FindingDory: A Benchmark to Evaluate Memory in Embodied Agents

Karmesh Yadav*[1], Yusuf Ali*[1], Gunshi Gupta[2], Yarin Gal[2], and Zsolt Kira[1]

[1] Georgia Tech, Atlanta, GA, USA
[2] University of Oxford, Oxford, UK
https://findingdory-benchmark.github.io

Abstract. Rapid progress in embodied AI is bringing robots closer to performing complex tasks in real-world environments. In these settings, robots must operate over long time horizons, making decisions based on experiences collected over hours or even days. Vision-language models (VLMs) have recently shown strong capabilities in planning and control, making them promising high-level controllers for embodied agents. However, current VLMs can process only a limited number of images at once, highlighting the need for more efficient mechanisms to manage long-term memory in embodied contexts. To meaningfully evaluate these models for long-horizon control, a benchmark must target scenarios where memory is essential. Existing long-video QA benchmarks neglect embodied challenges like object manipulation and navigation, which require low-level skills and fine-grained reasoning over past interactions. Moreover, effective memory integration in embodied agents involves both recalling relevant historical information and executing actions based on that information, making it essential to study these aspects together. In this work, we introduce FindingDory, a new benchmark for long-range embodied tasks in the Habitat simulator. FindingDory evaluates memory-centric capabilities across 60 tasks requiring sustained engagement and contextual awareness in an environment. The tasks can also be procedurally extended to longer and more challenging versions, enabling scalable evaluation of memory and reasoning. We further present baselines that integrate state-of-the-art closed-source and fine-tuned open-source VLMs with low-level navigation policies, assessing their performance on these memory-intensive tasks and highlighting key areas for improvement.

1 Introduction

Memory is a fundamental capability for intelligent agents that allows them to recall past experiences, adapt to dynamic environments, and make informed decisions over extended timescales. In humans and animals, it plays a crucial role in navigation, reasoning, and goal-directed behavior. As we strive to develop more capable embodied systems, equipping them with robust memory mechanisms is essential, particularly in settings where agents must process high-dimensional

* Equal contribution

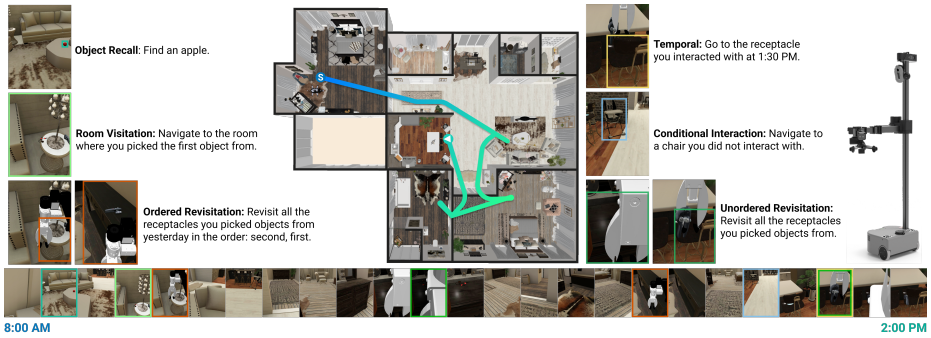


Fig. 1: FindingDory: We propose a benchmark for evaluating agents' memory by constructing scenarios that require reasoning over previously collected experiences. In these scenarios, an agent receives logs of prior interactions in indoor environments and must complete navigation, pick, and place instructions. The figure shows a sample episode: first, a privileged agent with full environment access executes randomly sampled mobile-manipulation tasks. Then, the memory-evaluating agent is placed in the same environment and tasked with completing instructions that rely on the privileged agent's logs to decide where to navigate or which object to retrieve.

multimodal inputs to interact with their surroundings. The ability to store and retrieve relevant information is the key to unlocking sophisticated behaviors, from household assistance to autonomous exploration.

Recent advances in Vision-Language Models (VLMs) have significantly improved high-level reasoning and planning for embodied tasks. These models leverage large-scale multimodal training to exhibit strong scene understanding and action-guided perception. However, most existing VLM applications focus on short-term or static tasks, such as image captioning or Visual Question Answering (VQA), which do not require sustained memory usage. Extending these models to long-horizon embodied control introduces new challenges: embodied agents must integrate observations over time, recall past experiences, and act accordingly within their environment.

While recent efforts have sought to extend long-context understanding in VLMs, primarily through video-based QA or long-document comprehension [25, 41, 50], these approaches fail to fully capture the complexities of long-term memory in decision-making. Many existing QA benchmarks either rely on multiple-choice formats, which enable guessing, or require extensive human annotation to curate question-answer pairs that require models to attend to long video sequences. In contrast, we propose embodied memory tasks such as *"navigating to a soft toy that you did not rearrange yesterday"* or *"find the second object picked up earlier"* that demand fine-grained recall and multi-hop reasoning over past observations. These tasks provide a more rigorous test of memory capabilities, requiring agents to track and respond to environmental changes rather than passively perceive static information.

In this paper, we introduce FindingDory, a benchmark designed to evaluate long-horizon memory in highly photorealistic simulated environments. Our benchmark consists of **60 diverse embodied tasks**, which we categorize along

various axes of memory requirements, requiring both long-range temporal and spatial reasoning while discouraging heuristic shortcuts or guessing. To ensure that performance is directly tied to memory utilization rather than simple perception, we curate scenarios where agents must recall past interactions to succeed. Additionally, our benchmark features dynamic environments - where mobile manipulator agents modify the scene - making it essential to reason over changing contexts. Furthermore, FindingDory is procedurally extensible, allowing tasks to scale in complexity, ensuring continued relevance as embodied AI progresses.

Unlike standard QA datasets that rely on subjective human annotations, our benchmark leverages photorealistic simulation to enable automatic evaluation of memory-based navigation. These tasks go beyond static recall since they require agents to perform complex spatio-temporal reasoning over past interactions. For instance, tasks such as *"navigate to any object that you interacted with yesterday"* demand not only accurate memory retrieval, but also strategic decision-making to identify and navigate to the *nearest* valid target from where the agent is currently positioned. In FindingDory, we introduce metrics that quantify task completion efficiency and not just successful recall.

Our key contributions are:

1. A benchmark for evaluating long-horizon memory in embodied decision-making, featuring 60 diverse embodied tasks in highly realistic indoor environments.
2. A comprehensive evaluation of VLM based high-level policies combined with low-level navigation policies, analyzing their performance and limitations in memory-intensive tasks.
3. A systematic and extensible evaluation framework with metrics that enable future advancements in memory-efficient embodied agents.

2 Related Work

In this section, we review related work on long-video question answering and embodied AI benchmarks that incorporate historical experience in their problem formulation. We defer the discussion of different approaches for augmenting policies with memory-like mechanisms to Appendix C.1.

2.1 Video QA benchmarks

Recent analyses of video QA benchmarks [6, 28] show that many tasks can be solved by attending to a few key frames, not full videos, suggesting that current benchmarks may favor keyframe selection over evaluating long-term memory.

Long-context evaluations for foundation models also focus on "needle-in-a-haystack" tasks, evaluating retrieval of specific details rather than integrating information across many segments [59]. While recent video benchmarks feature more challenging questions [11, 22, 23, 26, 28, 34, 40, 41, 46, 53, 60], they adopt

Table 1: *FindingDory vs.* other memory-focused embodied benchmarks.

Benchmark	Photorealism	Semantic Reasoning	Task Specification	Template Categories	# Resulting Tasks	Memory Length (π^*)	Isolates Memory
MemoryMaze [31]	✗	✗	One-hot	✗	1	500-1000 steps	✗
MemoryGym [32]	✗	✗	One-hot	✗	3	128-512 steps (Extensible)	✗
MultiON [47]	✓	✗	One-hot	✗	1	500 steps	✗
SIF [29]	✓	✓	Language	3	240 (`val/test`)	✗ (Uses semantic map)	✗
OpenEQA (Active) [26]	✓	✓	Language	✗	1600	75 steps	✗
GoatBench [20]	✓	✓	Lang + Image	✗	1800 − 3600 (`val`)	500 steps (Static Env)	✗
Excalibur [62]	✓	✓	Language	11	21	Requires exploration	✗
Ours	✓	✓	Language	11	≈ 6000 (`val`)	500-3500 steps (Extensible)	✓

multiple-choice QA formats which are not suitable for evaluating fine-grained decision-making over long histories. They may also allow models to succeed through guessing and are not extensible since they require substantial human annotation effort.

2.2 Embodied AI benchmarks

Various benchmarks have been proposed to evaluate vision-language models in embodied settings where agents must leverage their experience to improve future decision-making [47]. Many of these benchmarks focus on embodied question answering (EQA) [9]. For example, HM-EQA [35] requires agents to actively explore environments to gather information needed to answer questions. In contrast, OpenEQA [26] evaluates models both in an active exploration setting and in an episodic-memory setting, where questions must be answered using pre-collected observation histories. Notably, Open-EQA found that a *blind* VLM baseline (no visual input) still performed well, suggesting that multiple-choice evaluation can introduce biases and reduce the need for genuine memory recall.

Other benchmarks study the situated instruction following setting where agents must gather information to interpret ambiguous instructions [24, 29, 36]. However, many allow agents to sidestep memory challenges by precomputing relevant objects and locations based on recent history, effectively turning potentially memory-intensive tasks into structured instruction-following problems. Memory-Maze and Memory-Gym [31, 32] focus on partially observable decision-making with history-based state modeling, but their simplistic visuals and small state spaces limit their relevance to complex real-world memory settings.

In Tab. 1, we compare FindingDory to the most related benchmarks within embodied AI and RL with respect to the following desired properties:

1. **Photorealism**: Measures the visual complexity of environments, affecting the meaningful evaluation of foundation models trained on real-world data.
2. **Scope of reasoning needed over memory**: Distinguishes between benchmarks that require recalling diverse semantic information v.s. a narrow set of attributes (e.g. object locations).
3. **Task Specification**: Evaluates whether tasks are defined using flexible language-based descriptions, enabling richer memory challenges.
4. **Memory horizon length**: Adapting the definition from [30], this measures the temporal scope over which memory must be maintained for optimal

task performance. Unlike static benchmarks requiring localized frame recall, dynamic environments demand reasoning over events distributed across extended observation sequences. This metric becomes difficult to quantify when tasks require substantial exploration phases.

5. **Memory isolation**: Whether benchmark performance isolates memory evaluation from confounding factors, particularly exploration requirements, which represent orthogonal challenges to memory recall and reasoning.

Our benchmark isolates memory evaluation by avoiding confounding related to exploration and focuses on careful and extensible task curation with automated metrics, requiring agents to recall and reason over long-horizon environmental observations.

3 FindingDory

In this section, we introduce our embodied memory benchmark, FindingDory, built on top of the **Habitat simulator** [37]. Habitat offers photorealistic visuals, diverse household scenes, and object-rich environments, making it a practical and transferable testbed for evaluating memory-augmented embodied agents. Since many vision-language models (VLMs) are pretrained on real-world image distributions, they can directly interpret observations from Habitat, enabling realistic and scalable evaluation of long-term decision-making. We first describe the overall structure of FindingDory and the suite of tasks it comprises, then provide implementation details.

3.1 Benchmark Design and Tasks

The goal of FindingDory is to evaluate an agent's ability to use prior mobile-manipulation history to complete future tasks efficiently. To achieve this, we design a two-phase setup: an *experience collection phase*, where a scripted oracle agent generates object rearrangement episodes through pick-and-place interactions, and an *evaluation phase*, where the evaluated agent must use the resulting egocentric observation history to recall and act on remembered object-receptacle states. This decomposition isolates memory from exploration, enabling standardized evaluation.

During the **experience collection phase**, an oracle agent picks up and places objects onto designated receptacles (e.g., a table, chair, or bed), introducing meaningful state changes into the environment. To ensure that these logs provide reliable supervision for downstream memory evaluation, we generate a large pool of candidate episodes, prevalidate object-receptacle placements with physics checks, and discard episodes with unstable placements or failed oracle trajectories (see Appendix B.1). Episodes span 400–3500 frames with 2–11 pick-and-place interactions (see Appendix B.2). Once collection ends, the **evaluation phase** begins: the agent is given access to its recorded history (images, pose, past actions) and tasked with following instructions that require integrating long-term memory into decision-making.

Tasks are instantiated from a set of **templated instructions** designed to probe diverse memory challenges. These cover **spatial**, **temporal**, and **semantic** reasoning: e.g., recalling which object was moved, when it was moved, or how it was interacted with; identifying rooms visited or farthest locations; and reasoning over attributes such as shape, color, or material. We also introduce **time-usage memory** tasks (e.g., longest room exploration, object with longest manipulation). A smaller subset of **attribute-based tasks** involves referring to objects indirectly via observed properties (e.g., "go to the striped object you interacted with previously"), helping distinguish methods that can flexibly reconstruct fine-grained visual details from those relying on static summaries. In total, 60 templates are used to generate scene-specific task instances (see Tab. 5). We further group tasks into interpretable categories for analysis in Tab. 2.

Single vs. Multiple Goals. Most tasks involve single-goal instructions, but some require reaching multiple destinations in sequence. Even single-goal tasks can be challenging, demanding agents recall what was *not* interacted with or trace object movements across rearrangement states.

Extensibility. The benchmark is procedurally extensible. Increasing the complexity of the experience collection phase (e.g., the number of interactions or episode length) naturally expands the temporal reasoning horizon and memory dependencies required for successful task completion. Additionally, noise can be introduced during experience collection to simulate imperfect observations or actions, further increasing task difficulty (see Appendix F for details).

3.2 Implementation details

Simulation Setup and Dataset Composition. We build FindingDory using the Habitat 2.0 simulator [37, 42] and scenes from the Habitat Synthetic Scenes Dataset (HSSD) [19]. Our benchmark uses 107 scenes from the training split and 30 from the validation split, with 1,478 training and 100 validation episodes. We populate these scenes with the same object sets used in OVMM [56]. Across these episodes, we include 839/247 object instances spanning 84/72 distinct categories in the `train/val` splits, respectively. Objects are rearranged during the experience collection phase to induce temporally grounded memory challenges. Additionally, agents interact with 17/16 distinct receptacle categories in the `train/val` splits (e.g., tables, chairs, beds). To support attribute-based tasks (see Tab. 2), we generate semantic descriptions for object instances using GPT-4o. Specifically, we prompt the model with multi-view images of each object and extract attributes such as shape, color, material, and functionality. The generated descriptions are manually reviewed for quality and then used to instantiate the five object-attribute task templates. We evaluate these semantic-attribute tasks only on the validation split, providing a held-out test of whether agents can generalize beyond templates available for training. In total, our benchmark includes 79,213 training and 5,876 validation task instances. More details in Appendix B.

Agent Setup. We adopt the Hello Robot Stretch embodiment [18] in our simulation. The agent is equipped with RGB and depth sensors (resolution

Table 2: Task categories with example instructions and associated challenges. See Appendix I for more examples.

	Task Category	Example Instructions	Memory Requirement
Spatial	Object Recall	Navigate to a *{category}*.	Object was previously seen or needs exploration.
	Interaction	Navigate to any object that you did not interact with yesterday.	Differentiate between interacted & uninteracted objects.
	Conditional Interaction	Navigate to a *{receptacle_category}* you picked an object from.	Object categories, object-receptacle relationships.
	Object Attribute	Navigate back to a *{target_color}* colored object that you interacted with yesterday.	Object attributes other than category.
	Spatial Relationship	Navigate to the object which you interacted with which is the farthest from your current location.	Spatial relationships to objects.
	Room Visitation	Navigate to a room that you did not visit yesterday.	Room boundaries and past interactions.
Temporal	Interaction Order	Navigate to the object you interacted with immediately after *{object_category}*.	Sequence of objects interactions.
	Time-Based	Navigate to the object that you interacted with at *{HH:MM}* yesterday.	Time of interaction & final position of objects.
	Duration Tracking	Navigate to the object which took the longest time to rearrange.	Duration of interactions.
Multi-Goal	Unordered Revisitation	Revisit all the receptacles you picked objects from yesterday.	Past locations of multiple objects and best visitation order.
	Ordered Revisitation	Revisit all the objects you interacted with yesterday in specific order.	Similar to unordered revisitation but with strict sequencing.

640 × 480) and a GPS+Compass sensor. The agent can execute one of four discrete navigation actions: `MOVE_FORWARD` (0.25m), `TURN_LEFT` / `TURN_RIGHT` (by 10°) and a `STOP` action. During the experience collection phase, the agent can also manipulate its arm joints (extension, lift, yaw, pitch, roll) and activate a `manipulation_mode` for pick-and-place actions, which rotates the base (90°) and directs the head camera toward the gripper.

Procedural Generation and Verification using PDDL. We operate over a large set of diverse task instructions (see Tab. 2) by encoding them into templates using the PDDL specification system [2, 43]. Specifically, each template specifies *entities* (objects and receptacles) with *properties* such as interaction order, target color, and start/goal receptacle. These entities are combined using *predicates* that map to pythonic functions which can be evaluated against any simulator state. *Goal conditions* compose these predicates with *quantifiers* (e.g., `EXISTS`, `FOR_ALL_UNORDERED`) to express complex logical relationships between multiple predicates. The defined instructions templates are *procedurally* expanded by sampling valid entity bindings based on the entity properties which serve as constraints. The bindings are finally substituted into the goal conditions to generate the FindingDory dataset and enable scalable, automated verification for each episode without hand-coding low-level rules.

4 Experiments

We evaluate several baseline approaches on FindingDory and assess their ability to complete memory-intensive tasks using a combination of vision-language reasoning and navigation. Below, we describe the baseline architecture, implementation details, and evaluation metrics.

4.1 Evaluated Approaches

To establish strong baselines, we adopt a hierarchical architecture (App. Fig. 16) comprising two modules: (1) a high-level goal selection module, which processes the interaction history to select a sequence of goal frames, and (2) a low-level navigation policy, which executes actions to reach the selected frame. This architecture enables reasoning over long-horizon multimodal experience while delegating fine-grained control to a learned policy. Inspired by prior work [51], we avoid directly predicting actions using VLMs due to their poor performance in continuous control, and instead restrict them to high-level goal prediction.

High-Level Goal Selection. High-level goal selection requires grounding the instruction in the full interaction history before choosing a goal frame. The agent must often infer intermediate context such as which object was picked, where it was placed, which receptacle it was moved from, what happened before or after another event, and which valid target is closest to the agent's current location. Thus, our tasks require temporal, conditional, spatial, and multi-event reasoning over history, rather than simple visual matching. We evaluate three approaches: a Video VLM agent that directly reasons over interaction videos, Structured Memory agents that summarize video chunks into textual or multi-view images before reasoning, and a Supervised Fine-Tuned model trained to directly predict goal frames from interaction history.

Video LM Agent. This agent receives the full interaction video along with a task instruction. Each frame in the video is annotated with its index and the time of day at which it was captured. The agent is tasked with predicting the index of a frame from the video that corresponds to a viable goal state. Specifically, the predicted frame should represent a viewpoint from which, if the agent were to navigate back to the same camera position, it could successfully complete the task. Although the model outputs a single frame index, selecting the correct one often requires reasoning over multiple temporally scattered observations, including frames where the agent may not be actively interacting. Since different VLMs are capable of handling different numbers of frames in context, we subsample the frames passed to each model based on what empirically performs best. We evaluate both proprietary (Gemini-2.0-Flash [44], GPT-4o) and open-source (Qwen2.5-VL [4], Gemma-3 [45], GLM-4.1V-Thinking [14]) vision-language models capable of processing long video contexts. See Appendices D.1 to D.3 for more details.

Structured Memory Agent. Similar to [3, 29], we consider an explicit, memory-building baseline which breaks down the video frames into chunks and

uses a VLM to generate relevant textual summaries for each video chunk. After this guided "captioning" of the entire video sequence, an LLM selects goal frame indices by reasoning over the generated textual descriptions. In addition, we also evaluate 3D-Mem [54] that incrementally builds an efficient, structured 3D representation of the environment using informative snapshot images (instead of text) to store salient information. See Appendix D.4 and Appendix D.5 respectively for details on implementation.

Supervised Fine-Tuning (SFT). We additionally explore a supervised fine-tuning approach where the high-level VLM is trained to predict goal frame indices, given the task prompt and interaction history. Since multiple frames may validly represent the goal (e.g., several consecutive frames showing an object being placed), we supervise the model on the full list of acceptable target frames. At test time, a single frame is sampled from the list of predictions and is considered correct if it matches any of these valid frames. This setup enables the model to learn task-specific cues from examples and serves as a stronger baseline than zero-shot prompting, particularly for tasks requiring subtle temporal grounding. We provide additional details in Appendix D.6.

Low-Level Navigation. To execute the high-level plan, we use a low-level policy that takes the selected goal frame and outputs discrete actions: `MOVE-FORWARD`, `TURN-RIGHT`, `TURN-LEFT`, and `STOP`. Our primary policy is the LSTM-based model from OVRL-v2 [52], trained to reach image goals from egocentric RGB-D observations (see Appendix E.1). We also evaluate a mapping-based policy that deterministically navigates to the location corresponding to the goal frame using a global map, without relying on learned visual features (see Appendix E.2).

Solvability. The action space of all our high-level baselines is constrained to frames observed during experience collection, which is insufficient for some tasks. For example, in the task *"Find an apple that you did not interact with yesterday"*, if the agent never closely approaches that apple during experience collection, then no frame in the history can serve as a goal frame. This reflects a limitation of the baseline's goal-parameterization rather than a limitation of the benchmark itself. To establish an upper bound on task difficulty for current baselines, we quantify the percentage of episodes that could theoretically be solved through optimal frame selection from interaction history, assuming perfect navigation (teleportation) to the selected waypoint. This upper bound is the dashed line (`Oracle`) in Fig. 2.

4.2 Evaluation Metrics

We benchmark agent performance on the FindingDory tasks using **Low Level Success Rate** (`LL-SR`), assessing whether the agent finds the right target entities, and **Low Level Success-weighted-by-Path-Length** (`LL-SPL`), evaluating if the agent selects and navigates to the optimal entity using the shortest possible path. Since we use a hierarchical baseline, we additionally introduce a **High-Level Policy Success Rate** (`HL-SR`), and **High-Level Policy SPL** (`HL-SPL`) which measures the accuracy and efficiency of the high-level policy in

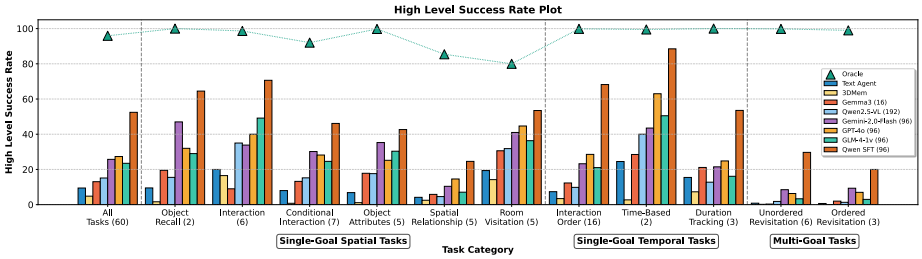


Fig. 2: FindingDory High-Level Success Rate Plot: We evaluate variety of baselines on the high-level task of FindingDory. The numbers in parentheses for each VLM denote the optimal frame budget used the model. Proprietary VLMs such as `Gemini-2.0-flash` and `GPT-4o` show little success across most FindingDory task categories. Agents struggle especially on multi-goal tasks where multiple subgoals must be achieved sequentially from the interaction videos collected during Phase 1. The `Qwen-SFT` baseline (see Appendix D.6), trained to predict ground-truth frame indices, performs best but still reaches only $\approx 50\%$ on high-level tasks.

selecting the correct and closest goal frames for navigation. To further analyze failure modes, we also report two relaxed variants: **Distance-to-Goal-Only SR** (`DTG-SR`), which relaxes the requirement that the target object must be visible in the selected frame, and **Semantic Coverage SR** (`SC-SR`), which ignores spatial proximity. Additional metrics, success criteria, and threshold specifics are provided in Appendix B.3. For the 3D-Mem baseline, we report evaluations on 95 validation episodes on which the memory-snapshot construction pipeline completed within the max allocated time of 6 hrs per episode.

5 Results

In this section, we analyze the performance of the previously outlined baselines on FindingDory tasks, presenting results grouped according to the task categories described in Sec. 3.1.

State-of-the-Art VLMs struggle on the FindingDory benchmark. We first evaluate different VLMs on the high-level task and report success rates in Fig. 2, observing consistently low performance across all task subsets (see Tab. 4 for numbers). GPT-4o achieves the highest success at 27.3%, followed by Gemini-2.0-Flash at 25.7%, and GLM-4.1V-Thinking at 23.5%. The non-reasoning open-source models Qwen2.5-VL and Gemma-3 score between 13% and 16%. The text-based agent performed worse at 9.45%, underscoring that textual memory representations alone are insufficient for disambiguating visual goals or reasoning over complex visual histories. The 3D-Mem [54] baseline performed similarly with an overall `HL-SR` of 6.9% (see Appendix D.5 for detailed discussion). Notably, the supervised fine-tuned (SFT) variant of Qwen outperforms all frozen VLMs, achieving an average improvement of 25% across tasks, with considerable improvement on temporal and multi-goal tasks. A more detailed breakdown of the gap between the zero-shot and SFT version of Qwen by task category and interaction count is provided in Appendix H.

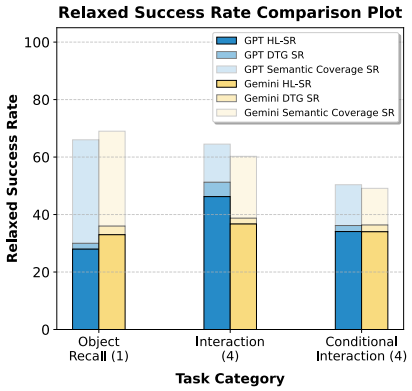
In single-goal tasks, all methods struggle particularly with the spatial relationship subset, with GPT-4o achieving only 14.5% success. These tasks require reasoning about relative distances between entities and the agent, an ongoing challenge noted in prior work [53]. In contrast, room identification tasks yield the highest performance, likely due to the small number of distinct rooms and the lenient success criteria of reaching any location within the correct room. Performance on conditional interaction and object attribute tasks is comparable, suggesting that current VLMs handle both category-based queries (e.g., "apple") and attribute-based descriptions (e.g., "red food") with similar effectiveness.

Temporal tasks pose additional challenges to the existing VLMs, with models consistently failing to identify the correct order of interactions, since it requires multi-hop reasoning. Duration tracking is similarly difficult as it requires inferring elapsed duration based on specific points in the video. In contrast, for the time-based tasks, we observe that baselines perform better, as these tasks only require single frame retrieval based on the timestamps in the question, without additional reasoning.

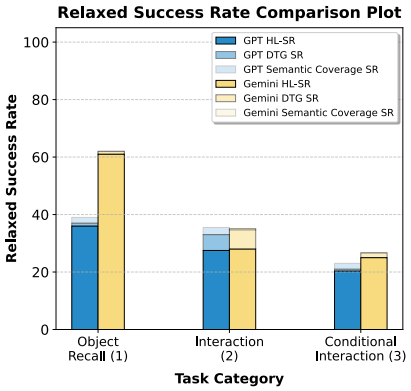
Poor performance on multi-goal tasks. From Fig. 2, we observe that all VLM baselines achieve near-zero performance on most multi-goal tasks. This indicates that VLMs struggle to track and recall multiple object–receptacle interactions. In our multi-goal evaluation, we employ a single-shot prompt, requiring the high-level VLM agent to predict all target frame indices in one response, primarily to avoid repeated inferences over long sequences given the large number of frames. Qualitatively, we find that VLMs often fail to predict even the correct number of targets needed to solve a task. While our trained Qwen model performs somewhat better, reaching around 20% accuracy, it still falls far short of the attainable 99% success rate on this task.

VLMs can identify target receptacles but not precisely localize them. We analyze VLM performance on tasks where the valid entities are receptacles versus where they are objects. As shown in Fig. 3a, the `Semantic Coverage-SR (SC-SR)` metric (see Appendix B.3) is substantially higher for receptacle-based tasks. This indicates that VLMs often select frames where the receptacle is visible, but the agent remains far from it. In contrast, distance-threshold success (`DTG-SR`) improves negligibly, suggesting that very few cases exist where the agent is close to a receptacle but not facing it. For smaller objects, no such improvement is observed (see Fig. 3b), since small objects are rarely visible from longer distances. Taken together, these results highlight a key limitation of current VLMs: they can semantically detect large targets but fail to construct spatially grounded representations necessary for precise localization.

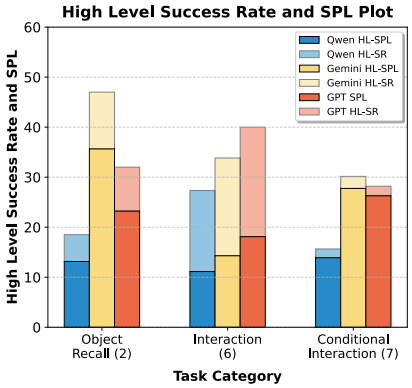
VLMs struggle with identifying the closest entity that solves the task. In Fig. 3c we report the `HL-SPL` metric (Sec. 4.2) which captures the efficiency with which the high level VLM-based modules select goal frames. During the experience collection phase, the agent can interact with multiple object instances of the same category due to which a task such as *"Navigate to a squeezer you interacted with"* could have multiple valid entities that the agent can navigate to. We observe that for such tasks there is a wide gap between `HL-SR` and



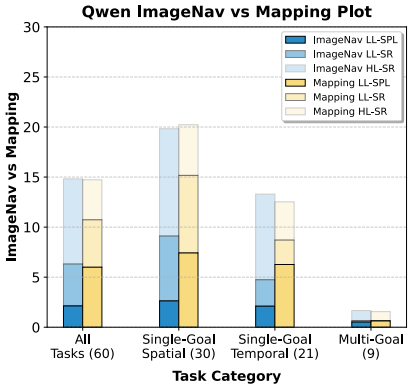
(a) Metric: Receptacle tasks.



(b) Metric: Object tasks.



(c) HL-SR vs HL-SPL.



(d) Hierarchical policy.

Fig. 3: (a, b) Relaxed metrics for object and receptacle tasks. Models show substantial gains on receptacle-based tasks under relaxed metrics, whereas tasks targeting small objects exhibit little improvement. **(c) Success Rate vs. SPL.** For tasks where the VLM may select among multiple correct frames, all models sometimes choose suboptimal ones, reducing the `HL-SPL` metric. **(d) Hierarchical policy on FindingDory.** When Qwen is paired with an ImageNav or Mapping agent, overall SR and SPL drop markedly (as reflected in `LL-SR/LL-SPL`).

`HL-SPL` (upto 50% in case of interaction-based tasks) for all model variants. This suggests that although the VLMs can correctly recognize a valid entity to solve the task, they still fail to perform a fine-grained spatial analysis of the multiple valid entities being rearranged dynamically in the video. Thus, they cannot spatially locate the valid entity nearest to its current position.

Policy decomposition causes performance degradation. To evaluate the hierarchical agent introduced in Sec. 4.1, we use `Qwen2.5-VL-7B` as the high-level VLM module and pair it with either an ImageNav policy or a mapping-based policy as the low-level agent. The low-level policy is triggered only when the high-level VLM correctly identifies a subgoal. As shown in Fig. 3d, `LL-SR` drops sharply for the ImageNav policy. The same policy achieves 78% success on its native ImageNav validation episodes (see Appendix E.1), confirming it per-

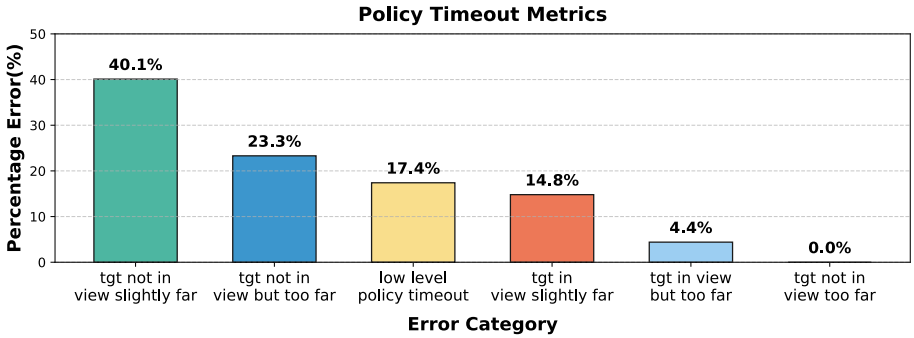


Fig. 4: Failure modes of the ImageNav policy.

forms well when goal images match its training distribution. The degradation arises from distribution shift: VLM-selected goals often come from pick–place interactions with atypical viewpoints and limited visual cues, unlike the goals used during ImageNav training. By contrast, the mapping-based navigation policy is more robust, suffering only a 25% relative drop in SR. However, the SPL of both methods remains in the single digits, underscoring the importance of FindingDory in exposing these shortcomings and motivating the development of more reliable memory and navigation strategies.

Low level navigation policy failures. From Figure 4, we observe that the low-level image-goal navigation policy can fail due to invoking the termination (stop) action incorrectly or a navigation routine timeout (max budget of 1000 steps including the data collection phase). To further understand this behavior, we analyze if the navigation policy actually looks at the target entity when the stop action is invoked.

Interestingly, we find that in several instances, the target entity is visible from a slightly larger distance threshold (2.0 m) in comparison to the ImageNav policy success threshold during training (which is 1.0 m for all supplied goal frames; see Appendix E.1 for details).

This often happens when VLMs select goal frames in which the target entity is visible in the 2D image but not necessarily closest in 3D space, and thus the goal frame coordinates are slightly further away than the threshold distance.

Leveraging additional frames. We study the effect of input length on high-level performance by evaluating agents on subsampled videos of varying sizes (Fig. 5). This allows us to test whether models benefit from longer histories or merely pick up sparse visual cues. Interestingly, frozen VLMs, although capable of processing long contexts, show little to no improvement with more frames. Their performance often degrades at higher frame counts, suggesting difficulty in extracting relevant signals from densely packed, unfiltered inputs.

In contrast, the fine-tuned model (Qwen SFT) achieves clear gains when trained with longer videos. This shows that with appropriate supervision, models can move beyond shallow matching to exploit richer temporal structure in interaction histories. Nonetheless, even with full finetuning on longer (*subsampled*) videos, a large gap remains relative to maximum achievable performance

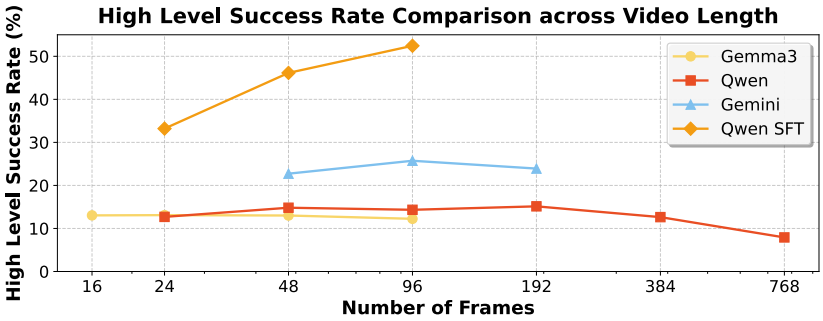


Fig. 5: Performance of different VLMs at different at different sub-sampling rate. Frozen VLMs show little improvement with longer contexts, while the fine-tuned Qwen model benefits from additional frames.

(see Appendix B.4). This underscores the need for future methods that enable VLMs to extract and store in a compressed manner (i.e. via memory mechanisms) task-relevant information more effectively from long video sequences within the context window.

Effect of noisy experience collection. We additionally evaluate robustness to non-oracle experience logs by injecting noise during the experience collection phase. This leads to substantial performance drop for both the zero-shot Qwen agent and the supervised fine-tuned (SFT) model. This experiment demonstrates the extensibility of FindingDory to more challenging settings, and shows that the results reported in the main evaluation represent an approximate upper bound to real world settings (see Appendix F).

Cross-Benchmark Results. In this experiment, we study whether spatiotemporal memory supervision from simulator-based FindingDory data can transfer beyond the simulator to egocentric real-world video reasoning benchmarks. To evaluate this, we use VSI-Bench [53], a benchmark designed to measure spatiotemporal reasoning in VLMs on egocentric real-world videos. We augment the `Video-R1-CoT-165k` dataset introduced in [10] with 40,000 samples from FindingDory.

For effective co-training, we construct chain-of-thought traces for the entire training split of FindingDory (see Appendix B.5). We provide training and evaluation details in Appendix G. We evaluate the trained models on VSI-Bench, with results shown in Tab. 3. Co-training with FindingDory improves the SFT baseline by 1.7% in MCQ accuracy. Furthermore, RL fine-tuning with GRPO on the combined dataset improves performance by 1.85% over the Video-R1-only RL checkpoint (MCQ accuracy). These consistent gains suggest that the

Method	MCQ Acc. (↑)	Num. Acc. (↑)
Zero-Shot	33.69	21.47
Video-R1-Only SFT	33.33	32.67
Video-R1 + FindingDory SFT	35.06	32.35
Video-R1-Only GRPO	33.57	33.9
Video-R1 + FindingDory GRPO	**35.42**	**34.82**

Table 3: `VSI-Bench` results. Models trained with FindingDory video data show consistent gains.

memory supervision provided by FindingDory can improve spatiotemporal reasoning on real-world egocentric video benchmarks.

Qualitative Examples. We present multiple example responses from various VLMs in Appendix I. Overall we observe that closed models perform better than open-source counterparts on single-goal tasks (see Figs. 23, 26 and 27) but all models suffer on multi-goal tasks (see Fig. 28). We also provide example agent video trajectories on our project website: findingdory-benchmark.github.io.

6 Conclusion

We introduced FindingDory, a benchmark for evaluating long-horizon memory in embodied agents within a highly photorealistic indoor simulation. Our findings reveal limitations in VLM context scaling, particularly in handling long observation sequences required for memory-intensive tasks. FindingDory provides informative metrics that disentangle different aspects of memory reasoning, allows for procedural task generation, and efficiency-focused failure analysis beyond simple frame selection accuracy. Its extensibility allows for longer temporal dependencies, supporting research on context scaling and memory efficiency in multimodal models. Our experiments also highlight challenges with hierarchical policies motivating the need for better architectures that closely integrate long-context VLM capabilities with generalist embodied policies.

References

1. Explore vision capabilities with the gemini api. https://ai.google.dev/gemini-api/docs/vision?lang=python
2. Aeronautiques, C., Howe, A., Knoblock, C., McDermott, I.D., Ram, A., Veloso, M., Weld, D., Sri, D.W., Barrett, A., Christianson, D., et al.: Pddl| the planning domain definition language. Technical Report, Tech. Rep. (1998)
3. Anwar, A., Welsh, J., Biswas, J., Pouya, S., Chang, Y.: ReMEmbR: Building and Reasoning Over Long-Horizon Spatio-Temporal Memory for Robot Navigation. arXiv e-prints arXiv:2409.13682 (Sep 2024). https://doi.org/10.48550/arXiv.2409.13682
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A., Wijmans, E.: Objectnav revisited: On evaluation of embodied agents navigating to objects. arXiv preprint arXiv:2006.13171 (2020)
6. Buch, S., Eyzaguirre, C., Gaidon, A., Wu, J., Fei-Fei, L., Niebles, J.C.: Revisiting the "Video" in Video-Language Understanding. arXiv e-prints arXiv:2206.01720 (Jun 2022). https://doi.org/10.48550/arXiv.2206.01720
7. Chang, M., Gervet, T., Khanna, M., Yenamandra, S., Shah, D., Min, S.Y., Shah, K., Paxton, C., Gupta, S., Batra, D., Mottaghi, R., Malik, J., Singh Chaplot, D.: GOAT: GO to Any Thing. arXiv e-prints arXiv:2311.06430 (Nov 2023). https://doi.org/10.48550/arXiv.2311.06430
8. Chaplot, D.S., Gandhi, D.P., Gupta, A., Salakhutdinov, R.R.: Object goal navigation using goal-oriented semantic exploration. Advances in Neural Information Processing Systems **33**, 4247–4258 (2020)
9. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–10 (2018)
10. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
11. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075 (2024)
12. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)
13. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
14. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv e-prints pp. arXiv–2507 (2025)
15. Huang, C., Mees, O., Zeng, A., Burgard, W.: Visual language maps for robot navigation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 10608–10615. IEEE (2023)

16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Kahatapitiya, K., Ranasinghe, K., Park, J., Ryoo, M.S.: Language Repository for Long Video Understanding. arXiv e-prints arXiv:2403.14622 (Mar 2024). https://doi.org/10.48550/arXiv.2403.14622
18. Kemp, C.C., Edsinger, A., Clever, H.M., Matulevich, B.: The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 3150–3157. IEEE (2022)
19. Khanna*, M., Mao*, Y., Jiang, H., Haresh, S., Shacklett, B., Batra, D., Clegg, A., Undersander, E., Chang, A.X., Savva, M.: Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. arXiv preprint (2023)
20. Khanna*, M., Ramrakhya*, R., Chhablani, G., Yenamandra, S., Gervet, T., Chang, M., Kira, Z., Chaplot, D.S., Batra, D., Mottaghi, R.: Goat-bench: A benchmark for multi-modal lifelong navigation (2024)
21. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles (2023)
22. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
23. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. arXiv e-prints arXiv:2311.17005 (Nov 2023). https://doi.org/10.48550/arXiv.2311.17005
24. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=IDJx97BC38
25. Ma, Y., Zang, Y., Chen, L., Chen, M., Jiao, Y., Li, X., Lu, X., Liu, Z., Ma, Y., Dong, X., Zhang, P., Pan, L., Jiang, Y.G., Wang, J., Cao, Y., Sun, A.: Mmlongbench-doc: Benchmarking long-context document understanding with visualizations (2024), https://arxiv.org/abs/2407.01523
26. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., Yadav, K., Li, Q., Newman, B., Sharma, M., Berges, V., Zhang, S., Agrawal, P., Bisk, Y., Batra, D., Kalakrishnan, M., Meier, F., Paxton, C., Sax, S., Rajeswaran, A.: Openeqa: Embodied question answering in the era of foundation models. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
27. Majumdar, A., Yadav, K., Arnaud, S., Ma, J., Chen, C., Silwal, S., Jain, A., Berges, V.P., Wu, T., Vakil, J., et al.: Where are we in the search for an artificial visual cortex for embodied intelligence? Advances in Neural Information Processing Systems **36**, 655–677 (2023)
28. Mangalam, K., Akshulakov, R., Malik, J.: EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. arXiv e-prints arXiv:2308.09126 (Aug 2023). https://doi.org/10.48550/arXiv.2308.09126
29. Min, S.Y., Puig, X., Singh Chaplot, D., Yang, T.Y., Rai, A., Parashar, P., Salakhutdinov, R., Bisk, Y., Mottaghi, R.: Situated Instruction Following. arXiv e-prints arXiv:2407.12061 (Jul 2024). https://doi.org/10.48550/arXiv.2407.12061

30. Ni, T., Ma, M., Eysenbach, B., Bacon, P.L.: When do transformers shine in RL? decoupling memory from credit assignment. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=APGXBNkt6h
31. Pašukonis, J., Lillicrap, T.P., Hafner, D.: Evaluating long-term memory in 3d mazes. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=yHLvIlE9RGN
32. Pleines, M., Pallasch, M., Zimmer, F., Preuss, M.: Memory gym: Partially observable challenges to memory-based agents. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=jHc8dCx6DDr
33. Ramakrishnan, S.K., Wijmans, E., Kraehenbuehl, P., Koltun, V.: Does spatial cognition emerge in frontier models? arXiv preprint arXiv:2410.06468 (2024)
34. Rawal, R., Saifullah, K., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813 (2024)
35. Ren, A.Z., Clark, J., Dixit, A., Itkina, M., Majumdar, A., Sadigh, D.: Explore until confident: Efficient exploration for embodied question answering. In: arXiv preprint arXiv:2403.15941 (2024)
36. Sashank Dorbala, V., Goyal, P., Piramuthu, R., Johnston, M., Ghanadhan, R., Manocha, D.: S-EQA: Tackling Situational Queries in Embodied Question Answering. arXiv e-prints arXiv:2405.04732 (May 2024). https://doi.org/10.48550/arXiv.2405.04732
37. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. ICCV (2019)
38. Saxena, S., Buchanan, B., Paxton, C., Liu, P., Chen, B., Vaskevicius, N., Palmieri, L., Francis, J., Kroemer, O.: GraphEQA: Using 3d semantic scene graphs for real-time embodied question answering. In: 9th Annual Conference on Robot Learning (2025), https://openreview.net/forum?id=Yy9EVIajH5
39. Sethian, J.A.: A fast marching level set method for monotonically advancing fronts. proceedings of the National Academy of Sciences **93**(4), 1591–1595 (1996)
40. Song, D., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking mllms in long context. arXiv preprint arXiv:2404.18532 (2024)
41. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., et al.: Moviechat: From dense token to sparse memory for long video understanding. arXiv preprint arXiv:2307.16449 (2023)
42. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D.S., Maksymets, O., et al.: Habitat 2.0: Training home assistants to rearrange their habitat. In: NeurIPS (2021)
43. Szot, A., Schwarzer, M., Agrawal, H., Mazoure, B., Metcalf, R., Talbott, W., Mackraz, N., Hjelm, R.D., Toshev, A.T.: Large language models as generalizable policies for embodied tasks. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=u6imHU4Ebu
44. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
45. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
46. Wang, C., Ning, R., Pan, B., Wu, T., Guo, Q., Deng, C., Bao, G., Wang, Q., Zhang, Y.: Novelqa: A benchmark for long-range novel question answering (2024)

47. Wani, S., Patel, S., Jain, U., Chang, A., Savva, M.: Multion: Benchmarking semantic map memory using multi-object navigation. NeurIPS (2020)
48. Werby, A., Huang, C., Büchner, M., Valada, A., Burgard, W.: Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In: First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024 (2024)
49. Xie, Q., Min, S.Y., Zhang, T., Xu, K., Bajaj, A., Salakhutdinov, R., Johnson-Roberson, M., Bisk, Y.: Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation. arXiv e-prints arXiv:2409.18313 (Sep 2024). https://doi.org/10.48550/arXiv.2409.18313
50. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge (2025), https://arxiv.org/abs/2501.13468
51. Xu, Z., Chiang, H.T.L., Fu, Z., Jacob, M.G., Zhang, T., Lee, T.W.E., Yu, W., Schenck, C., Rendleman, D., Shah, D., Xia, F., Hsu, J., Hoech, J., Florence, P., Kirmani, S., Singh, S., Sindhwani, V., Parada, C., Finn, C., Xu, P., Levine, S., Tan, J.: Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs. In: 8th Annual Conference on Robot Learning (2024), https://openreview.net/forum?id=JScswMfEQ0
52. Yadav, K., Majumdar, A., Ramrakhya, R., Yokoyama, N., Baevski, A., Kira, Z., Maksymets, O., Batra, D.: Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav. https://arxiv.org/abs/2303.07798 (2023)
53. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. arXiv preprint arXiv:2412.14171 (2024)
54. Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., Gan, C.: 3d-mem: 3d scene memory for embodied exploration and reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17294–17303 (June 2025)
55. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., Kira, Z., Savva, M., Chang, A., Chaplot, D.S., Batra, D., Mottaghi, R., Bisk, Y., Paxton, C.: Homerobot: Open vocabulary mobile manipulation. In: 7th Annual Conference on Robot Learning (2023)
56. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., et al.: Homerobot: Open-vocabulary mobile manipulation. arXiv preprint arXiv:2306.11565 (2023)
57. Yokoyama, N., Ha, S., Batra, D., Wang, J., Bucher, B.: Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 42–48. IEEE (2024)
58. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering (2023)
59. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024), https://arxiv.org/abs/2406.16852
60. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264 (2024)
61. Zhou, K., Zheng, K., Pryor, C., Shen, Y., Jin, H., Getoor, L., Wang, X.E.: Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In: International Conference on Machine Learning. pp. 42829–42842. PMLR (2023)

62. Zhu, H., Kapoor, R., Min, S.Y., Han, W., Li, J., Geng, K., Neubig, G., Bisk, Y., Kembhavi, A., Weihs, L.: Excalibur: Encouraging and evaluating embodied exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14931–14942 (June 2023)