

TrajRAG: Retrieving Geometric-Semantic Experience for Zero-Shot Object Navigation

Yiyao Wang^{1,2}, Sixian Zhang^{1,2}, Keming Zhang^{1,2}, Xinhang Song^{1,2,*}, Songjie Du², Shuqiang Jiang^{2,3}

¹State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing

²University of Chinese Academy of Sciences, Beijing, ³Institute of Computing Technology, Chinese Academy of Sciences, Beijing

{yiyao.wang, sixian.zhang, keming.zhang, xinhang.song}@vip1.ict.ac.cn,
 dusongjie25@mails.ucas.ac.cn, sqjiang@ict.ac.cn

Table 1. The selected object categories used in TrajRAG.

Categories	chair, couch, potted plant, bed, dining-table, tv, oven, sink, refrigerator, bottle, cabinet, cushion, chest of drawers, towel, shower, bathtub, counter, fireplace, seating, clothes
------------	---

1. Object Categories in TrajRAG

As shown in Tab. 1, the object categories employed in TrajRAG are derived from a systematic integration and semantic deduplication of 6 target object classes from HM3Dv1 and 21 from MP3D. Specifically, categories with highly similar semantic meanings were merged, while specialized classes prone to frequent misdetection (e.g., picture, mirror) were excluded to form the final set of selected object categories.

2. Ablation on Number of Retrieved Trajectories K

As shown in Fig. 1, we conduct an ablation study on the number of retrieved trajectories K in TrajRAG. When $K = 0$, the LLM based solely on the topo-polar candidate-path representation of the current episode, without accessing any historical trajectories. Increasing K progressively incorporates more scene-specific prior knowledge into the LLM prompts, resulting in consistent improvements in both the Success Rate (SR) and Success Weighted by Path Length (SPL).

The optimal performance is achieved at $K = 32$, beyond which no further gains are observed. For $K < 32$, the systematic improvement with larger K confirms that retrieved trajectories effectively enhance the LLM’s ability to reason in embodied settings, despite its Internet-scale pre-training. For $K > 32$, performance plateaus, which we attribute to two factors: 1) the increasing context length ex-

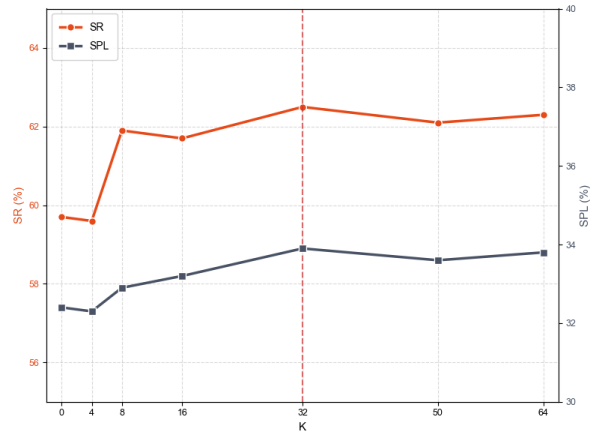


Figure 1. Ablation on the number of retrieved trajectories K in TrajRAG (HM3Dv1). Orange circles denote SR (left axis); dark-blue squares denote SPL (right axis).

ceeding the LLM’s effective reasoning horizon, 2) and the growing influence of object detection errors (discussed in Section 4). Overall, these results validate that trajectory retrieval in TrajRAG contributes systematically to navigation performance.

3. Real-World Robot Evaluation

3.1. Experimental Setup

We validate our approach in a fully furnished 120 m² apartment that contains everyday furniture classes—bed, sofa, TV, coffee table, chair, dining table, potted plant, toilet, etc.—providing a realistic multi-room test field. The hardware platform is a Hello Robot Stretch 3. Its 340 mm wide differential-drive base carries two active wheels and one passive omni-wheel, allowing smooth passage through standard doorways while being unable to negotiate stairs or large elevation changes. An Intel RealSense D435i depth camera, mounted on the mast, delivers 1280 (height)

*Corresponding author.

Table 2. Real-World Robot Evaluation Results

Target	VLFM	TrajRAG
bed	66.67	83.33
nightstand	55.56	72.22
dining table	77.78	77.78
chair	88.89	94.44
refrigerator	38.89	66.67
Avg.	65.56	78.89

× 720 (width) images with a 58° horizontal field of view and serves as the sole exteroceptive sensor.

3.2. Evaluation Results

We evaluate navigation performance using five common household object categories as targets: bed, nightstand, dining table, chair, and refrigerator. For each target category, we select multiple starting positions across different rooms, resulting in a total of 6 distinct starting coordinates. At each starting point, the agent performs 3 navigation episodes, yielding 18 episodes per object category.

The experimental results in Tab. 2 demonstrate that our TrajRAG consistently outperforms the VLFM [1] baseline across most target categories. By leveraging historical navigation experiences, TrajRAG enhances the agent’s capacity to reason about target object locations and the likely room layout. This reasoning enables TrajRAG to attain superior performance.

3.3. Video Demo

We provide a video demonstration of TrajRAG deployed on a real robot, included as `demo.mp4` in the supplementary material.

The clip simultaneously visualizes the onboard RGB stream and the incrementally built semantic map. Consecutive segments are selected to expose the agent’s trajectory-selection logic in real time, corroborating the effectiveness of our approach in physical environments.

4. Failure Cases

We analyze the failure cases of our method and categorize them into three primary types: Infeasible (49.33%), Active Stop (31.21%), and Stepout (19.46%).

Infeasible refers to scenarios where the target does not exist on the agent’s starting floor. Active Stop (autonomous navigation termination) includes False Positive (24.81% of total failures, incorrect target navigation) and No Frontier (6.40%, exhausted frontier exploration). Stepout (step limit exhaustion) comprises Stuck (6.93%, trapped at fixed waypoints) and False Negative (12.53%, resulting from missed object detections, excessively large scene scales, and inefficient path planning).

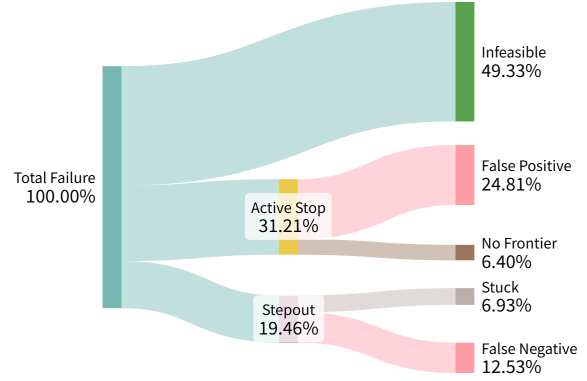


Figure 2. Failure cases of TrajRAG on HM3Dv1. ‘Infeasible’ means the target does not exist on the agent’s starting floor. ‘False Positive’ means the agent navigates to an incorrect target, while ‘No Frontier’ means the target is not found after exploring all frontiers. ‘Stuck’ means the agent navigates to a fixed waypoint and exhausts its steps in a small space. ‘False Negative’ covers other cases where the agent exhausts its steps without finding the target.

Overall, Infeasible is the dominant failure mode, with False Positive and False Negative as key sub-types, highlighting the need to prioritize target existence verification and robust target recognition.

5. Prompts

The prompt for the LLM to perform reasoning in selecting candidate paths is presented below, where ‘{ }’ serve as placeholders that are filled with information from the current episode and the retrieved entries.

```
You are an embodied intelligent robot currently
situated in an indoor environment. Your task
is to find a target object.
Environmental information will be provided as
trajectories that include the path start
point, key waypoints, and geometric & object
observations around each point.

The path you have traveled from the start point
to your present location is:
{history_paths}

The candidate paths you may choose from at this
moment are:
{current_paths}
Your goal is to find the object: {target}

From past experience, the following successful
paths are available for reference when making
your decision.
{knowledge_paths}

Your response must be a single JSON object with
the following structure:
```

```
{ {  
  "reasoning": "",  
  "selected_path": "",  
}
```

References

- [1] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48. IEEE, 2024. [2](#)