

Describe Anything Anywhere At Any Moment

Supplementary Material

A. Place Extraction Details

This section further clarifies and illustrates our place extraction approach as introduced in Sec. 3.D). An overview of the process is shown in Fig. A.1.

A volumetric local occupancy map is automatically maintained for 3D reconstruction in the active window of Khronos [54], shown in Fig. A.1a. The occupancy map is convolved with a bounding box of the robot to compute a 2D traversability field, shown in Fig. A.1b. The traversability field is then tessellated by inscribing maximal rectangles such that each rectangle contains only traversable space, shown in Fig. A.1c. To ensure approximately uniform coverage, we further impose a maximum side length constraint (of $2m$ in this example). Each side is further classified into bordering traversable, unknown, or intraversable space. Finally, the places graph is computed as the centroid of each rectangle in Fig. A.1d. Two places are considered connected if adjacent sides of the two rectangles are labeled as traversable boundaries.

For semantic lifting, the obtained place nodes are first projected along the Z axis onto the reconstructed floor to obtain a physical surface point, and then projected into frames observing that point to obtain the features as well as semantic descriptions.

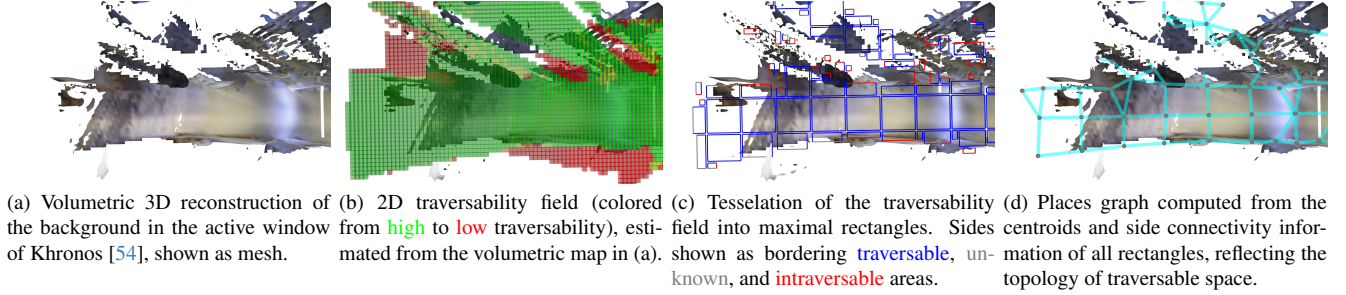


Figure A.1. Illustration of the steps of our traversability place extraction algorithm. The different stages are shown in figures (a)-(d) for a top-down view of the robot moving on a street.

B. Retrieval-augmented Reasoning

In order for an LLM to interface with the 4D scene graph, we use the LLM as a tool-calling agent. The LLM can use the following tools to retrieve information about the scene graph:

- `semantic_search`: Given a natural language of a subject, the tool returns the N ($=10$) most similar fragments based on cosine similarity of the CLIP and sentence embedding features. Returned information includes `description`, `position`, and `observation_timeline` (list of start- and end-observations).
- `fragments_in_radius`: Given a position, the tool returns fragments within a small radius (same information as `semantic_search`).
- `region_information`: The tool returns summaries about the regions in the environment. Information includes the region description, as well as entry and exit positions and times.
- `fragments_in_region`: Given a region ID and a query description, the tool performs `semantic_search` within a region.
- `agent_trajectory`: Given a start and end position, the tool returns N ($=10$) equally spaced poses (position + heading) along the agent trajectory.

C. Example of frame selection quality-score

Fig. C.2 shows a mock-example of the frame selection outlined in Sec. 3-B. Given are three sequential images (1-3) with observed fragments `car`, `tree`, `hydrant`, `person`. Let $\epsilon = 0$. Then, Eq. (1) with $K^* = 1$ will eliminate frame 3 as it does not display all subjects. Thus only a single frame can be selected in Eq. (2). Eq. (2) will then maximize q_i and therefore select frame 2 for all fragments `car`, `tree`, `hydrant`, `person`.

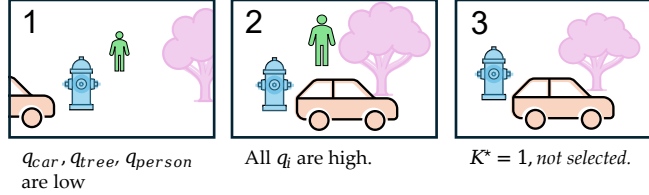


Figure C.2. Mock-example for the frame selection heuristic.

D. Ablation on retrieval tools

We evaluate the benefit of each retrieval tool in Appendix B by holding out the tool when evaluating the method. The results can be seen in Tab. 7. In particular, when running without tools pertaining to regions, positional accuracy drops due to questions referencing region landmarks (e.g., “Where did you see the stairs in the building?”) ($n = 7$) and temporal accuracy drops considerably, as questions like “when did you enter/leave the building?” or “how long have you been outside?” ($n = 15$) become much harder to answer for the LLM-Agent. For retrieval without agent trajectory information, questions like “Was the robot driving on the left side of the sidewalk?” or “When did you turn around for the first time?” ($n = 8$) become impossible to answer. For queries about object relationships (e.g., “Can you find me somewhere to sit near the stairs?”) ($n = 15$), querying objects within a radius becomes useful.

Table 7. Ablation study on retrieval-tools of DAAAM on OC-NaVQA.

	Question Accuracy \uparrow	Positional Error [m] \downarrow	Temporal Error [min] \downarrow
DAAAM + GPT-5-mini	0.711	41.75	1.792
w/o region_information & fragments_in_region	0.707	48.93	3.576
w/o agent_trajectory	0.672	42.67	2.543
w/o fragments_in_radius	0.692	43.67	2.188