SPARK: SPATIAL PERCEPTION AND REASONING BENCHMARK A SCAVENGER HUNT GAME FOR LLM AGENT

Sihan Ren, Siyuan Xie, Jae Won Kim, Zitong Hu, Heekyung Lee, Markus En Cheng Lim

Benchmark Track, 4 Units.

1 EXTENDED ABSTRACT

Vision-and-Language Navigation (VLN) in real-world environments presents a significant challenge in AI research due to the complexity and variability of real-world settings. Existing approaches have largely relied on controlled or synthetic environments, limiting their applicability to realistic scenarios. To address this gap, we introduce a novel benchmark inspired by the Scavenger Hunt game at UC Berkeley, designed to evaluate the capabilities of embodied agents in navigating complex urban street environments. Our benchmark assesses agents on their ability to perceive, understand, and reason about their surroundings, as well as to execute tasks efficiently through multimodal reasoning and decision-making.

The proposed framework is built on a graph-based representation derived from Google Maps, where each vertex corresponds to a navigable location, with five panoramic images offering 180-degree views. This setup enables a detailed evaluation of agents' spatial memory and reasoning skills. We integrated an oracle system that allows agents to engage in free-form dialogue, enabling them to ask strategic questions when faced with uncertainty. The framework measures both the quantity and quality of these inquiries, encouraging agents to adopt deliberate and efficient information retrieval strategies. By fostering critical thinking and incremental information retrieval, our benchmark promotes the development of long-term memory and decision-making capabilities in embodied agents.

In a typical task, the agent receives an initial prompt containing both the target destination and a general direction. This prompt is intentionally obfuscated, resembling a clue in a Scavenger Hunt, requiring the agent to first interpret and break down the task. Based on its understanding, the agent engages in ad-hoc planning to determine its subsequent actions. This process emphasizes the agent's ability to handle ambiguity, perform incremental reasoning, and dynamically adapt its navigation strategy to achieve the specified goal.

To evaluate the performance of the agent, we defined a set of metrics that capture various aspects of its decision-making and reasoning capabilities:

- **Stop Distance:** Measures the agent's decision-making accuracy regarding when to stop. This metric evaluates whether the agent demonstrates overconfidence (hallucinations) or underconfidence in assessing progress toward the goal. Lower values indicate better performance.
- **Efficiency:** Assesses how efficiently the agent completes successful tasks, with higher efficiency reflecting fewer unnecessary actions and optimal task execution.
- **Question Quality:** Evaluates the agent's ability to ask meaningful questions to gather necessary information. Higher scores signify better reasoning and communication skills, highlighting the agent's capacity for strategic inquiry.
- **Action Reasoning:** Measures whether the agent's actions are grounded in available information rather than internal assumptions. High-quality reasoning ensures purposeful actions aligned with task requirements.
- **Success Rate:** Reflects the overall effectiveness of the agent in solving the tasks, capturing its ability to achieve specified goals within the given constraints.

These metrics collectively provide a comprehensive evaluation of the agent's ability to navigate complex environments, reason effectively, and achieve task objectives.

We ran several experiments utilising state of the art multimodal agents like GPT-4o, Claude 3.5 Sonnet, and other open source models, and we found that GPT-4o outperforms the rest of the models. However, overall success rate and efficiency stil lremains low across the board.

We have identified and hypothesized several potential factors contributing to task failure, including circular reasoning, overly broad questioning, repetitive inquiries, and other related issues, which will be discussed in detail below.

Our benchmark advances embodied AI research by challenging agents with realistic tasks that require strategic reasoning, spatial memory, efficient decision-making and multi-modal cross inference capabilities. Experimental results validate the framework's ability to evaluate key capabilities, with metrics offering insights into performance strengths and areas for improvement. By building on real-world scenarios, this work paves the way for developing adaptable, efficient agents for complex real-world applications like urban navigation and assistive robotics.