OceanGym: Evaluating Language-Grounded Embodied Agents in Underwater Environments

Anonymous ACL submission

Abstract

We introduce OCEANGYM, the first comprehensive benchmark for underwater embodied agents, designed to advance AI in one of most demanding real-world environments. OCEANGYM features eight realistic task domains and a unified agent framework driven by Multi-modal Large Language Models (MLLMs), which integrates perception, memory, and sequential decision-making. Agents must follow natural language instructions, interpret optical and sonar inputs, autonomously explore complex environments, and achieve long-horizon goals. Extensive experiments reveal a significant performance gap between state-of-the-art MLLM-driven agents and human experts, underscoring persistent difficulties in perception, planning, and adaptation underwater. By offering a high-fidelity, rigorously designed platform, OCEANGYM establishes a testbed for developing robust embodied AI and enabling sim-to-real transfer to Autonomous Underwater Vehicles (AUVs), marking a decisive step toward agents capable of operating in one of Earth's last unexplored frontiers.

1 Introduction

As Richard S. Sutton famously noted, we are entering an "era of experience" (Silver and Sutton, 2025). Embodied agents equipped with language models (Zhao et al., 2023; OpenAI, 2024) are emerging as a central paradigm for intelligent systems, as they accumulate and leverage experience through continuous interaction to close the perception-decision-action loop in physical or simulated environments (Gupta et al., 2021; Ding et al., 2024; Liu et al., 2025). Unlike static decision or generative models, these agents must integrate rich multimodal sensory streams and execute continuous-control policies to achieve long-horizon objectives. This necessitates a unified treatment of perceptual representation, planning, online inference, and sequential policy optimization (Fung et al., 2025). Significant progress has been demonstrated across diverse domains (Anderson et al., 2018; Ma et al., 2025b; Gao et al., 2024).

In contrast, underwater embodied agents remain largely unexplored despite their critical scientific and engineering importance (Visbeck, 2018; Kelly et al., 2022; Zheng et al., 2023; Li et al., 2024b; Gao et al., 2025a). Deploying embodied agents in marine environments offers unique opportunities for ocean exploration, offshore resource development, environmental monitoring, and subsea rescue operations. These tasks impose stringent requirements on the robustness and reliability of autonomous platforms, making the development of agents capable of functioning under real marine conditions a key bridge between simulated research and practical deployment (Ma et al., 2025a).

Challenges. Underwater embodied agents face distinct challenges that set them apart from overland and aerial systems. *Perceptually*, poor visibility and low-light conditions, combined with the inherent limitations of optical sensors, compel reliance on sonar, inertial measurements, and other sparse modalities (Li et al., 2024c; Aubard et al., 2025). These heterogeneous and noisy observations complicate sensor fusion and perception. *Environmentally*, deep-sea and offshore settings are largely unexplored, with unstable localization, absent prior knowledge, and dynamic currents. The lack of prior knowledge prevents effective environmental modeling, requiring agents to reason under circumstances of extreme partial observability and uncertainty (Sariman et al., 2025). Together, these factors constrain the development of underwater agents, leaving their capabilities in early stages.

Building OceanGym. To address these challenges, we introduce OCEANGYM, an open environment benchmark for underwater embodied agents. OCEANGYM constructs a comprehensive marine environment spanning approximately 800m × 800m (length × width), with dynamically ad-

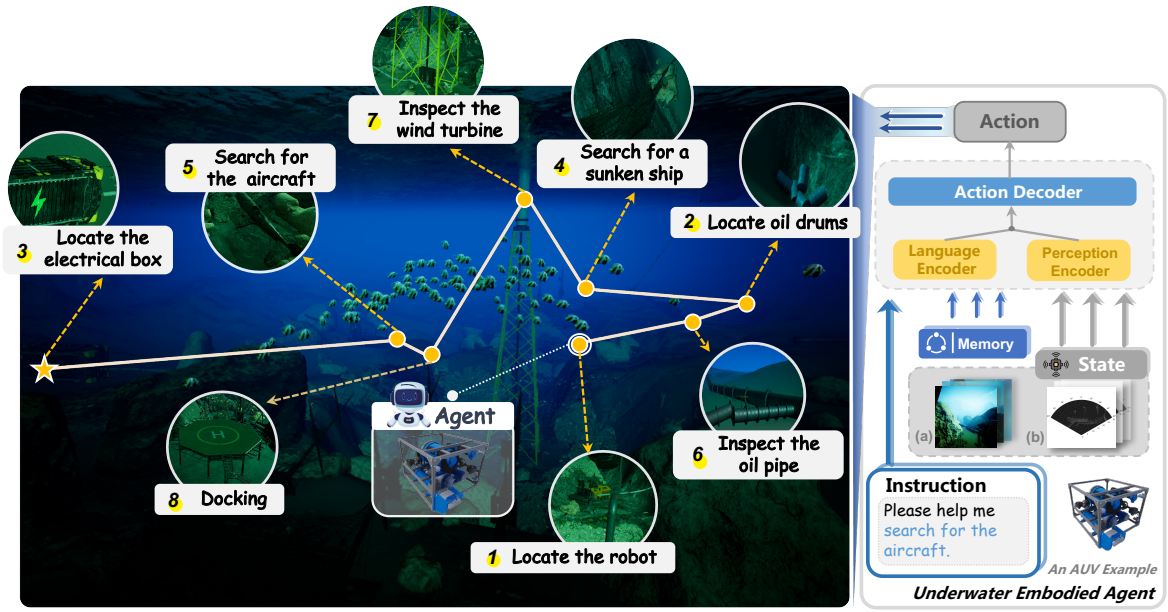

Figure 1: OCEANGYM introduces a unified **agent framework** across **8 real-world underwater scenarios**. The agent interprets instruction, fuses optical and sonar imagery, and controls Autonomous Underwater Vehicles (AUVs).

justable depth to simulate varying lighting conditions. The platform incorporates eight distinct task domains designed to reflect real-world operational requirements. Additionally, it provides a multimodal agent framework that integrates perception, memory, and action decision-making capabilities for controlling AUVs. The benchmark unifies perception and decision-making in simulated underwater scenarios, where agents must infer target states from contextual cues or multi-view sensor data and execute complex behaviors such as search, inspection, and docking. By simulating these environments, OCEANGYM enables systematic evaluation of models' capabilities in underwater embodied settings and offers a initial pathway for transferring learned skills to real-world AUVs.

Benchmark Results and Insights. Extensive experiments demonstrate performance gaps between MLLMs and human experts, especially under low-visibility conditions. Agents often exhibit difficulties in accurately interpreting sonar data, distinguishing objects in complex scenes, and maintaining consistent decision-making strategies over prolonged missions. Additional limitations emerge in memory retention and adaptability when objects are occluded or environmental conditions change.

2 OceanGym

OCEANGYM is a high-fidelity embodied underwater environment that simulates a realistic ocean setting with diverse scenes. As illustrated in Figure 2, OCEANGYM establishes a robust benchmark for evaluating autonomous agents through a series of challenging tasks, encompassing various perception analyses and decision-making navigation. OCEANGYM facilitates these evaluations by enabling MLLM-driven agents with multi-modal perception and parameterized action spaces.

2.1 OceanGym Environment

We develop OCEANGYM atop Unreal Engine (UE) 5.3 (Epic Games, 2025), providing a comprehensive set of underwater environments, including both natural terrains and engineered structures. The environment features several semantic regions such as open water, seabed plains, underwater cliffs, pipeline networks, wreckage sites, and energy infrastructure zones. Each region is modeled with realistic physical and geometric properties, incorporating elements like oil pipelines, chemical waste barrels, submerged shipwrecks, electrical equipment, wind turbine foundations, and aircraft debris (more details in Appendix E). We simulate varying lighting conditions by adjusting underwater depth, considering two representative scenarios: shallow (50m) and deep (500m) water. In deep water, optical sensing depends on artificial lighting with a visibility range of about 0–10m. Moreover, OCEANGYM is scalable: users can customize the environment by adjusting depths for more complex lighting, introducing new objects, or designing additional tasks within the existing framework, thereby extending diversity and difficulty.

2.2 Underwater Embodied Agents

We model the agent's control–perception loop as a Partially Observable Markov Decision Process (POMDP) enhanced with contextual memory. At each time step t, the agent processes the task specification $\mathcal{T} = (I_{\text{target}}, c)$, where I_{target} is a visual reference image of the target and c provides its textual identity and features. It also considers language

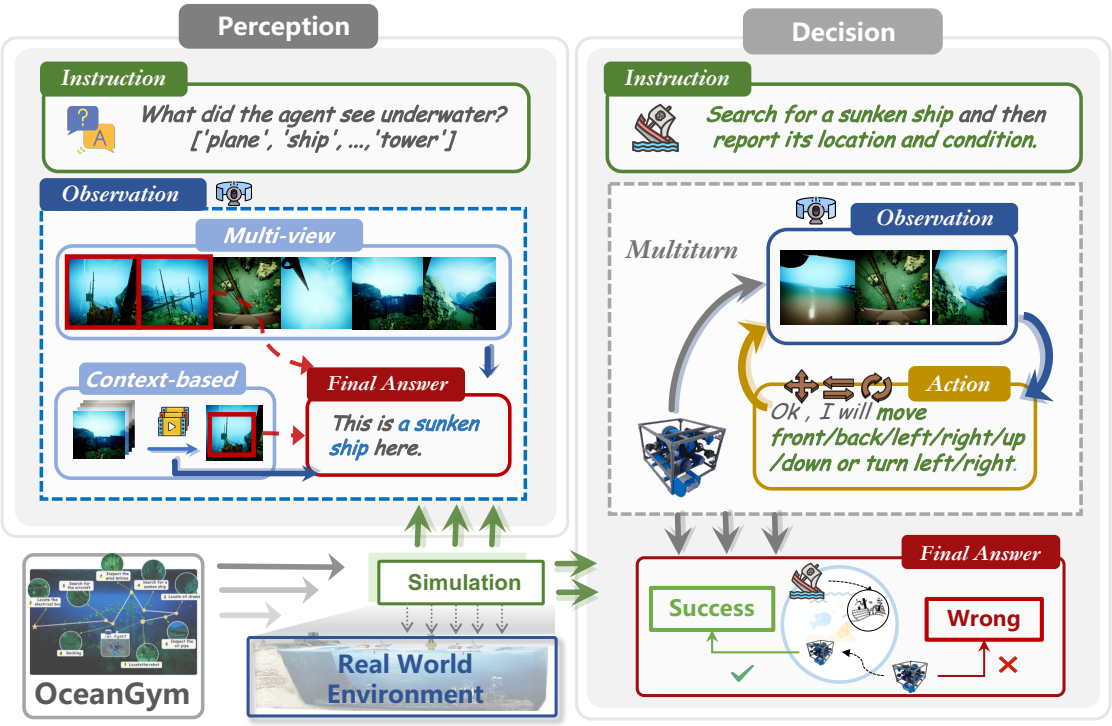

Figure 2: **OceanGym Tasks.** OceanGym comprises **Perception Tasks** (divided into **Multi-view Perception** and **Context-based Perception** settings) and **Decision Tasks** for evaluating embodied agents.

instruction L, synchronized observations O_t, and its memory state m_t. These elements collectively shape the agent's perception and objectives.

With the above information, the agent must generate either a textual perception response y_t for perception tasks, or determine a control action a_t for decision tasks. Here, $a_t \in \mathcal{A}$ is a discrete action selected from the action space $\mathcal{A}$. A decision trajectory is described by $\sigma = (O_1, a_1, s_1, m_1, \ldots, m_{t-1}, O_t, a_t, s_t)$. In this sequence, O_i represents the observations, a_i the actions, s_i the states, and m_i the memory states at each time step i. The episode concludes when the target is achieved or when the maximum time t_{max} is exhausted. The ultimate reward is based on the successful score of the task, as defined in §2.5.

State and Observation. The agent's state at time t is given by $s_t = \{(x_t, y_t, z_t), (\phi_t, \theta_t, \psi_t)\}$, where (x_t, y_t, z_t) represent the agent's 3D positional coordinates, and $(\phi_t, \theta_t, \psi_t)$ denote the roll, pitch, and yaw angles, respectively. At each timestep, the agent receives synchronized RGB and sonar images from sensors oriented in six different directions. The directions are defined by the set $\mathcal{D}_{\text{sensor}} = \{\text{f}, \text{b}, \text{l}, \text{r}, \text{u}, \text{d}\}$, corresponding to `front`, `back`, `left`, `right`, `up`, and `down`. The RGB images from these directions are denoted as $O_t^R = \{o_{t,d}^R\}_{d \in \mathcal{D}_{\text{sensor}}}$, and the sonar images are represented similarly as $O_t^S = \{o_{t,d}^S\}_{d \in \mathcal{D}_{\text{sensor}}}$. Therefore, the complete observation at time t can be expressed as a combination of both image sets, $O_t = (O_t^R, O_t^S)$.

Action Space. The agent's action direction set is defined as $\mathcal{D}_{\text{action}} = \{f, b, l, r, u, d, rl, rr\}$, which encompasses both directional and rotational movements. Directional actions include translations along the primary axes: `forward` (f), `backward` (b), `left` (l), `right` (r), `up` (u), and `down` (d). Rotational actions consist of `rotate left` (rl) and `rotate right` (rr). At each timestep t, the agent selects an action $a_t \in \mathcal{A}$ from this discrete set and applies a control magnitude $\delta \in \mathbb{R}_{\geq 0}$ to determine the execution intensity.

Memory. Memory systems play a crucial role in storing and structuring past information, thereby enhancing the agent's resilience in dynamic and partially observable environments (Xi et al., 2025; Liu et al., 2023; Zhong et al., 2024; Wu et al., 2024; Maharana et al., 2024). OceanGym agent maintains an explicit memory m_t, structured as a sliding window that records the last K steps:

$$m_t = \{(d_{t-k}, a_{t-k}) \mid k = 1, 2, \ldots, K\} \quad (1)$$

Within this memory structure, d_{t-k} denotes the textual description at time $t-k$, and a_{t-k} represents the corresponding action executed. The sliding window size K is implemented primarily to prevent the context length from exceeding the model's maximum input capacity. The default window size is large enough to capture the necessary historical information for most tasks in our benchmark. The perception module $\mathcal{P}_\theta$, modeled as an MLLM, generates a summary based on the current context and the interaction history $\{(O_k, a_k)\}_{k=t-K}^{t}$:

$$d_t = \mathcal{P}_\theta\big(\{(O_k, a_k)\}_{k=t-K}^{t}\big) \quad (2)$$

This summary is subsequently used to refresh the memory: $m_{t+1} = \text{update}(m_t, d_t, a_t)$.

Memory-augmented Markov Process. To maintain the Markov property while incorporating memory, we introduce an augmented state $\tilde{s}_t = (s_t, m_t)$. The state transition is modeled as:

$$p(\tilde{s}_{t+1} \mid \tilde{s}_t, a_t, \delta) \tag{3}$$

where $p(\cdot \mid \cdot)$ represents the augmented state transition function of the environment. This function captures both the evolution of memory, ensuring that the system remains Markovian despite the added complexity of memory integration.

Agent Policy. The agent policy is a multimodal, memory-augmented mapping parameterized by an MLLM with parameter vector θ:

$$\pi_\theta(a_t, y_t \mid L, O_t, m_t, \mathcal{T}, \delta) \tag{4}$$

Concretely, for perception tasks, we sample an answer $y_t \sim \pi_\theta(y \mid L, O_t, m_t, \mathcal{T}, \delta)$, and for decision tasks, we sample an action $a_t \sim \pi_\theta(a \mid L, O_t, m_t, \mathcal{T}, \delta)$. An episode terminates at time T when the agent either outputs a STOP command (for decision tasks) or provides a final answer to the question (for perception tasks) or when the maximum time $t_{\max}$ is reached. The policy, combined with the memory-augmented transition dynamics, induces the trajectory distribution:

$$\mathbb{P}_\theta(\sigma \mid L, \mathcal{T}) = \prod_{t=1}^{T-1} \pi_\theta(a_t, y_t \mid L, O_t, m_t, \mathcal{T}, \delta) \times p(\tilde{s}_{t+1} \mid \tilde{s}_t, a_t, \delta) \tag{5}$$

where σ represents the trajectory of the agent through the state space over time, influenced by the specified policy π_θ and the transition model $p(\tilde{s}_{t+1} \mid \tilde{s}_t, a_t, \delta)$.

Running Example: Shipwreck Area

Perception Task: (1) Multi-view perception setting. The agent receives perception images (visual and sonar) from different sensors at the same time to determine the target, such as whether it is a shipwreck. (2) Context-based perception setting. The agent analyzes images one by one along a trajectory from a fixed viewpoint to identify the target.

Decision Task: The agent receives a task instruction, such as "Search for a sunken ship," and then explores for 30 minutes to complete it.

2.3 OceanGym Perception Tasks

The perception tasks are categorized into two settings: **Multi-View Perception** and **Context-based Perception**. These tasks primarily use RGB images as input, with sonar data added in certain experiments to enhance perception. The data for each setting are collected by human operators and designed to evaluate different aspects of MLLMs' perceptual abilities. There are a total of 85 scenes. More details in Appendix D.

Multi-view Perception Setting. This setting evaluates the agent's ability to interpret visual information from multiple synchronized viewpoints. At each timestep t, the agent captures a set of six simultaneous RGB images, denoted as $O_t^R = \{o_{t,d}^R\}_{d \in \mathcal{D}_{\text{sensor}}}$, where d refers to the different sensor orientations: `front`, `back`, `left`, `right`, `up`, and `down`. The objective is to consistently identify and localize underwater objects across these varied viewpoints. This setting examines whether objects visible from certain angles can be correctly perceived when the visual inputs from all directions are sequentially processed by the MLLM, thereby evaluating robustness to viewpoint variations.

Context-based Perception Setting. This setting assesses the agent's ability to perceive and interpret sequential observations gathered during navigation. At each timestep t, the agent captures an RGB image o_t^R from a fixed orientation, forming a chronological sequence $O_{1:m}^R = \{o_t^R\}_{t=1}^m$, where m is the total number of timesteps. The agent must track and understand changes over time, ensuring consistent and accurate identification and localization of underwater objects. This evaluation emphasizes temporal consistency and the agent's capacity to build a coherent perception from a stable visual perspective in dynamic underwater environments.

2.4 OceanGym Decision Tasks

Decision Task Definition. Decision tasks evaluate decision-making in continuous 3D environments, where agents must integrate multimodal sensory input with task specifications. Each episode begins from an initial state $s_0 = \{(x_0, y_0, z_0), (\phi_0, \theta_0, \psi_0)\}$ and requires the agent to reach the target defined by $\mathcal{T}$. The agent must combine sensory observations O_t, temporal memory, and goal information to execute precise maneuvers in cluttered, low-visibility environments. Key parameters of the task include the decision in-

terval t_{interval} and the task's limited duration $t_{\max}$[1]. The decision interval t_{interval} determines how frequently the agent makes decisions and executes actions. The total task duration $t_{\max}$ sets the temporal constraint, within which the agent must meet its objectives, thereby influencing the planning and movement strategies employed by the agent. Compared with grid-based navigation benchmarks, this task emphasizes continuous control and realistic underwater environment, reflecting the challenges of autonomous exploration and inspection tasks.

Decision Task Design. To evaluate the decision-making capabilities of MLLMs in marine environments, we design eight representative task scenarios that are commonly used in actual underwater operations (more details in Appendix E). The task construction methods are divided into two categories: detection tasks and tracking tasks. Detection tasks focus on assessing the ability of MLLMs to locate specific underwater objects, including searching for large targets such as sunken ships or aircraft wreckage, and smaller targets like scientific research robots. Tracking tasks focus on evaluating the ability of MLLMs to perform inspection and monitoring tasks underwater, including scenarios like pipeline inspection and platform approaches. To further investigate the performance in challenging environments, four representative tasks are conducted under low light deep-sea conditions. For each task, the starting position is selected to vary the difficulty, as tasks become more challenging when the starting point is far from the target, when the target is initially out of view, or when the agent initially faces away from the goal. In our experiments, we follow a consistent initial positioning strategy: the first two starting positions are the same across all tasks to ensure reproducibility, while the third starting position is randomly set within the operational area to test how well the agent adapts to different starting conditions.

2.5 Evaluation Metrics

Perception Task Evaluation. We evaluate model performance using exact match accuracy. Let y_i denote the ground-truth answer and $\hat{y}_i$ represent the model's predicted answer for the i-th sample.

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\hat{y}_i = y_i] \tag{6}$$

[1]By default, t_{interval} takes 30 seconds and $t_{\max}$ takes 0.5 hours in decision tasks.

For multiple-choice items, y_i and $\hat{y}_i$ are treated as sets and equality requires an exact set match.

Decision Task Evaluation. We evaluate decision tasks using a distance-based scoring method. Each episode ends when the agent issues a STOP command or reaches the time limit $t_{\max}$. For a task with n evaluation points, let $\mathbf{p}_i$ be the i-th target location. If the target is detected, we use the closest position from the agent's trajectory to $\mathbf{p}_i$; otherwise, we use the agent's final position. The Euclidean distance is computed as $d_i = \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2$, and the score for each point is defined as:

$$S_i = \begin{cases} 100, & d_i \le \tau_1, \\ 100\,\dfrac{\tau_2 - d_i}{\tau_2 - \tau_1}, & \tau_1 < d_i \le \tau_2, \\ 0, & d_i > \tau_2, \end{cases} \tag{7}$$

where the distance thresholds are set to $\tau_1 = 30$ meters and $\tau_2 = 100$ meters by default. The total score is a weighted sum as $S_{\text{total}} = \sum_{i=1}^{n} w_i S_i$, where w_i are task-specific weights[2].

3 Experiments

3.1 Experimental Settings

To thoroughly evaluate the perception and decision capabilities of MLLMs in underwater environments, we conduct experiments using a variety of models [3]. Among the open-source models, we assess MiniCPM-V-4.5 (Yao et al., 2025), GLM-4.5V (Team et al., 2025) and Qwen2.5-VL-7B (Bai et al., 2025). For proprietary models, we test GPT-4o-mini and GPT-5.2 (OpenAI, 2024), as well as Gemini-2.5-Flash and Gemini-3-Pro (Gemini Team, 2024). **We run each task three times and report the average results.** Human volunteers in the laboratory provide answers based on perception task instructions, and operate robots through keyboards for decision-making tasks.

3.2 Main Results

Perception Results. The results for perception tasks are summarized in Table 1. In shallow, well-illuminated water environments, Gemini-3-Pro achieves the strongest overall performance among the evaluated MLLMs, with an average

[2]For a single-point task $w_1 = 1.0$; for two points $(w_1, w_2) = (0.6, 0.4)$; for three points $(w_1, w_2, w_3) = (0.6, 0.2, 0.2)$.

[3]We prefer small MLLMs that can run on edge devices.

Model	Shallow Water Environment (High Illumination)					Deep Water Environment (Low Illumination)				
	Multi-View Perception		Context-based Perception		Avg	Multi-View Perception		Context-based Perception		Avg
	Vision	+Sonar	Vision	+Sonar		Vision	+Sonar	Vision	+Sonar	
GLM-4.5V	52.73	**56.36**	46.67	63.33	54.77	36.36	30.91	20.00	33.33	30.15
GPT-4o-mini	34.55	34.55	20.00	33.33	30.61	14.55	20.00	3.33	6.67	11.14
Gemini-2.5-Flash	29.09	30.91	50.00	33.33	35.83	9.09	5.45	20.00	30.00	16.14
Qwen2.5-VL-7B	**58.18**	43.64	56.67	70.00	57.12	27.27	20.00	33.33	33.33	28.48
Minicpm-4.5	52.73	43.64	36.67	23.33	39.09	29.09	23.64	**43.33**	13.33	27.35
GPT-5.2	56.36	**56.36**	36.67	**73.33**	55.68	36.36	36.36	13.33	33.33	21.60
Gemini-3-Pro	45.45	54.55	**66.67**	70.00	**59.17**	**40.00**	**38.18**	36.67	**36.67**	**37.88**
Human	100.00	100.00	100.00	100.00	100.00	94.55	98.18	86.67	90.00	92.35

Table 1: Performance of perception tasks across different models and conditions. Values represent accuracy percentages (%). Bold indicates the highest value in each column, underline indicates the second highest. Adding sonar means using both RGB and sonar images.

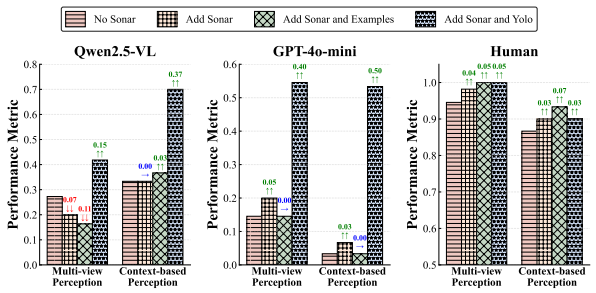

Figure 3: Performance comparison between human and MLLMs after adding sonar and sonar reference examples for objects in deep water environments.

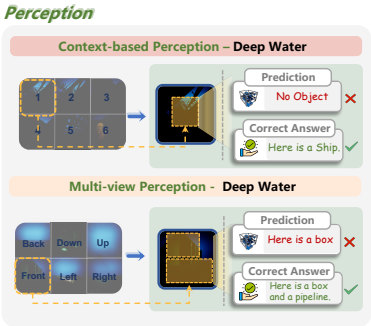

Figure 4: Case analysis in perception tasks. Agents are susceptible to errors under challenging conditions such as under low-light, multi-object, or occluded conditions.

accuracy of 59.17%, followed by Qwen2.5-VL-7B (57.12%) and GLM-4.5V (54.77%). Under deep water conditions with low illumination, all models exhibit significant performance degradation, though Gemini-3-Pro emerges as the most robust (37.88% average accuracy), followed by GLM-4.5V (30.15%), Qwen2.5-VL-7B (28.48%), and Minicpm-4.5 (27.35%). Notably, incorporating sonar data does not consistently improve performance across models or tasks. The performance gap between MLLMs and human level indicates substantial room for improvement in underwater perception tasks and degraded visual inputs.

Decision Results. Performance on decision tasks is shown in Table 2. Several tasks resulted in zero scores, indicating extreme difficulty due to small object size or time constraints. GPT-4o-mini achieves the best average performance in both shallow (18.4%) and deep water (14.8%) environments, with GLM-4.5V ranking second under shallow conditions (16.1%) and deep water conditions (11.5%). Performance declines markedly in deep water, where Gemini-2.5 and Qwen2.5-VL-7B both average 6.4%. Human performance substantially outperforms all models, reaching 100% in shallow water and 69.6% in deep water, underscoring the gap between MLLM-driven agents and humans.

3.3 Analysis

MLLM agents struggle to exploit sonar data for enhanced underwater perception, in stark contrast to humans who leverage it effectively. To investigate the role of sonar data in deep-water environments, we compare the performance of human experts with the two MLLMs, Qwen2.5-VL and GPT-4o-mini, on perception tasks. Specifically, we either let the models directly comprehend sonar images or provide them with human-annotated interpretations as prompts. As shown in Figure 3, human experts consistently benefit from incorporating sonar data across tasks. By contrast, MLLMs exhibit only limited gains when using raw sonar images, and this gap becomes even more pronounced when reference sonar images of each object are introduced. This limitation likely stems from current MLLMs' fundamental difficulty in interpreting sonar imagery and underwater perceptual data (Xie et al., 2022; Zheng et al., 2023; Xu et al., 2025; Aubard et al., 2025), combined with potential constraints in the sonar simulation within OceanGym. Notably, when employing a YOLO model (Redmon et al., 2016) specifically trained on sonar data as auxiliary perception tools, we observe performance improvements, suggesting that specialized vision models may currently outperform general-

Task	Model				Human
	GLM-4.5V	GPT-4o-mini	Gemini-2.5	Qwen2.5-VL-7B	
Shallow Water Environment (High Illumination)					
Locate the robot	$6.6_{\pm 19.83}$	$\mathbf{8.9}_{\pm 10.1}$	$0.0_{\pm 0.00}$	$7.8_{\pm 13.5}$	100
Locate the oil drums	$10.7_{\pm 16.52}$	$\mathbf{11.1}_{\pm 19.2}$	$3.5_{\pm 6.0}$	$5.7_{\pm 9.8}$	100
Locate the electrical box	$7.9_{\pm 17.08}$	$\mathbf{36.6}_{\pm 21.9}$	$15.9_{\pm 27.4}$	$8.7_{\pm 15.0}$	100
Search for a sunken ship	$5.9_{\pm 5.04}$	$13.4_{\pm 19.3}$	$\mathbf{20.5}_{\pm 14.3}$	$10.3_{\pm 10.3}$	100
Search for the aircraft	$\mathbf{25.0}_{\pm 6.10}$	$16.9_{\pm 17.8}$	$11.7_{\pm 15.6}$	$7.8_{\pm 10.0}$	100
Inspect oil pipe	$\mathbf{37.8}_{\pm 17.88}$	$27.1_{\pm 23.6}$	$18.3_{\pm 15.8}$	$30.8_{\pm 25.2}$	100
Inspect the wind turbine	$20.3_{\pm 28.89}$	$13.9_{\pm 14.33}$	$\mathbf{25.1}_{\pm 22.1}$	$14.7_{\pm 17.0}$	100
Docking	$14.9_{\pm 13.20}$	$19.2_{\pm 33.28}$	$\mathbf{19.4}_{\pm 33.6}$	$8.3_{\pm 7.2}$	100
Average	$16.1_{\pm 15.6}$	$\mathbf{18.4}_{\pm 19.9}$	$14.4_{\pm 16.1}$	$11.8_{\pm 13.7}$	100
Deep Water Environment (Low Illumination)					
Locate oil drums	$\mathbf{10.6}_{\pm 21.35}$	$5.6_{\pm 9.69}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	40.8
Search for a sunken ship	$2.9_{\pm 2.16}$	$\mathbf{12.8}_{\pm 14.48}$	$8.2_{\pm 14.1}$	$3.4_{\pm 5.8}$	100
Inspect the oil pipe	$\mathbf{32.5}_{\pm 5.86}$	$15.8_{\pm 15.5}$	$6.6_{\pm 11.4}$	$21.7_{\pm 25.3}$	78.2
Inspect the wind turbine	$0.0_{\pm 0.0}$	$\mathbf{25.1}_{\pm 16.0}$	$10.6_{\pm 10.0}$	$0.4_{\pm 0.6}$	100
Average	$11.5_{\pm 7.3}$	$\mathbf{14.8}_{\pm 13.9}$	$6.4_{\pm 8.8}$	$6.4_{\pm 8.4}$	69.6

Table 2: Performance in decision tasks requiring autonomous completion by MLLM-driven agents.

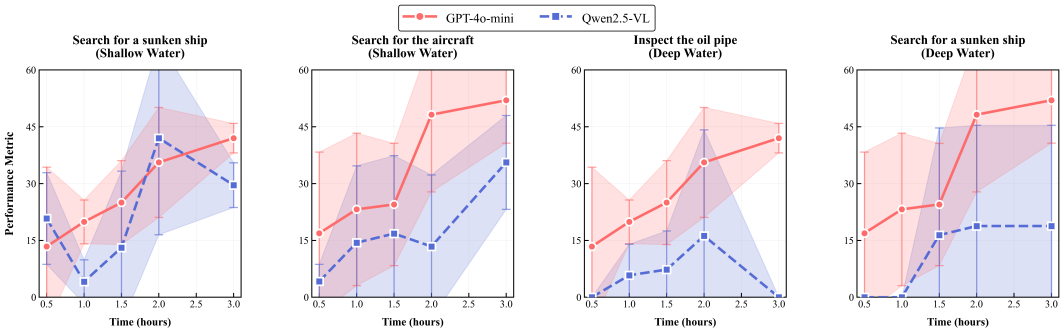

Figure 5: Scaling analysis performance over time in decision tasks.

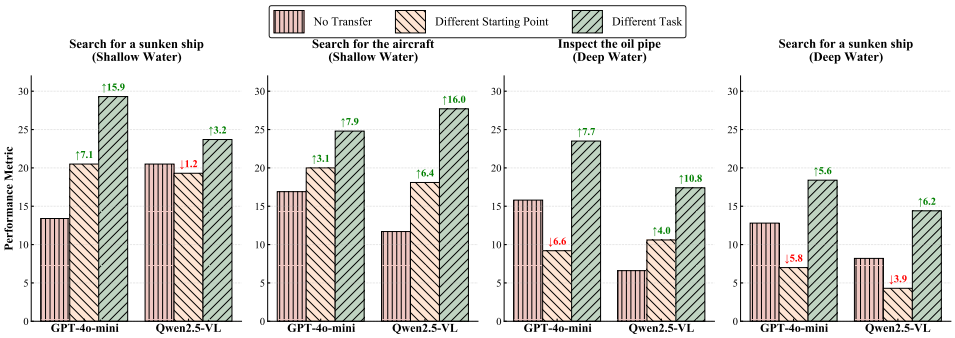

Figure 6: Impact of different memory transfer mechanisms on model performance.

purpose MLLMs in sonar data usage.

Extended exploration boosts agent knowledge and task performance, following a scaling law that eventually plateaus. We analyze the relationship between navigation performance and operational duration using the representative MLLMs, across both shallow- and deep-water scenarios. The performance was evaluated over durations of 0.5, 1, 1.5, 2, and 3 hours. As shown in Figure 5, performance initially improves with longer operation time, consistent with prior studies on test-time scaling (Zhang et al., 2025a; Zhu et al., 2025), but eventually plateaus. This plateau reflects inherent limitations in perception, memory, and reasoning, as well as a lack of intrinsic curiosity to explore new regions. These findings underscore the need to improve both fundamental MLLM capabilities and agent strategies, such as enhanced memory and long-horizon planning, to break through performance ceilings in embodied environments.

Memory transfer enables agents to leverage past experience to tackle new challenges. We investigate whether knowledge and experience accumulated from previous tasks (Hou et al., 2024; Hu et al., 2024a; Tan et al., 2025; Tang et al., 2025) can enhance performance in new tasks. Specifically, we explore using agents' previously explored trajectories as experiential input. Experiments are conducted in both shallow water and deep water environments, evaluating two transfer condi-

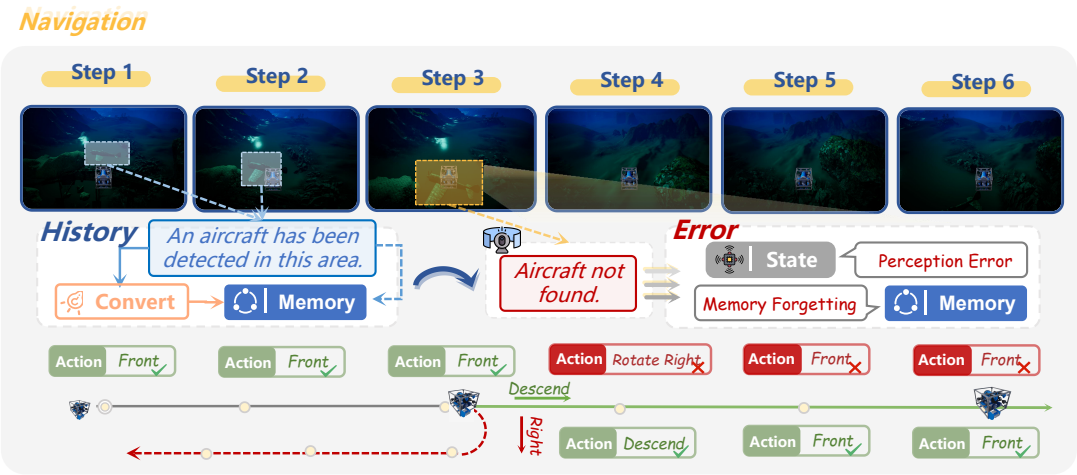

Figure 7: Case analysis in decision tasks.

tions: within-task transfer (different starting points) and cross-task transfer (different but related tasks). As shown in Figure 6, memory transfer improves decision-making performance in shallow water environments under both transfer conditions. However, in the more challenging deep water environment, only cross-task transfer demonstrates stable performance improvements, while within-task transfer shows limited benefits. This suggests that more appropriate prior experiences provide more robust guidance under perceptually degraded conditions. Transfer learning helps compensate for perceptual limitations by providing informed priors about environmental structure and effective navigation strategies. These findings underscore the importance of developing adaptive memory retrieval mechanisms that can selectively leverage relevant past experiences to enhance decision-making in autonomous underwater agents operating under diverse environmental conditions.

Case analysis. We present perception case analyses and illustrate failure cases in Figure 4, mainly due to: **(1) Occlusions**, where targets are partially blocked; **(2) Multi-object Scenes**, causing identification and localization ambiguities; and **(3) Low Illumination**, which severely reduces vision-based perception accuracy. Figure 7 shows common decision task failures, primarily from: **(1) Perception Errors**, where inaccurate detection leads to wrong actions; and **(2) Memory Forgetting**, where the agent cannot retain crucial past information, such as visited locations or previous decisions.

Sim-to-Real transfer analysis. To bridge simulation and reality, we deploy physical reference objects from OceanGym's object modeling into a real marine environment, enabling correlation between simulated and real-world performance. An AUV equipped with a sonar data acquisition system collects sonar measurements. As shown in Figure 9, **the YOLO model trained in simulation enhances GPT-4o-mini's ability to interpret real-world sonar data**. However, it exhibits limited generalization to objects not included in the simulation. As illustrated in Figure 10, **aligning the model's output actions with real operational commands allows prompts from the simulated environment to directly control real AUVs**. Furthermore, with environment memory that mirrors reality, the model can rapidly follow language commands to locate or reach designated targets.

4 Related Work

Embodied intelligence relies on simulation platforms across terrestrial (Chang et al., 2017; Wu et al., 2018; Manolis Savva et al., 2019), aerial (Shah et al., 2018; Lee et al., 2025; Wang et al., 2024a), and marine domains (Potokar et al., 2022; Song et al., 2025; Chu et al., 2025). The emergence of MLLMs has advanced embodied agents by integrating visual understanding (OpenAI, 2024; Bai et al., 2025; Gemini Team, 2024). Existing embodied benchmarks cover indoor (Anderson et al., 2018; Wu et al., 2018), urban (Chen et al., 2019; Gao et al., 2024), aerial (Gao et al., 2025b; Cai et al., 2025) and real-world (Zhao et al., 2025; Koh et al., 2024) settings, yet a unified benchmark for underwater multimodal embodied agents remains absent. More related work in Appendix B.

5 Conclusion

We introduce OceanGym, the first bechmark environment designed for underwater embodied agents, hopping it can bridge the gap between simulated research and real-world embodied robots. OceanGym addresses the unique challenges of marine environments through a high-quality simulation platform featuring realistic underwater dynamics, multi-modal sensing (RGB and sonar), and diverse task scenarios. Our experiments reveal significant limitations in current MLLMs.

Limitations

While OceanGym provides a valuable testbed for underwater embodied agents, several limitations should be acknowledged. First, OceanGym leverages Unreal Engine (UE) 5.3 (Epic Games, 2025) for realistic underwater environment rendering and physical simulation, while utilizing HoloOcean's (Potokar et al., 2022) cluster-based multipath ray-tracing algorithm to simulate multibeam sonar. Although UE plugins, as exemplified by those shown in Figure 20, can be used to simulate water flow, buoyancy, lighting, water interaction etc, they cannot fully replicate the real underwater environment, as factors such as ocean currents, salinity, marine life, and geological changes are not accurately captured. Future work may leverage generative models (Ball et al., 2025) or physics-informed machine learning to incorporate these complexities. The optical and sonar images still differ from those in the real world, particularly since sonar simulation introduces errors. We will continue to refine the system to reduce these discrepancies, noting that real-world sonar itself is also subject to noise and inaccuracies. In addition, the environment is large and requires considerable computational resources. These limitations highlight opportunities for future work to expand task coverage, improve physical realism, and optimize computational efficiency.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sara Aldhaheri, Yang Hu, Yongchang Xie, Peng Wu, Dimitrios Kanoulas, and Yuanchang Liu. 2025. Underwater robotic simulators review for autonomous system development. *Preprint*, arXiv:2504.06245.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin Aubard, Ana Madureira, Luís Teixeira, and José Pinto. 2025. Sonar-based deep learning in underwater robotics: Overview, robustness, and challenges. *IEEE Journal of Oceanic Engineering*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Philip J Ball, J Bauer, F Belletti, and 1 others. 2025. Genie 3: A new frontier for world models.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Hengxing Cai, Jinhan Dong, Jingjun Tan, Jingcheng Deng, Sihang Li, Zhifeng Gao, Haidong Wang, Zicheng Su, Agachai Sumalee, and Renxin Zhong. 2025. Flightgpt: Towards generalizable and interpretable UAV vision-and-language navigation with vision-language models. *CoRR*, abs/2505.12835.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539.

Yaran Chen, Wenbo Cui, Yuanwen Chen, Mining Tan, Xinyao Zhang, Dongbin Zhao, and He Wang. 2023. Robogpt: an intelligent agent of making embodied long-term decisions for daily instruction tasks. *arXiv preprint arXiv:2311.15649*.

Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, and 1 others. 2025. Embodiedeval: Evaluate multimodal llms as embodied agents. *arXiv preprint arXiv:2501.11858*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Shuguang Chu, Zebin Huang, Yutong Li, Mingwei Lin, Ignacio Carlucho, Yvan R. Petillot, and Canjun Yang. 2025. Marinegym: A high-performance reinforcement learning platform for underwater robotics. *Preprint*, arXiv:2503.09203.

Ronghao Dang, Yuqian Yuan, Yunxuan Mao, Kehan Li, Jiangpin Liu, Zhikai Wang, Xin Li, Fan Wang, and Deli Zhao. 2025. Rynnec: Bringing mllms into embodied world. *Preprint*, arXiv:2508.14160.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, and 1 others. 2024. Understanding world or predicting future? a comprehensive survey of world models. *arXiv preprint arXiv:2411.14499*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *CoRR*, abs/2406.05756.

Epic Games. 2025. Unreal engine.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, and 1 others. 2025. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel Bear, Dan Gutfreund, David D. Cox, and 5 others. 2021. Threedworld: A platform for interactive multi-modal physical simulation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. 2024. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. *Preprint*, arXiv:2410.09604.

Yuan Gao, Ruiqi Shu, Hao Wu, Fan Xu, Yanfei Xiang, Ruijian Gou, Qingsong Wen, Xian Wu, and Xiaomeng Huang. 2025a. Neuralom: Neural ocean model for subseasonal-to-seasonal simulation. *arXiv preprint arXiv:2505.21020*.

Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, and 4 others. 2025b. Openfly: A comprehensive platform for aerial vision-language navigation. *Preprint*, arXiv:2502.18041.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Mingning Guo, Mengwei Wu, Jiarun He, Shaoxian Li, Haifeng Li, and Chao Tao. 2025. BEDI: A comprehensive benchmark for evaluating embodied agents on uavs. *CoRR*, abs/2505.18229.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721.

Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, Rongtao Xu, and Shibiao Xu. 2025. Multimodal fusion and vision-language models: A survey for robot vision. *CoRR*, abs/2504.02477.

Yining Hong, Rui Sun, Bingxuan Li, Xingcheng Yao, Maxine Wu, Alexander Chien, Da Yin, Ying Nian Wu, Zhecan James Wang, and Kai-Wei Chang. 2025. Embodied web agents: Bridging physical-digital realms for integrated agent intelligence. *CoRR*, abs/2506.15677.

Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. " my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2024a. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model.

Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, Yuhuai Li, Shengze Xu, Shawn Wang, Xinchen Xu, Shuofei Qiao, Kun Kuang, Tieyong Zeng, Liang Wang, Jiwei Li, and 9 others. 2024b. Os agents: A survey on mllm-based agents for general computing devices use. `https://github.com/OS-Agent-Survey/OS-Agent-Survey/`.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. 2023. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.

Rachel Kelly, Laura G Elsler, Andrei Polejack, Sander van der Linden, Kajsa Tönnesson, Sarah E Schoedinger, Francesca Santoro, Gretta T Pecl, Michael Palmgren, Patrizio Mariani, and 1 others. 2022. Empowering young people with climate and ocean science: Five strategies for adults to consider. *One Earth*, 5(8):861–874.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. 2024. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, and 1 others. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Jungdae Lee, Taiki Miyanishi, Shuhei Kurita, Koya Sakamoto, Daichi Azuma, Yutaka Matsuo, and Nakamasa Inoue. 2025. Citynav: A large-scale dataset for real-world aerial navigation. *Preprint*, arXiv:2406.14240.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, and 1 others. 2021. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Jose Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, and 16 others. 2024a. BEHAVIOR-1K: A human-centered, embodied AI benchmark with 1, 000 everyday activities and realistic simulation. *CoRR*, abs/2403.09227.

Yun Li, Yiming Zhang, Tao Lin, Xiangrui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. 2025. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *Preprint*, arXiv:2503.23765.

Zhe Li, Ronghui Xu, Jilin Hu, Zhong Peng, Xi Lu, Chenjuan Guo, and Bin Yang. 2024b. Ocean significant wave height estimation with spatio-temporally aware large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3892–3896.

Zikang Li, Zhuojun Xie, Puhong Duan, Xudong Kang, and Shutao Li. 2024c. Dual spatial attention network for underwater object detection with sonar imagery. *IEEE Sensors Journal*, 24(5):6998–7008.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Huaping Liu, Di Guo, and Angelo Cangelosi. 2025. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys*, 57(7):1–36.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023. Think-in-memory: Recalling and post-thinking enable llms with long-term memory.

Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024b. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *Preprint*, arXiv:2407.06886.

Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. 2023. Fmb: a functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, page 02783649241276017.

Dong Ma, Ye Li, Teng Ma, and António M Pascoal. 2025a. The state of the art in key technologies for autonomous underwater vehicles: A review. *Engineering*.

Yunsheng Ma, Wenqian Ye, Can Cui, Haiming Zhang, Shuo Xing, Fucai Ke, Jinhong Wang, Chenglin Miao, Jintai Chen, Hamid Rezatofighi, and 1 others. 2025b. Position: Prospective of autonomous driving-multimodal llms world models embodied intelligence ai alignment and mamba. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1010–1026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. 2024. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*.

Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiao yong Wei, Shanru Lin, Hui Liu, Philip S. Yu, and Qing Li. 2025. A survey of webagents: Towards next-generation ai agents for web automation with large foundation models. *Preprint*, arXiv:2503.23350.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Easton Potokar, Spencer Ashford, Michael Kaess, and Joshua G Mangelson. 2022. Holoocean: An underwater robotics simulator. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3040–3046. IEEE.

Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. *arXiv preprint arXiv:2310.05861*.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502.

Yanyuan Qiao, Haodong Hong, Wenqi Lyu, Dong An, Siqi Zhang, Yutong Xie, Xinyu Wang, and Qi Wu. 2025. Navbench: Probing multimodal large language models for embodied navigation. *Preprint*, arXiv:2506.01031.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788. IEEE Computer Society.

Cagatay Sariman, Ahmed Hallawa, and Anke Schmeink. 2025. Ur-earl: A framework for designing underwater robots using evolutionary algorithm-driven reinforcement learning. *Ocean Engineering*, 321:120402.

Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, pages 621–635. Springer.

Haochen Shi, Zhiyuan Sun, Xingdi Yuan, Marc-Alexandre Côté, and Bang Liu. 2024. OPEx: A component-wise analysis of LLM-centric agents in embodied instruction following. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 622–636, Bangkok, Thailand. Association for Computational Linguistics.

Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21440–21455. Association for Computational Linguistics.

David Silver and Richard S Sutton. 2025. Welcome to the era of experience. *Google AI*, 1.

Jingyu Song, Haoyu Ma, Onur Bagoren, Advaith V. Sethuraman, Yiting Zhang, and Katherine A. Skinner. 2025. Oceansim: A gpu-accelerated underwater robot perception simulation framework. *Preprint*, arXiv:2503.01074.

Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. 2024. Towards long-horizon vision-language navigation: Platform, benchmark and method. *arXiv preprint arXiv:2412.09082*.

Xiaoyu Tan, Bin Li, Xihe Qiu, Chao Qu, Wei Chu, Yinghui Xu, and Yuan Qi. 2025. Meta-agent-workflow: Streamlining tool usage in llms through workflow construction, retrieval, and refinement. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 458–467, New York, NY, USA. Association for Computing Machinery.

Xiangru Tang, Tianrui Qin, Tianhao Peng, Ziyang Zhou, Daniel Shao, Tingting Du, Xinming Wei, Peng Xia, Fang Wu, He Zhu, Ge Zhang, Jiaheng Liu, Xingyao Wang, Sirui Hong, Chenglin Wu, Hao Cheng, Chi Wang, and Wangchunshu Zhou. 2025. Agent kb: Leveraging cross-domain experience for agentic problem solving. *Preprint*, arXiv:2507.06229.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9568–9578. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129(1):246–266.

Martin Visbeck. 2018. Ocean science research is key for a sustainable future. *Nature communications*, 9(1):690.

Chen Wang, Fei Xia, Wenhao Yu, Tingnan Zhang, Ruohan Zhang, C. Karen Liu, Li Fei-Fei, Jie Tan, and Jacky Liang. 2025a. Chain-of-modality: Learning manipulation programs from multimodal human videos with vision-language-models. *Preprint*, arXiv:2504.13351.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. 2024a. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *Preprint*, arXiv:2410.07087.

Xiangyu Wang, Donglin Yang, Ziqin Wang, Hohin Kwan, Jinyu Chen, Wenjun Wu, Hongsheng Li, Yue Liao, and Si Liu. 2024b. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *Preprint*, arXiv:2410.07087.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018. Building generalizable agents with a realistic and rich 3d environment.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 9 others. 2025. The rise and potential of large language model based agents: a survey. *Sci. China Inf. Sci.*, 68(2).

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, and 1 others. 2020. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11097–11107.

Kaibing Xie, Jian Yang, and Kang Qiu. 2022. A dataset with multibeam forward-looking sonar for underwater object detection. *CoRR*, abs/2212.00352.

Wei Xu, Cheng Wang, Dingkang Liang, Zongchuang Zhao, Xingyu Jiang, Peng Zhang, and Xiang Bai. 2025. Nautilus: A large multimodal model for underwater scene understanding. *Preprint*, arXiv:2510.27481.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. 2025b. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *Preprint*, arXiv:2502.09560.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. 2025c. Mmsi-bench: A benchmark for multi-image spatial intelligence. *Preprint*, arXiv:2505.23764.

Fanglong Yao, Yuanchang Yue, Youzhi Liu, Xian Sun, and Kun Fu. 2024. Aeroverse: Uav-agent benchmark suite for simulating, pre-training, finetuning, and evaluating aerospace embodied world models. *CoRR*, abs/2408.15511.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2025. Minicpm-v: A gpt-4v level mllm on your phone. *Nat Commun 16, 5509 (2025)*.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025a. What, how, where, and how well? A survey on test-time scaling in large language models. *CoRR*, abs/2503.24235.

Yifan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. 2025b. Mme-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Baining Zhao, Jianjie Fang, Zichao Dai, Ziyou Wang, Jirong Zha, Weichen Zhang, Chen Gao, Yue Wang, Jinqiang Cui, Xinlei Chen, and Yong Li. 2025. Urbanvideo-bench: Benchmarking vision-language models on embodied intelligence with video data in urban spaces. *Preprint*, arXiv:2503.06157.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. 2022a. Vlmbench: A compositional benchmark for vision-and-language manipulation. *Advances in Neural Information Processing Systems*, 35:665–678.

Kaizhi Zheng, Xiaotong Chen, Odest Chadwicke Jenkins, and Xin Wang. 2022b. Vlmbench: A compositional benchmark for vision-and-language manipulation. *Advances in Neural Information Processing Systems*, 35:665–678.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *CoRR*, abs/2408.09429.

Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. 2023. Marinegpt: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Gengze Zhou, Yicong Hong, and Qi Wu. 2024a. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7641–7649.

Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, Hongxin Zhang, Yilun Du, Joshua B. Tenenbaum, and Chuang Gan. 2024b. HAZARD challenge: Embodied decision making in dynamically changing environments. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

King Zhu, Hanhao Li, Siwei Wu, Tianshun Xing, Dehua Ma, Xiangru Tang, Minghao Liu, Jian Yang, Jiaheng Liu, Yuchen Eleanor Jiang, Changwang Zhang, Chenghua Lin, Jun Wang, Ge Zhang, and Wangchunshu Zhou. 2025. Scaling test-time compute for LLM agents. *CoRR*, abs/2506.12928.

Appendix

A The Use of Large Language Models (LLMs)

We confirm that LLMs are used only as an auxiliary tool to assist in refining wording and sentence structure. Their application in experiments is strictly confined to scientific research purposes, and all such uses have been clearly documented in the Experimental Settings. No additional reliance on LLMs has been involved in this work.

B More Related Works

Embodied Simulations. Embodied intelligence describes artificial intelligence systems whose intelligent behavior emerges through continuous physical and sensory interactions with the environment (Gupta et al., 2021; Ding et al., 2024; Shi et al., 2024). Simulation platforms are essential for advancing such systems across ground, aerial, and marine domains (Liu et al., 2024b; Han et al., 2025; Aldhaheri et al., 2025). In ground applications, platforms like Matterport3D (Chang et al., 2017), House3D (Wu et al., 2018), and Habitat (Manolis Savva et al., 2019) provide realistic indoor and outdoor environments for navigation, scene understanding, and human-robot interaction research. Aerial robotics benefits from simulators such as AirSim (Shah et al., 2018), CityNav (Lee et al., 2025) and OpenUAV (Wang et al., 2024a), which offer high-fidelity simulations with accurate physics and sensor models. Similarly, in the marine domain, simulation platforms such as HoloOcean (Potokar et al., 2022), OceanSim (Song et al., 2025), and MarineGym (Chu et al., 2025) provide specialized capabilities for modeling hydrodynamic effects and underwater dynamics. With the development of embodied intelligence, an increasing variety of simulation environments (Kolve et al., 2017; Puig et al., 2018; Xiang et al., 2020; Gan et al., 2021; Li et al., 2021; Nasiriany et al., 2024; Zhou et al., 2024b; Hong et al., 2025) have emerged to meet specific tasks, needs, or scenarios.

MLLM-driven Embodied Agents. Building upon the rapid advancement of LLMs (Achiam et al., 2023; Touvron et al., 2023; Chiang et al., 2023; Yang et al., 2025a), the emergence of MLLMs (OpenAI, 2024; Bai et al., 2025; Meta AI, 2024; Liu et al., 2024a; Gemini Team, 2024; Team et al., 2025; Wang et al., 2025b) has further strengthened agent capabilities by incorporating visual understanding for multimodal perception. Despite impressive results in various agent applications (Hu et al., 2024b; Ning et al., 2025), MLLM-driven agents still face substantial challenges in real-world and simulated embodied environments. Key difficulties persist in spatial cognition (Prasad et al., 2023; Du et al., 2024; Tong et al., 2024; Shiri et al., 2024; Zheng et al., 2024; Dang et al., 2025; Yang et al., 2025c; Li et al., 2025), task planning (Chen et al., 2023; Huang et al., 2023; Zhou et al., 2024a) , object navigation (Wang et al., 2024b; Khanna et al., 2024; Guo et al., 2025; Qiao et al., 2025; Cheng et al., 2025), and robotic manipulation (Zheng et al., 2022a; Yang et al., 2025b; Wang et al., 2025a). To evaluate agent capabilities, embodied benchmarks have been developed across diverse settings, including indoor (Anderson et al., 2018; Wu et al., 2018), urban (Chen et al., 2019; Caesar et al., 2020; Vasudevan et al., 2021; Gao et al., 2024), aerial (Yao et al., 2024; Gao et al., 2025b; Cai et al., 2025), specialised (Zheng et al., 2022b; Luo et al., 2023; Song et al., 2024; Li et al., 2024a) and real-world (Zhao et al., 2025; Koh et al., 2024; Zhang et al., 2025b) scenarios.

C Applications of OceanGym and Future Directions

Competitive Arena for Foundation Models. OceanGym provides a rigorous testbed for evaluating foundation models and embodied agent frame-

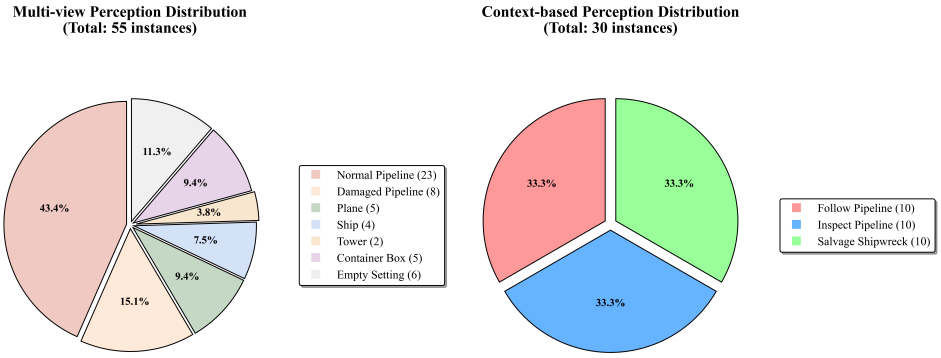

Figure 8: Statistics of perception tasks.

works, with a particular focus on memory mechanisms. Future work can leverage OceanGym to optimize prompt design, improve memory utilization, enhance base model capabilities, and advance research in task planning and sub-task decomposition for complex underwater missions.

Underwater Data Synthesis Platform. The framework enables the generation of synthetic underwater simulation data to improve both perception and decision-making capabilities of autonomous agents. This data can be effectively used for training models, and given the inherent difficulties in acquiring real-world underwater data, serves as a valuable complement to limited real-world datasets.

Reinforcement Learning Testbed. OceanGym offers a rich environment for reinforcement learning research, providing comprehensive feedback signals for training autonomous behaviors. Researchers can leverage the platform's diverse reward mechanisms, customizable environmental configurations, and real-time performance metrics to systematically evaluate and benchmark Reinforcement Learning algorithms under varying underwater conditions.

Sim-to-Real Bridge. The platform represents an initial step toward enabling the transfer of trained models to real-world AUVs, serving as a promising link between simulation and deployment. As illustrated in Figure 9, YOLO models trained in OceanGym's simulated environment demonstrate measurable performance improvements when tested on real sonar data, though generalization remains limited for objects not included in the simulation. Furthermore, Figure 10 shows that by aligning model outputs with standardized robotic operating commands, the same prompts and environment memory used in simulation can directly guide real AUVs to follow language instructions and locate target points. Through this early exploration of connecting virtual testing with real-world deployment, OceanGym offers a foundation for reducing reliance on expensive and hazardous field trials, accelerating development cycles, and enhancing the reliability and robustness of autonomous underwater systems.

D Perception Task Statistics

Figure 8 presents the statistical distribution of different perception settings analyzed in our dataset. The dataset consists of 85 sets of data, which include 55 sets focusing on Multi-view Perception and 30 sets on Context-based Perception. Within the Multi-view Perception data, 55 sets are categorized as follows: 23 sets involve normal pipelines, 8 sets entail damaged pipelines, 5 sets are related to planes, 4 sets concern ships, 2 sets focus on towers, 5 sets involve container boxes, and 6 sets do not feature any specific dominant object. For the Context-based Perception data, the 30 sets are evenly divided among three distinct sub-tasks, each comprising 10 sets. These sub-tasks involve the agent following pipelines, inspecting pipelines for potential damage, and scanning around shipwrecks.

E Decision Task Details

Decision-making tasks require an embodied agent to accomplish a given objective through a series of decisions. Every objective built using intricate 3D assets based on real-world references, ensuring accurate representation of structural and material characteristics. Figure 11 illustrates the perceptual input at one specific state during such a task.

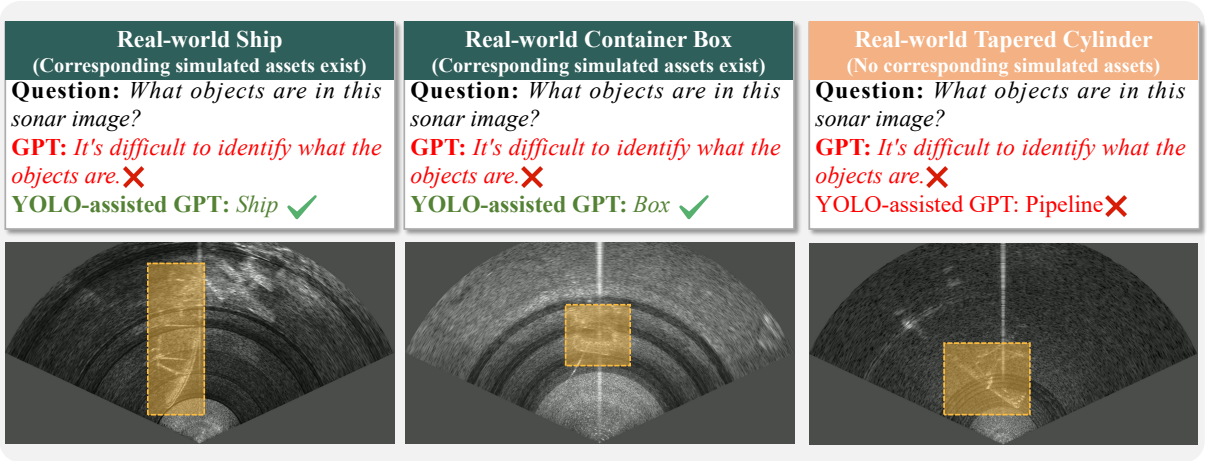

Figure 9: We evaluate whether YOLO models trained in simulated environments can enhance real-world performance by testing them on actual sonar data. The results demonstrate that while the YOLO-assisted GPT-4o-mini approach yields measurable performance improvements for certain objects modeled in OceanGym, the models exhibit limited generalization capability for objects not included in the simulation.

Figure 10: By matching the model's output actions with real operation commands, such as those standardized in the MOOS-IvP or ROS, the same prompts used in the simulated environment can directly control real AUVs. Moreover, using environment memory that mirrors reality as input enables the model to rapidly follow language instructions to reach or locate target points. In the picture, using simulated environment prompts and memory to help the real AUV find 3 target points.

Locate the robot. Locate and approach the mining robot in a complex underwater environment within an abandoned subsea research zone characterized by variable terrain, low visibility, and artificial structures. The operational protocol mandates an initial memory check for the target's coordinates; if available, the system engages in direct coordinate-based navigation. Absent prior data, the robot utilizes its six camera feeds for visual comparison against a reference image, identifying the target by its distinct shape, structure, and color. A systematic exploration pattern, such as a grid or linear search, is then executed. Throughout the mission, all encountered special objects and artificial structures are documented. Maintaining a strict minimum standoff distance of 10 meters from all rocks and obstacles is the highest priority, superseding all other actions. The vehicle must remain within the predefined operational boundaries at all times, and all reports must exclusively detail artificial structures, ignoring any marine life.

Inspect the oil pipe. Locate and identify the abandoned subsea oil pipeline network situated in a central zone where pipelines may be partially buried and serve as potential navigation references. The procedure begins with a query of the robot's memory for known pipeline coordinates, initiating direct navigation if the data is present. Without prior coordinates, the robot employs its camera feeds to detect linear structures and surface features that match the reference imagery of a pipeline. This is followed by a systematic exploration of the area to comprehensively document all artificial structures and special objects. A critical safety requirement is to maintain a safe distance from all obstacles, executing immediate directional changes upon hazard detection. All reporting must focus solely on artificial structures, with biological entities entirely omitted from logs.

Locate oil drumss. Locate and identify oil drums or barrels submerged in an environment where they may be partially buried or scattered within sediment under conditions of poor visibility. The first action is a memory scan for stored coordinates of oil drums, proceeding with direct waypoint navigation if the search is successful. If no coordinates exist, the robot must use its camera systems to identify cylindrical objects and any visible mark-

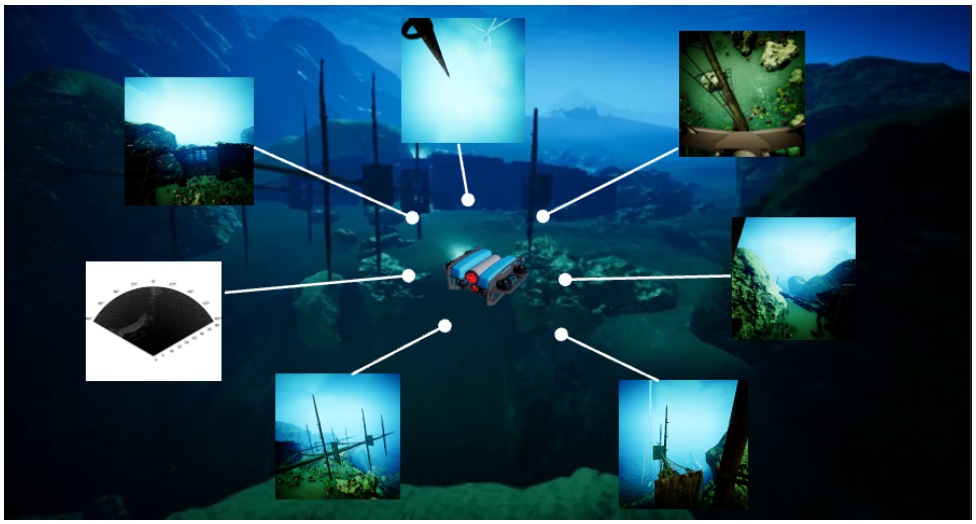

Figure 11: A state case of a decision-making task.

Figure 12: Target object for the “Locate the robot” task.

Figure 13: Target object for the “Inspect the oil pipe” task.

Figure 14: Target object for the “Locate oil drums” task.

Figure 15: Target object for the “Search for a sunken ship” task.

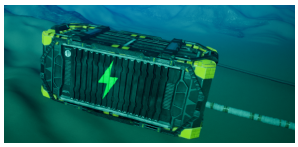

Figure 16: Target object for the “Locate the electrical box” task.

Figure 17: Target object for the “Inspect the wind turbine” task.

Figure 18: Target object for the “Search for the aircraft” task.

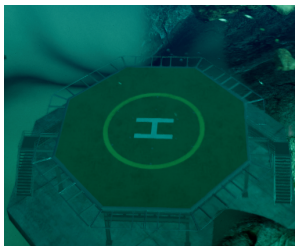

Figure 19: Target object for the “Docking” task.

ings that align with the target description. A methodical search pattern is then conducted across the operational area, with all special objects documented. Strict obstacle avoidance protocols are continuously enforced, and the robot’s trajectory must never exceed the designated operational boundaries. Reports are confined to artificial structures and special objects only.

Search for a sunken ship. Locate and identify sunken shipwrecks, which are typically structurally complex entities that may be partially buried or obscured by various underwater obstacles. The mission initiates with an access of the robot’s memory for any known coordinates of shipwrecks, utilizing them for direct navigation if available. In the absence of positional data, the robot relies on its camera feeds to recognize large structural features and surface details that correspond to the reference images of a shipwreck. A systematic exploration is subsequently performed to document all special objects within the area. A safe distance from all obstacles must be maintained throughout the operation, and the vehicle is required to stay within its prescribed operational limits. All marine life is systematically ignored and excluded from reporting.

Locate the electrical box. Locate and identify underwater electrical boxes, which are often partially buried in sediment and possess distinctive structural features. The operational sequence starts with a retrieval attempt from the robot’s memory

for the coordinates of electrical boxes, followed by direct navigation to any located waypoints. Without prior coordinate data, the robot must analyze its camera feeds to identify the target based on its specific shape, structural characteristics, and any identifiable markings. A thorough and systematic exploration of the zone is then carried out, with all special objects recorded. The mission must adhere to strict obstacle avoidance procedures and remain within the defined operational boundaries at all times. All communications and reports are restricted to artificial structures and special objects.

Inspect the wind turbine. Locate and identify underwater wind power station structures, which are large installations featuring multiple pillars and mechanical components. The robot first searches its internal memory for stored coordinates of the wind power station, navigating directly to the location if the data is found. If the coordinates are not located, the system uses its camera arrays to identify the major structural and mechanical elements that match the reference documentation. A systematic exploration pattern is executed to document every special object in the vicinity. A safe buffer distance from all obstacles is perpetually maintained, and the robot's path must comply strictly with the operational boundaries.

Search for the aircraft. Locate and identify underwater aircraft wreckage, which can be complex and potentially dispersed across different areas of the seafloor. The initial phase involves a memory check for any existing coordinates related to aircraft wreckage, with immediate navigation initiated upon a successful find. If no data is available, the robot switches to using its visual feeds to identify key structural features and surface details that are consistent with the target wreckage. A comprehensive systematic search is then conducted, ensuring all special objects are documented. Strict obstacle avoidance is paramount, and the vehicle must operate entirely within the set boundaries. Reports are exclusively to contain information on artificial structures and special objects.

Docking. Locate and identify an underwater landing platform marked with a distinctive "H" symbol, a structure with a regular form that provides a reliable navigation reference. The robot's first action is to consult its memory for the platform's coordinates, proceeding with direct navigation if the information is available. Should the coordinates be absent, the platform must be identified visually through the camera feeds by recognizing the "H" marking and the overall platform structure. This is followed by a systematic exploration to document all special objects in the area. A safe distance from all obstacles must be maintained, and the operation is confined to the approved boundaries. All reporting is limited to artificial structures and special objects, with no mention of biological activity.

F Utilization of Unreal Engine Plugins

Our simulation environment is compatible with Unreal Engine plugins. As shown in Figure 20, we leverage several UE plugins to simulate water interaction effects.

G The Impact of Different Prompts on Perception Tasks

Due to the difficulty in finding a prompt that is suitable for all MLLMs, we test the impact of different prompts on the model. As shown in Table 3, we find that the impact was relatively small in deep water environment. Prompt1 is the prompt used in the main experiment, and prompt2 is the best prompt for GPT-4o-mini during the testing process.

H Prompt for OceanGym

Figures 21–23 show the perception prompts, while Figure 24 details the navigation prompt.

Table 3: Performance of perception tasks across different prompts.

Model	Shallow Water Environment (High Illumination)					Deep Water Environment (Low Illumination)				
	Multi-View Perception		Context-based Perception		Avg	Multi-View Perception		Context-based Perception		Avg
	Vision	+Sonar	Vision	+Sonar		Vision	+Sonar	Vision	+Sonar	
GPT-4o-mini(prompt1)	34.55	34.55	20.00	33.33	30.61	14.55	20.00	3.33	6.67	11.14
GPT-4o-mini(prompt2)	54.55	45.45	40.00	30.00	42.5	20.00	20.00	10.00	0.00	12.5

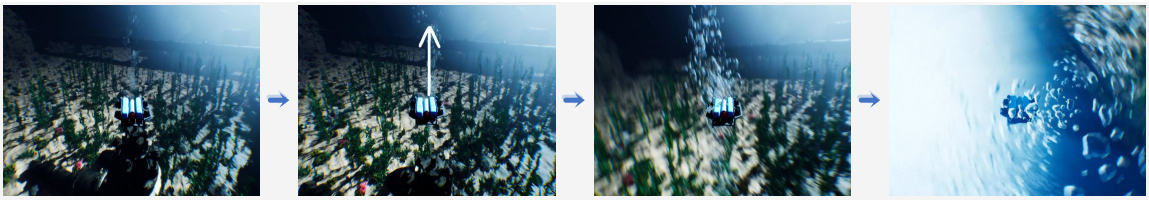

Figure 20: Examples of water flow interaction simulated using Unreal Engine plugins. An upward water flow is configured in the environment, exerting influence on an AUV as it passes through the affected region.

Prompt for Perception Tasks

[RGB Image]
You are an assistant that analyzes an image and checks which of the following options appear in it.
Options:**[Options]**
Instructions:
- Carefully examine the image, even the corners.
- You can choose single or multiple options, if none of the options appear, just return an empty list.
- For multiple-choice questions, no points will be awarded for incomplete selections, over-selections, or incorrect selections.
- The output must be a valid list (only list, no explanation, no extra text).

Figure 21: Prompt template for perception tasks without sonar data.

Prompt for Perception Tasks (Add Sonar)

[Sonar Image]
This sonar image can be used as a reference to assist in identifying the next color image.

[RGB Image]
You are an assistant that analyzes an image and checks which of the following options appear in it. Before that, I have already provide you a sonar image to help you choose the correct one.
Options:**[Options]**
Instructions:
- Only when you find it difficult to recognize the color image, I suggest you refer to the previous sonar image together.
- Carefully examine the image, even the corners.
- You can choose single or multiple options, if none of the options appear, just return an empty list.
- For multiple-choice questions, no points will be awarded for incomplete selections, over-selections, or incorrect selections.
- The output must be a valid list (only list, no explanation, no extra text).

Figure 22: Prompt template for perception tasks with sonar reference.

Prompt for Perception Tasks (Add Sonar and Examples)

[Object A Sonar Image]
This sonar image example is **[Object A]**.

[Object B Sonar Image]
This sonar image example is **[Object B]**.

...

[Sonar Image]
This sonar image can be used as a reference to assist in identifying the next color image.

[RGB Image]
You are an assistant that analyzes an image and checks which of the following options appear in it. Before that, I have already provide you a sonar image to help you choose the correct one.
Options:**[Options]**
Instructions:
- Only when you find it difficult to recognize the color image, I suggest you refer to the previous sonar image together.
- Carefully examine the image, even the corners.
- You can choose single or multiple options, if none of the options appear, just return an empty list.
- For multiple-choice questions, no points will be awarded for incomplete selections, over-selections, or incorrect selections.
- The output must be a valid list (only list, no explanation, no extra text).

Figure 23: Prompt template for perception tasks with sonar examples.

Prompt for Navigation Tasks

You are an expert pilot for an Autonomous Underwater Vehicle (AUV), designated as the "Control Expert". Your mission is to navigate a complex underwater environment to complete specific tasks. You will receive data from six cameras and location sensors. Your decisions must be precise, safe, and strategic.
1. Tactical Briefing for the Area of Operations
Before the mission begins, you must internalize the following intelligence about the operational area. This context is vital for interpreting sensor data and forming a macro-level strategy.

...

3. Mission Briefing and Sensor Data
Task Description: **[Task Description]**
Target Object Name: **[Object Name]**
Target Object Reference Image: **[Object Image]**
Target Object Description: **[Object Description]**

...

5. Survey Navigation Commands
Available Commands: ‘ascend‘, ‘descend‘, ‘move left‘, ‘move right‘, ‘move forward‘, ‘move backward‘, ‘rotate left‘, ‘rotate right‘, ‘stop‘.
Command Execution: You must only issue ONE command per turn from the list above.

...

Remember:
Conduct comprehensive reconnaissance! Systematic coverage = priority! Use efficient exploration patterns! Catalog all special objects! Maintain exploration momentum! Always use format! Ignore all marine life! One continuous line between markers!

Figure 24: Prompt template for navigation tasks.