goal generation and question and answering tasks on a scene graph (15,944 places and 314 objects) constructed on real data from a camp environment across a dataset of 182 examples.

References

- [1] N. Hughes et al. Foundations of Spatial Perception for Robotics: Hierarchical Representations and Real-Time Systems. *The International Journal of Robotics Research*, 2024.
- [2] M. B. Peterson et al. ROMAN: Open-Set Object Map Alignment for Robust View-Invariant Global Localization. In *Robotics: Science and Systems (RSS)*, 2025.
- [3] A. Ray et al. Structured Interfaces for Automated Reasoning with 3D Scene Graphs, 2025. To be presented at ICRA 2026.

Concept Of Operations



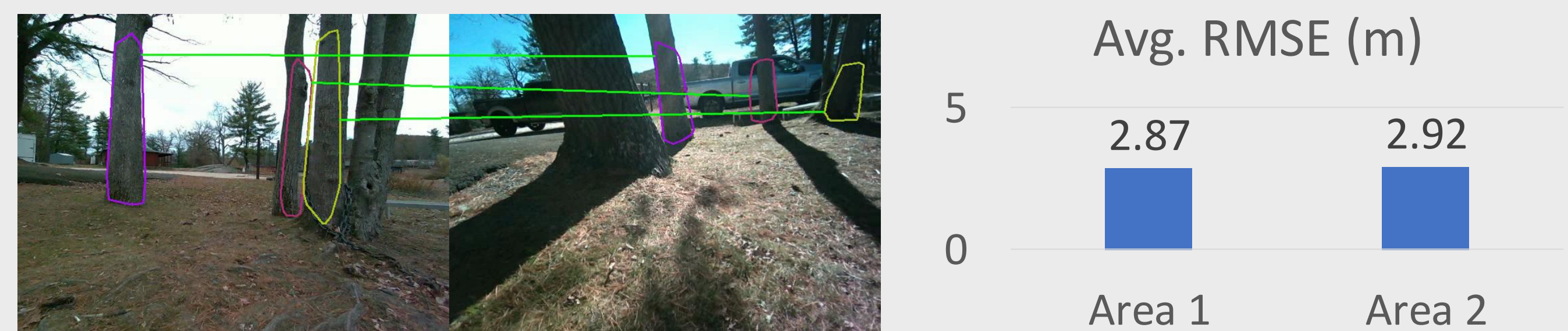
In the mapping phase, a 3D scene graph is constructed from RGB-D sensor data along with sparse submaps for view-invariant localization [2]. The user enters a mission, and the agent grounds the mission to a set of PDDL goals for a TAMP planner. The resulting plan is then executed within the prior map without the need for GPS positioning.

Operational Areas



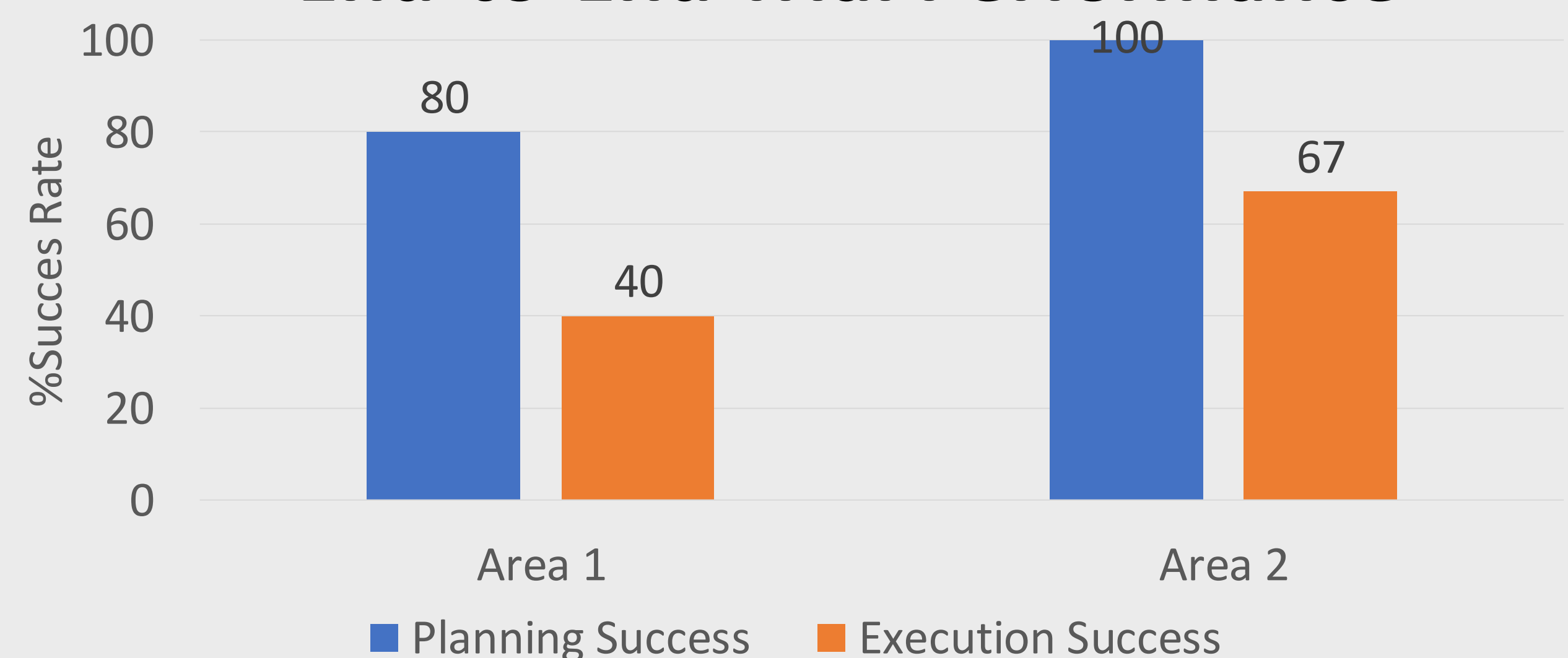
Scene graphs were built in two operational areas. Robot trajectories are drawn using RTK GPS for reference --- GPS was not used during execution.

Localization Performance



Robust view-invariant localization is performed using sparse segment matching [2]. Performance is averaged across 7.6km of travel from experimental trials.

End-to-End Trial Performance



Trial success rate with 5 trials in Area 1 and 9 trials in Area 2. In Area 1, there were 0.4 interventions per trial on average and in Area 2 there were 0.22. Interventions include operations such as teleoperating robots around an undetected obstacle and providing additional clarification about the mission in the chat interface. The end-to-end performance improves on the initial goal generation via clarification interventions.