360CityArena: A Realistic Virtual Urban Navigation Benchmark for Embodied Agents

Kenta Watanabe[1], Atsuyuki Miyai[1], Mizuki Takenawa[1], Kiyoharu Aizawa[1], and Toshihiko Yamasaki[1]

The University of Tokyo, Tokyo, Japan
{k_watanabe,miyai,yamasaki}@cvm.t.u-tokyo.ac.jp
{takenawa,aizawa}@hal.t.u-tokyo.ac.jp
https://360mm-team.github.io/360CityArena/

Abstract. We present 360CityArena, a benchmark for evaluating the urban exploration capabilities of Embodied Agents within a photorealistic environment constructed from 360° videos. Existing outdoor benchmarks either lack sufficient photorealism or complexity, resulting in a considerable gap from real-world urban environments. 360CityArena is built on a realistic reconstruction of the Akihabara district in Tokyo, Japan, using 602 360° video segments covering 85 streets, and consists of 175 meticulously human-crafted tasks. It encompasses three task categories: Environment Understanding, Path Reasoning, and Spatial Reasoning, covering fundamental abilities required for urban exploration, such as localization, landmark search, path planning, and relational spatial reasoning, thereby enabling comprehensive evaluation in realistic urban scenes. Our evaluation using state-of-the-art LMM-based agents shows that even the strongest model, Gemini-2.5 Flash, performs far below human level (human: 77.3% vs. Gemini-2.5 Flash: 17.1%), revealing substantial challenges that remain in city-scale embodied navigation and reasoning. 360CityArena provides a necessary and challenging testbed for photorealistic urban-district navigation and spatial reasoning.

Keywords: Embodied AI · Urban Navigation Benchmark · Photorealistic Virtual Environments

1 Introduction

Imagine a future where an AI assistant is a natural part of city life. It can help you when you get lost, guide blind people to where they need to go, and provide support whenever someone needs it. This kind of assistant would make cities more accessible, reduce the difficulties of traveling, and help create a more inclusive society. One promising approach toward achieving this vision is Embodied AI, which performs complex tasks based on perception of its surrounding environment [1, 8, 10, 26]. Realizing this vision will require major advances in evaluating and testing such systems across a wide range of tasks within realistic urban environments that capture the visual complexity of real streets.

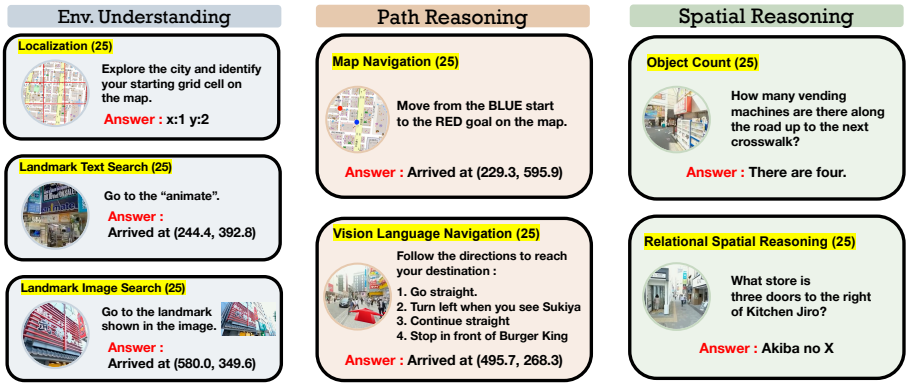

Fig. 1: Examples in each task type in 360CityArena. (i) Environment Understanding tasks include *Localization*, *Landmark Search with Language*, and *Landmark Search with Image*, where the agent infers its location or navigates to a specified landmark. (ii) Path Reasoning tasks evaluate the agent's ability to plan and execute routes, such as following map-based paths or vision-language navigation. (iii) Spatial Reasoning tasks assess relational understanding and quantitative perception, including identifying spatial relations between landmarks and counting objects in the environment.

A central problem in city exploration research is the lack of benchmarks that evaluate agents in environments that realistically reflect wide-area navigation in real urban spaces. Existing 3D simulators fall short in reconstructing city scenes with sufficient photorealism and complexity [9, 12, 17, 44]. Although environments built from Google Street View provide realistic imagery [6, 11, 30, 49], they lack dynamic elements and do not offer continuous, fully navigable spaces. To address these issues, recent studies have explored converting real-world urban videos into simulation environments, enabling more realistic and interactive experiences [45]. However, the video clips used in this method are short, making them unsuitable for exploring extended real street networks. In parallel, recent work studies navigation from different modalities and viewpoints, such as aerial VLN over cities and map-only evaluation of route planning [22, 34, 46]. These lines of work are complementary, but they still do not capture ground-level, egocentric city exploration in realistic street scenes with rich dynamics. To further advance the field, we need a benchmark that covers tasks within highly realistic and dynamic urban street environments, while supporting smooth egocentric navigation.

In this paper, we propose **360CityArena**, a benchmark for assessing Embodied Agents in a photorealistic real-world urban district represented as an interconnected pose graph of 360° video trajectories. 360CityArena is built upon a Realistic Virtual World environment [41] that recreates Akihabara, a major urban district in Tokyo, and includes 175 manually created tasks. The environment spans approximately 750 meters north–south and 650 meters east–west, covering 85 interconnected streets. Each task is grounded in real city structures

and landmarks, encouraging agents to perform practical, exploration-based activities.

As shown in Figure 1, 360CityArena consists of three task types: (i) *Environment Understanding*, (ii) *Path Reasoning*, and (iii) *Spatial Reasoning*. (i) *Environment Understanding* targets visual recognition and self-localization, measuring an agent's perceptual understanding of the environment. (ii) *Path Reasoning* focuses on path planning and decision-making ability within the real spatial layout of the environment. (iii) *Spatial Reasoning* evaluates geospatial reasoning through navigation. At a finer level, the benchmark systematically measures seven distinct capabilities. Furthermore, each task is categorized into Easy, Medium, and Hard, making it possible to evaluate agents according to difficulty level. Through these systematically constructed tasks, 360CityArena enables a comprehensive assessment of an Embodied Agent's ability to perceive, reason, and act in real-world urban environments.

Our experiments show that state-of-the-art LMM-based agents remain far below human performance across all tasks, especially on Map Navigation, Object Count, and Relational Spatial Reasoning. This demonstrates both the value of 360CityArena as a comprehensive benchmark for realistic urban settings and the need for stronger city-scale agents. We also evaluate removing explicit self-location information (*e.g.* a map). Performance does not degrade consistently and even improves in some tasks, suggesting that the manner of providing location priors can affect exploration strategies and failure modes, informing future model design.

The contributions of our paper are summarized as follows:

- **Proposing 360CityArena**: We introduce **360CityArena**, a photorealistic urban-district benchmark built on a photorealistic Virtual World reconstruction of Tokyo's Akihabara district from 360° videos. The benchmark provides seven task types across three categories: Environment Understanding, Path Reasoning, and Spatial Reasoning, enabling comprehensive evaluation of embodied agents in a realistic urban environment with an interconnected street network. Beyond evaluation, 360CityArena also serves as a practical guideline for constructing training data in photorealistic virtual worlds for embodied AI.
- **Benchmarking with Recent LMMs**: We build embodied agents based on several proprietary LMMs and open-source vision-language models, and benchmark them on 360CityArena. The results reveal substantial performance gaps compared to humans across all tasks and clear performance degradation as task difficulty increases.
- **Comprehensive Analysis**: Through extensive qualitative and quantitative analyses, we investigate how input modality (language versus images), the presence of self-location information, and model- and task-specific failure patterns influence agent behavior and performance. These analyses offer insights into the roles of visual cues and self-localization in urban embodied tasks.

Table 1: Comparison of urban navigation environments – realism, structural complexity, dynamics, interactivity, and exploration capability.

Environment	Category	Photo-realism	Structural complexity	Dynamics	Interaction	District-scale exploration	Motion
EmbodiedCity [12]	3D simulator	Low	Low	Medium	✓	✓	Continuous
MetaUrban [44]	3D simulator	Low	Medium	Medium	✓	✓	Continuous
CARLA [9]	3D simulator	Low	Low	Low	✓	✓	Continuous (clip)
Vid2Sim [45]	Video-to-sim	High	High	Medium	✓	✗	Continuous
StreetLearn [30]	GSV-based	High	High	Low	✗	✓	Discrete
360CityArena (Ours)	360° video	High	High	High	✗	✓	Continuous (trajectory)

2 Related Work

Embodied Agent. An Embodied Agent is an AI system that follows language instructions and completes tasks by perceiving the environment through camera views and sensor information. Recent advances in LMMs have greatly improved multimodal integration and reasoning, making them a strong foundation for embodied agents [15, 16, 24, 56]. Embodied-agent research spans real robots [19, 39, 52], indoor simulators [21, 37, 51], and outdoor simulators [6, 45]. While real robots are most realistic, data collection is costly and often environment-specific, limiting generalization. Simulation environments therefore attract attention for their diversity and reproducibility [10]. Indoor simulators ease data collection but have limited scale and simple layouts [37], motivating a shift toward larger, more complex outdoor settings with dynamic elements such as pedestrians and vehicles [30, 49].

Navigation Agents in Outdoor Environments. Recent advances in simulation environments have broadened embodied navigation research beyond indoor settings [21, 28, 35–37, 51] to large-scale outdoor environments. Outdoor navigation tasks are commonly distinguished by how the goal is specified: point-goal navigation, where the agent is given target coordinates [25, 30, 44]; image-goal navigation, where the target is provided as a reference image [17, 18]; object-goal navigation, where the goal is described in language (*e.g.* a landmark) [5, 14]; and Vision-Language Navigation (VLN), where the agent follows natural-language instructions [6, 23, 49]. Some work instead evaluates route planning directly on maps without embodied simulation [34, 46, 49]. In parallel, CityNav [22] studies real-world navigation from an *aerial* viewpoint, which is complementary but differs from *ground-level* exploration with egocentric street scenes. Separately, map understanding has been explored outside embodied navigation, including text-based map representations for grounding LLM planning (Tag Map [54]) and VLM benchmarks for advanced map queries (MAPWise [32]).

While point-goal, image-goal, and object-goal tasks emphasize efficient target reaching via visual or semantic cues, VLN additionally demands higher-level reasoning to resolve linguistic ambiguity and landmark references. However, aerial-VLN datasets and map-only evaluations do not jointly capture egocentric navigation and urban reasoning in realistic street environments. To address this gap, we introduce an urban-district benchmark that unifies these tasks with (i) photo-

realistic, dynamic street-level observations, (ii) multi-task diagnostics spanning environment understanding, path reasoning, and spatial reasoning, and (iii) controlled comparisons between language- and image-goal landmark search under matched conditions.

Real-world Simulation Environment. Prior work has explored 3D scene reconstruction, Google Street View–based environments, and 3D simulators for building real-world simulation settings. For 3D scene reconstruction, methods built on Neural Radiance Fields [29] and Gaussian Splatting [20] have led to numerous extensions [4, 47, 48, 53]. Despite these advancements, most approaches remain limited to producing static scenes and do not provide fully interactive environments [45]. City-scale reconstructions in conventional 3D simulators can procedurally generate diverse variations and support rich agent–environment interactions, but fall short in photorealism and structural complexity [12,17,27,44]. Environments constructed from Google Street View (GSV) enable diverse and highly realistic city-scale representations that are difficult to achieve with traditional 3D simulation [6,11,30,42,49]. However, these environments from GSV do not contain dynamic elements, and the inherent discontinuity between panoramas creates a gap from real-world continuous navigation [30]. To address this, a recent effort has attempted to convert real urban videos into interactive simulation environments [45]. While this approach is promising, this environment relies on short video clips, making them unsuitable for large-scale city exploration. Takenawa *et al.* [41] introduced a Realistic Virtual World (RVW) generated from large collections of 360° videos, providing a photorealistic and dynamic city-scale environment with real pedestrian and vehicle motions. We leverage this RVW to build a benchmark for realistic, city-scale exploration by embodied agents. While navigation within each filmed trajectory is continuous and smooth, discontinuities arise at trajectory boundaries, and the environment does not support physical interaction because it is constructed from pre-recorded video. We summarize key properties of representative real-world urban environments in Table 1.

3 The 360CityArena

We introduce 360CityArena, a novel benchmark meticulously designed to evaluate an agent's ability to explore a realistic urban district across a broad range of tasks (see Figure 2). The benchmark is built on a Realistic Virtual World reconstructed from the actual Akihabara district in Japan and consists of 175 tasks organized into three major categories and seven subcategories.

We first describe the Akihabara virtual environment in Section 3.1. We then provide detailed explanations of each task category in Section 3.2. The task construction process is outlined in Section 3.3, followed by the evaluation protocol in Section 3.4.

3.1 Akihabara Virtual Environment

The environment used in 360CityArena is a Realistic Virtual World of Akihabara constructed from 602 360° video segments over 85 streets [41]. The videos are

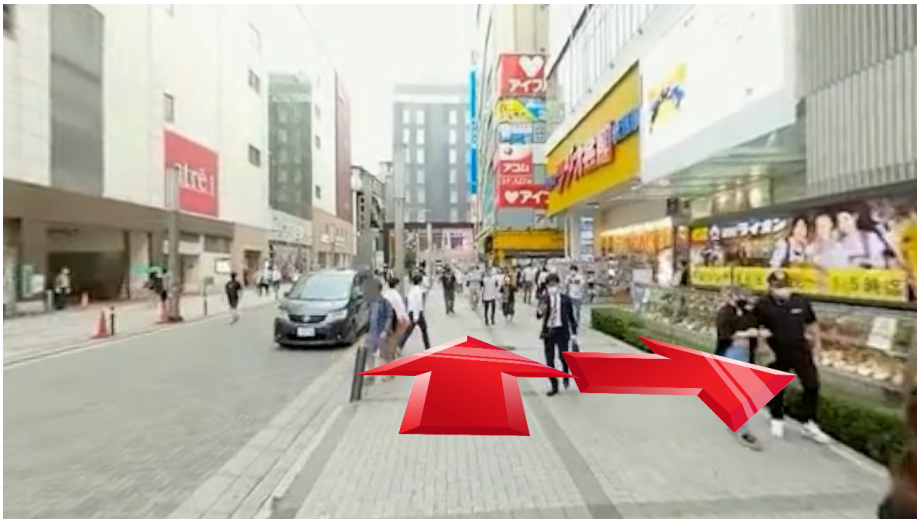

Fig. 2: Example of the visual observations in 360CityArena. The agent moves through the city while scanning its surroundings. At branching points, it must choose a route.

projected onto a spherical surface and organized into a navigable pose graph in Unity, enabling agents to traverse the captured area. Agents therefore navigate along prerecorded 360° video trajectories represented as a pose graph, rather than moving freely to arbitrary 3D positions or physically interacting with the environment. The resulting pose graph forms a single connected component with 193 nodes and 305 edges, and has a mean branching degree of 3.16.

In terms of geographic coverage, 360CityArena spans approximately 750 m north–south and 650 m east–west around Akihabara Station. Akihabara, originally an electronics district, has become a major commercial and cultural hub with dense electronics and anime/game-related retail and tourist attractions. The area features diverse streetscapes with narrow sidewalks, complex building layouts, abundant outdoor ads, and electronic signboards, creating a visually dense, information-rich scene. Each street is captured twice (one per direction), reflecting direction-dependent visual changes. These characteristics make Akihabara well suited for evaluating embodied agents' perception and navigation in a realistic urban simulation.

Although 360CityArena currently includes only one city, the representation is general. Following the Movie Map paradigm [40], similar city-scale environments can be built by geo-localizing 360° frames with GPS/SLAM and a map, identifying intersections, and connecting snippets into a traversable pose graph. Our city-agnostic graph augmented with 360° observations is directly compatible with such environments.

3.2 Task Categories

Our benchmark consists of three major categories and seven subcategories, with examples shown in Figure 1. Each subcategory contains 25 tasks. We describe each category in detail below.

1. Environment Understanding. This category evaluates an agent's ability to perceive and understand its surroundings, covering visual perception, memory, and language-grounded scene understanding. It includes three subcategories: *Localization*, *Landmark Search with Language* (object-goal), and *Landmark Search with Image* (image-goal). In *Localization*, the agent infers its initial position on a 5×5 grid from visual observations. Unlike GeoGuessr-style geolocalization, which predicts broad regions from scenes [13], our task operates at a much smaller scale and relies on town layout and local landmarks. The two Landmark Search tasks require navigating to a specified landmark, provided either in natural language or as an image; aside from the input modality, they share identical settings, enabling controlled comparison of modality effects.

2. Path Reasoning. This category evaluates an agent's ability to plan and explore based on spatial understanding. It measures not only the agent's planning and decision-making abilities, but also its ability to execute those plans in accordance with the spatial layout of the environment. The subcategories in this task are *Map Navigation* and *Vision-Language Navigation* (VLN). *Map Navigation* requires the agent to choose an optimal route using map information and navigate from a specified start point to a designated goal point. *Vision-Language Navigation* requires the agent to interpret and follow multi-step natural-language instructions and move to the target accordingly.

3. Spatial Reasoning. This category evaluates an agent's ability to understand the structural layout of objects and the positional relationships between landmarks in the environment. It measures core spatial perception and logical reasoning capabilities. The subcategories in this task are *Relational Spatial Reasoning* and *Object Count*. *Relational Spatial Reasoning* requires the agent to infer relative spatial relationships between landmarks. Given a reference landmark and a specified relation, the agent must identify the landmark that satisfies that relation. *Object Count* is a quantity-estimation task in which the agent must accurately count the number of specified objects within a designated area.

3.3 Benchmark Construction

Our construction involves eight annotators. We assigned two annotators to each of the seven sub-tasks. To ensure consistency in task quality across different tasks, one of the authors was assigned to all tasks. To ensure the quality of the benchmark, we selected annotators who had actually visited Akihabara in person. They interacted with the environment directly in Unity and created tasks that are verified to be solvable. The overall task construction process required approximately 60 hours.

Each task is labeled Easy, Medium, or Hard by annotators based on distance, instruction ambiguity, landmark visibility, and required exploration, rather than

fixed step-count thresholds. To support these labels, task-relevant statistics show monotonic increases from Easy to Hard: Map Navigation path length and decision points are 124/247/347 m and 3.3/6.4/9.0, Object Count ground-truth counts are 3.5/4.6/9.0, and VLN instruction steps are 4.4/5.9/7.4. Borderline cases may still involve subjectivity.

3.4 Evaluation Protocol

We define a set of evaluation criteria to assess agent performance. Our benchmark adopts four evaluation protocols, each designed to capture different aspects of task success. We detail each protocol and its corresponding task categories below.
1. exact_match. `exact_match` is applied to tasks whose outputs are explicitly defined as numerical values (grid coordinates) or textual strings [31, 55]. Under this metric, the agent's final textual output is evaluated, and the prediction is considered correct only when it exactly matches the answer text. In our benchmark, the *Localization* task uses this evaluation metric.
2. fuzzy_match. `fuzzy_match` leverages a language model (GPT-5 in our implementation) to assess whether the output is semantically equivalent to the ground truth [31, 55]. Importantly, the GPT-5 evaluator is fully isolated from the agent and only compares the final output with the ground-truth answer, preventing any information leakage. In the *Relational Spatial Reasoning* task, we employed this evaluation metric and additionally performed manual verification. The GPT-5 judge achieved 97.9% agreement with human majority votes ($\kappa = 0.937$), close to the inter-annotator agreement ($\kappa = 0.984$), supporting its use for evaluation automation.
3. coordinate_match. `coordinate_match` is used for tasks in which the agent must output its final position after moving through the environment. The coordinates are obtained directly from Unity. This metric is applied to *Map Navigation*, *VLN*, and *Landmark Search with Image / Language*. For all of these tasks, the agent's final position is evaluated by measuring the Euclidean distance to the target location, and the prediction is considered correct if it falls within a threshold of ε. For the distance threshold ε, we set $\varepsilon = 10\,\mathrm{m}$ for all tasks except *Map Navigation*, where a larger threshold of $\varepsilon = 20\,\mathrm{m}$ is used to account for noise caused by the marker's size on the map. A threshold-sensitivity check showed stable model rankings for 0.75×–1.5× thresholds (Kendall's $\tau \geq 0.89$), a human–best LMM gap above 36 pp even at $\varepsilon = 30\mathrm{m}$, and Map Navigation changes within ±2 pp for $\varepsilon = 15$–25 m.
4. mean_relative_accuracy (MRA). MRA is a flexible evaluation metric applied to tasks that involve numerical estimation [50]. In our benchmark, this corresponds to the *Object Count* task. MRA averages the relative accuracy across a range of confidence thresholds $\mathcal{C} = \{0.5, 0.55, \ldots, 0.95\}$:

$$\mathcal{MRA} = \frac{1}{10} \sum_{\theta \in \mathcal{C}} \mathbb{1} \left(\frac{|\hat{y} - y|}{y} < 1 - \theta \right), \tag{1}$$

where $\hat{y}$ denotes the predicted value and y the ground truth.

MRA accounts for the magnitude of error, so it provides a more appropriate evaluation than a simple binary correct/incorrect classification.

4 Navigation Agents

4.1 Problem Formulation

The environment and agent can be modeled as a partially observable Markov decision process (POMDP): $\mathcal{E} = (S, A, \Omega, T)$, where S represents the set of states, A represents the set of actions, Ω represents the set of observations. The transition function is defined as $T : S \times A \to S$, with deterministic transitions between states conditioned on actions. At each time step t, the environment is in some state s_t (*e.g.* a specific position and viewing direction). The agent receives a partial observation $o_t \in \Omega$, which consists of the visual input captured at time t and, optionally, the current positional information shown as an image with a marker indicating the agent's location and orientation on the map. The agent also maintains a memory buffer $M_t \in \mathcal{M}$ that stores important information from previous steps up to $t-1$. The agent then issues an action $a_t \in A$ conditioned on both o_t and the stored memory M_t, which results in a new state $s_{t+1} \in S$ and a new observation $o_{t+1} \in \Omega$ from the updated viewpoint. Simultaneously, relevant information from o_t and thoughts is written to the memory, updating it to M_{t+1}.

4.2 Baseline Agents

For our experiments, we evaluate multiple LMMs as baseline agents: GPT-5 [33], Claude Sonnet 4.5 (20250929) [2], Gemini 2.5 Flash [7], Qwen2.5-VL-32B-Instruct [3], and InternVL3.5 [43]. Among these models, GPT-5, Claude Sonnet 4.5 and Gemini-2.5 Flash are closed-source, while Qwen2.5-VL-32B-Instruct, InternVL3.5-8B, and InternVL3.5-38B are open-source. For inference with the open-source model, we used eight NVIDIA A100 80GB GPUs. In addition, all tasks were executed in Unity 6000.0.30f1 (Unity 6, 2023 LTS) running on macOS.

In our setting, the action space A consists of seven discrete actions: moving forward, tilting the viewpoint upward, tilting the viewpoint downward, rotating the viewpoint to the right, rotating the viewpoint to the left, resetting the viewpoint to align with the current heading direction, and outputting an answer. At branching points, the action space is augmented with movement actions corresponding to the available traversable directions, such as going forward, taking the right branch, or taking the left branch.

5 Experiment

5.1 Experimental Results

Human Performance. We conducted human evaluations after obtaining approval from our institution's Institutional Review Board (IRB). We recruited five

Table 2: Overall Results (%). Comparison of model and human performance across seven spatial and reasoning tasks, grouped into three major categories. Gemini-2.5 Flash achieves the highest overall performance, while the strongest model varies across individual tasks. All LMMs still fall far short of human performance.

	Environment Understanding			**Path Reasoning**		**Spatial Reasoning**	
	Loc	Landmark (Lang)	Landmark (Img)	Map Nav	VLN	Obj Count	Rel Reason
GPT-5	8.0	16.0	**48.0**	0.0	8.0	2.4	**32.0**
Claude Sonnet 4.5	4.0	4.0	16.0	**4.0**	4.0	10.8	8.0
Gemini-2.5 Flash	**12.0**	**28.0**	36.0	0.0	8.0	**24.0**	12.0
Qwen2.5-VL-32B-Instruct	4.0	16.0	20.0	0.0	0.0	18.8	4.0
InternVL3.5-8B	4.0	20.0	20.0	0.0	**12.0**	2.8	0.0
InternVL3.5-38B	0.0	16.0	0.0	0.0	**12.0**	7.2	4.0
Human	68.0	64.0	92.0	92.0	88.0	45.2	92.0

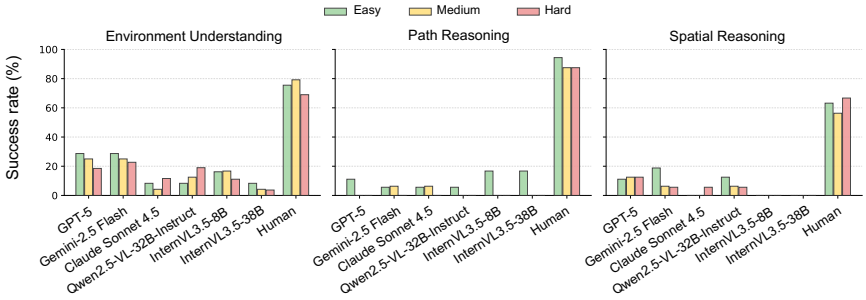

Fig. 3: Success rate by difficulty level across tasks and models (%). This graph shows model and human performance for each task, divided into three difficulty levels: Easy (E), Medium (M), and Hard (H).

participants, including both undergraduate and graduate students. We selected participants who had been to Akihabara before or who visit the area regularly. Since real-world deployment in a specific urban district can benefit from local familiarity, we report this result as a local-expert human baseline, which serves as an in-domain upper-bound reference rather than a generic human baseline.

As shown in Table 2, human participants achieved higher accuracy than current LMMs. For tasks such as *Map Navigation*, *Landmark Search with Image*, *Vision-Language Navigation*, and *Relational Spatial Reasoning*, humans reached around 90% accuracy. In contrast, performance on *Localization*, *Landmark Search with Language*, and *Object Count* was more modest, at 68%, 64%, and 45%, respectively. In Landmark Search, the task configurations for the Language and Image variants are identical except for the input modality. The fact that participants achieved much higher accuracy in the Image condition indicates that visual inputs contain substantially richer information than textual descriptions, including cues about appearance, location, and overall scene context.

Next, our main findings are as follows:

Table 3: Comparison with and without location information. We observe that the performance did not improve consistently across tasks; in some cases, accuracy decreased instead.

	Environment Understanding			Path Reasoning		Spatial Reasoning	
	Loc	Landmark (Lang)	Landmark (Img)	Map Nav	VLN	Obj Count	Rel Reason
GPT-5	8.0	16.0	48.0	0.0	8.0	2.4	32.0
GPT-5 (w/o location)	-	24.0	44.0	4.0	8.0	0.0	48.0

F1: LMMs perform far below human level. As shown in Table 2, all LMMs fall far short of human performance across the benchmark. Gemini-2.5 Flash achieves the highest overall performance among the evaluated models, while the strongest model varies across individual tasks (e.g., GPT-5 on *Landmark (Img)* and *Rel Reason*). These results indicate that, despite incremental progress, substantial gaps to human-level environment understanding and spatial reasoning remain.

F2: Image-based landmark search is consistently easier than language-based search. Across models, *Landmark Search with Image* generally outperforms *Landmark Search with Text* (e.g., GPT-5: 48.0 vs. 16.0; Claude: 16.0 vs. 4.0; Qwen: 20.0 vs. 16.0), suggesting that visual inputs provide concrete cues (appearance, textures, and surrounding context) that more directly support navigation than abstract textual descriptions. However, this improvement is not universal. InternVL does not show a clear gain from image inputs (8B: 20.0 vs. 20.0; 38B: 0.0 vs. 16.0), indicating limitations in effectively leveraging landmark images. By including both image- and language-based variants, our benchmark enables analysis of how models leverage different input modalities.

F3: Performance decreases as task difficulty increases. As shown in Figure 3, model performance generally declines with higher task difficulty. For example, GPT-5's accuracy in Environment Understanding drops from 28.0 $\rightarrow$ 25.0 $\rightarrow$ 18.5, and in Path Planning from 11.1 $\rightarrow$ 0.0 $\rightarrow$ 0.0 across the Easy, Medium, and Hard settings. This demonstrates that the benchmark allows meaningful comparison of model performance across different difficulty levels.

5.2 Analysis

Effect of Location Information. As shown in Table 3, we examined how explicitly providing the agent with its current location affects its spatial understanding and reasoning abilities, and found that adding a location prior yields inconsistent performance improvements across tasks. In Environment Understanding, Landmark (Language) accuracy dropped from 24.0% to 16.0%, while Landmark (Image) showed no meaningful change (within the margin of error). Path Reasoning tasks (Map Navigation and VLN) exhibited nearly no difference, suggesting static location information fails to aid path planning or instruction interpretation. In Spatial Reasoning, Object Count improved slightly,

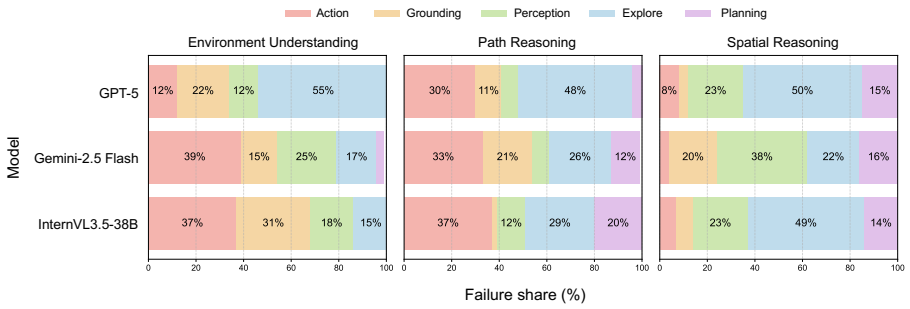

Fig. 4: Failure cause breakdown by task category across models (%). This figure shows the distribution of failure causes for each model across three task categories. Failures are categorized into five types: Action, Grounding, Perception, Explore, and Planning, and each stacked bar reports the percentage breakdown within the corresponding (model, task category) setting.

but Relational Reasoning fell from 48.0% to 32.0%. These performance degradations suggest agents struggle to align map-based location and relational information with real-world cues derived from visual inputs, rather than location information being inherently unhelpful. While prior work addresses map-based reasoning [34,38,46], the impact of the map-to-visual alignment process remains under-examined. Future research should independently evaluate this alignment capability and, where necessary, provide mechanisms for learning it.

Distinct Failure Patterns across Models. To better understand failure factors for three major models (GPT-5, Gemini-2.5-Flash, and InternVL3.5-38B), we used Gemini-3-Flash to analyze failures from viewpoint images, maps, and thought histories, assigning each case to one of five categories—Action, Grounding, Perception, Explore, and Planning—and then had humans verify the label and select the single dominant cause. Action denotes low-level action errors (*e.g.*, overshoot, wrong action, premature answer); Grounding, self-localization/orientation inconsistencies; Perception, object-recognition failures; Explore, stagnation or loops; and Planning, missed instructions or requirements. The resulting breakdown by model and task category is shown in Figure 4. These results should be interpreted as diagnosing integrated embodied urban performance, rather than isolating pure spatial reasoning: Action and Explore failures indicate that current LMM agents also suffer from controller and prompting brittleness when converting perception and reasoning into navigation decisions.

The results reveal distinct failure patterns reflecting model capabilities. First, GPT-5 contrasts sharply with other models: it shows minimal Action failures (12% in Environment Understanding) but dominant Explore failures (55%), indicating effective basic execution but poor exploration strategy. Conversely, Gemini-2.5-Flash and InternVL3.5-38B struggle primarily with Action failures (around 40%), pointing to deficits in low-level control. Second, Perception failures surge in Spatial Reasoning across all models (reaching 38% for Gemini),

t = 7. HottoMotto is on the right. t = 8. Look above.

t = 40. Look to the right. t = 41. Look below.

Fig. 5: Example of the agent's views for *Landmark Search with Language*, resulting in failure. The agent is instructed to search for "Jonathan". From t = 8 to t = 38, the gaze repeatedly moved up and down. At t = 39, the gaze briefly shifted to the right, but from t = 41 onward, it returned to the same up–down movement.

confirming that overlooking minute visual details is fatal for spatial tasks. Furthermore, Gemini displays consistent Grounding errors (15–21%) across categories, indicating persistent difficulty in aligning map data with the first-person perspective.

Case Study of Successes and Failures. For Landmark Search tasks in which the Language variant fails but the Image variant succeeds, we analyze individual cases using GPT-5, examining the agent's visual observations. The example shown here is a task in which the agent must locate Jonathan's restaurant.

In the *Landmark Search with Language* failure case in Figure 5, the agent passes by Hotto Motto and hypothesizes that Jonathan is located above it, prompting it to look upward. After failing to find the target, it shifts its gaze right and continues scanning. As a result, the agent stays in the same location, repeatedly changing viewpoints until the episode ends due to the step limit.

In contrast, as shown in Figure 6, in *Landmark Search with Image*, the target image provides cues such as its street-corner location and distinctive blue-and-white vertical facade with a red sign. The agent therefore disregards Hotto Motto, moves toward the major intersection, and successfully finds the target building.


t = 4. HottoMotto is on the right.


t = 5. Go past.


t = 16. Spot blue and white vertical stripes.


t = 29. Arrive at the destination.

Fig. 6: Example of the agent's views for *Landmark Search with Image*, resulting in success. The agent is instructed to search for "Jonathan" given in image form.

6 Limitations and Future Works

Restricted Exploration and Limited Interaction. Because 360CityArena is constructed from pre-recorded 360° videos, agents can explore only along captured trajectories rather than move freely to arbitrary locations. As a result, transitions across trajectory boundaries may introduce discontinuities that would not arise in fully interactive 3D simulators or in the real world. In our logs, boundary-transition actions accounted for 11.3% of all actions and were not overrepresented among Action or Explore failures; manual inspection also did not reveal clear discontinuity-induced failures. Nonetheless, we argue that situated visual navigation and spatial reasoning in a highly photorealistic urban setting remain challenging and important for current LMM-based agents. Extending the environment to support freer exploration and richer interaction is an important direction for future work.

Limited Urban Diversity. A limitation of the current benchmark is its restriction to a single urban district, which may limit urban diversity and conclusions regarding cross-regional generalization. However, we consider that 360CityArena provides a uniquely photorealistic, large-scale setting that facilitates more realistic urban evaluations than existing benchmarks. We leave cross-regional generalization as an important direction for future work.

7 Conclusions

We present 360CityArena, a benchmark designed to evaluate the urban exploration capabilities of Embodied Agents within photorealistic environments constructed from 360-degree videos. Our experimental results reveal that state-of-the-art LMM-based agents still face significant challenges in understanding and reasoning within city environments. We believe that providing a benchmark closer to real-world conditions will further advance the development of Embodied Agents that are applicable to real urban settings.

Acknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 25H01164, SIP Smart Disaster Prevention, JST BOOST Grant Number JPMJBS2418, and JST ASPIRE Program Grant No. JPMJAP2303. This study was conducted with approval from the relevant Ethics Committee (Approval No. UT-IST-RE-250508_6).

References

1. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
2. Anthropic: Claude Sonnet 4.5 system card. Tech. rep., Anthropic, PBC (2025)
3. 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. arXiv preprint arXiv:2502.13923 (2025)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In: CVPR. pp. 5470–5479 (2022)
5. Brahmbhatt, S., Hays, J.: DeepNav: Learning to navigate large cities. In: CVPR. pp. 5193–5202 (2017)
6. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments. In: CVPR. pp. 12538–12547 (2019)
7. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
8. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: CVPR. pp. 1–10 (2018)
9. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: CoRL. vol. 78, pp. 1–16 (2017)
10. Duan, J., Yu, S., Tan, H.L., Zhu, H., Tan, C.: A survey of embodied AI: From simulators to research tasks. TETCI **6**, 230–244 (2022)
11. Feng, J., Zhang, J., Liu, T., Zhang, X., Ouyang, T., Yan, J., Du, Y., Guo, S., Li, Y.: CityBench: Evaluating the capabilities of large language models for urban tasks. In: KDD. pp. 5413–5424 (2025)
12. Gao, C., Zhao, B., Zhang, W., Mao, J., Zhang, J., Zheng, Z., Man, F., Fang, J., Zhou, Z., Cui, J., et al.: EmbodiedCity: A benchmark platform for embodied agent in real-world city environment. arXiv preprint arXiv:2410.09604 (2024)
13. Haas, L., Skreta, M., Alberti, S., Finn, C.: PIGEON: Predicting image geolocations. In: CVPR. pp. 12893–12902 (2024)
14. Hong, Y., Sun, R., Li, B., Yao, X., Wu, M., Chien, A., Yin, D., Wu, Y.N., Wang, Z., Chang, K.W.: Embodied web agents: Bridging physical-digital realms for integrated agent intelligence. In: NeurIPS. vol. 38 (2025)
15. Huang, W., Abbeel, P., Pathak, D., Mordatch, I.: Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: ICML. vol. 162, pp. 9118–9147 (2022)
16. Ichter, B., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., Kalashnikov, D., Levine, S., Lu, Y., Parada, C., Rao, K., Sermanet, P., Toshev, A.T., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Yan, M., Brown, N., Ahn, M., Cortes, O., Sievers, N., Tan, C., Xu, S., Reyes, D., Rettinghouse, J., Quiambao, J., Pastor, P., Luu, L., Lee, K.H., Kuang, Y., Jesmonth, S., Jeffrey, K., Ruano, R.J., Hsu, J., Gopalakrishnan, K., David, B., Zeng, A., Fu, C.K.: Do as I can, not as I say: Grounding language in robotic affordances. In: CoRL. vol. 205, pp. 287–318 (2023)

17. Ji, Y., Zhu, Z., Zhao, Y., Liu, B., Gao, C., Zhao, Y., Qiu, S., Hu, Y., Yin, Q., Li, Y.: Towards autonomous UAV visual object search in city space: Benchmark and agentic methodology. arXiv preprint arXiv:2505.08765 (2025)
18. Jiao, J., He, J., Liu, C., Aegidius, S., Hu, X., Braud, T., Kanoulas, D.: LiteVLoc: Map-lite visual localization for image goal navigation. In: ICRA. pp. 5244–5251 (2025)
19. Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., Zhu, Y.: Vision-language-action models for robotics: A review towards real-world applications. IEEE Access **13**, 162467–162504 (2025)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM TOG **42**(4), 139 (2023)
21. 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. In: CVPR. pp. 16373–16383 (2024)
22. Lee, J., Miyanishi, T., Kurita, S., Sakamoto, K., Azuma, D., Matsuo, Y., Inoue, N.: CityNav: A large-scale dataset for real-world aerial navigation. In: ICCV. pp. 5912–5922 (2025)
23. Li, J., Padmakumar, A., Sukhatme, G., Bansal, M.: VLN-video: Utilizing driving videos for outdoor vision-and-language navigation. In: AAAI. vol. 38, pp. 18517–18526 (2024)
24. Li, M., Zhao, S., Wang, Q., Wang, K., Zhou, Y., Srivastava, S., Gokmen, C., Lee, T., Li, L.E., Zhang, R., Liu, W., Liang, P., Fei-Fei, L., Mao, J., Wu, J.: Embodied agent interface: Benchmarking LLMs for embodied decision making. In: NeurIPS. vol. 37, pp. 100428–100534 (2024)
25. Liu, X., Li, J., Jiang, Y., Sujay, N., Yang, Z., Zhang, J., Abanes, J., Zhang, J., Feng, C.: CityWalker: Learning embodied urban navigation from web-scale videos. In: CVPR. pp. 6875–6885 (2025)
26. Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., Lin, L.: Aligning cyber space with physical world: A comprehensive survey on embodied AI. IEEE/ASME TMECH **30**, 7253–7274 (2025)
27. Liu, Z., He, H., Alumootil, V., Pandya, A., Squicciarini, B., Wu, W., Zhou, B.: SidewalkBench: Benchmarking visual navigation on urban sidewalks. arXiv preprint arXiv:2606.16953 (2026)
28. Majumdar, A., Aggarwal, G., Devnani, B., Hoffman, J., Batra, D.: ZSON: Zero-shot object-goal navigation using multimodal goal embeddings. In: NeurIPS. vol. 35, pp. 32340–32352 (2022)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: ECCV. vol. 12346, pp. 405–421 (2020)
30. Mirowski, P., Grimes, M., Malinowski, M., Hermann, K.M., Anderson, K., Teplyashin, D., Simonyan, K., Zisserman, A., Hadsell, R., et al.: Learning to navigate in cities without a map. NeurIPS **31** (2018)
31. Miyai, A., Zhao, Z., Egashira, K., Sato, A., Sunada, T., Onohara, S., Yamanishi, H., Toyooka, M., Nishina, K., Maeda, R., et al.: WebChoreArena: Evaluating web browsing agents on realistic tedious web tasks. arXiv preprint arXiv:2506.01952 (2025)
32. Mukhopadhyay, S., Rajgaria, A., Khatiwada, P., Shrivastava, M., Roth, D., Gupta, V.: MAPWise: Evaluating vision-language models for advanced map queries. In: NAACL. pp. 9348–9378 (2025)
33. OpenAI: GPT-5 system card. Tech. rep., OpenAI (2025)

34. Paz-Argaman, T., Tsarfaty, R.: RUN through the streets: A new dataset and baseline models for realistic urban navigation. In: EMNLP-IJCNLP. pp. 6449–6455 (2019)
35. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A., Hlavac, M., Min, S.Y., Vondruš, V., Gervet, T., Berges, V.P., Turner, J.M., Maksymets, O., Kira, Z., Kalakrishnan, M., Malik, J., Chaplot, D.S., Jain, U., Batra, D., Rai, A., Mottaghi, R.: Habitat 3.0: A co-habitat for humans, avatars, and robots. In: ICLR (2024)
36. Savva, M., Chang, A.X., Dosovitskiy, A., Funkhouser, T., Koltun, V.: MINOS: Multimodal indoor simulator for navigation in complex environments. arXiv:1712.03931 (2017)
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. In: ICCV. pp. 9339–9347 (2019)
38. Schumann, R., Riezler, S.: Generating landmark navigation instructions from maps as a graph-to-text problem. In: ACL-IJCNLP. pp. 489–502 (2021)
39. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: ViNT: A foundation model for visual navigation. In: CoRL. vol. 229, pp. 711–733 (2023)
40. Sugimoto, N., Ebine, Y., Aizawa, K.: Building movie map – a tool for exploring areas in a city – and its evaluations. In: ACM MM. pp. 3330–3338 (2020)
41. Takenawa, M., Sugimoto, N., Wöhler, L., Ikehata, S., Aizawa, K.: Building and evaluating a realistic virtual world for large scale urban exploration from 360 ° videos. MTAP **85**(2), 149 (2026)
42. Wang, S., Liang, C., Gao, Y., Yu, E., Li, S., Li, J., Wang, H.: Cityseeker: How do VLMs explore embodied urban navigation with implicit human needs? In: ICLR (2026)
43. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
44. Wu, W., He, H., He, J., Wang, Y., Duan, C., Liu, Z., Li, Q., Zhou, B.: MetaUrban: An embodied AI simulation platform for urban micromobility. In: ICLR (2025)
45. Xie, Z., Liu, Z., Peng, Z., Wu, W., Zhou, B.: Vid2Sim: Realistic and interactive simulation from video for urban navigation. In: CVPR. pp. 1581–1591 (2025)
46. Xing, S., Sun, Z., Xie, S., Chen, K., Huang, Y., Wang, Y., Li, J., Song, D., Tu, Z.: Can large vision language models read maps like a human? arXiv preprint arXiv:2503.14607 (2025)
47. Xu, Q., Yi, X., Xu, J., Tao, W., Ong, Y.S., Zhang, H.: Few-shot NeRF by adaptive rendering loss regularization. In: ECCV. vol. 15124, pp. 125–142 (2024)
48. Yang, J., Pavone, M., Wang, Y.: FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In: CVPR. pp. 8254–8263 (2023)
49. Yang, J., Ding, R., Brown, E., Qi, X., Xie, S.: V-IRL: Grounding virtual intelligence in real life. In: ECCV. vol. 15103, pp. 36–55 (2024)

50. 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. In: CVPR. pp. 10632–10643 (2025)
51. Yang, R., Chen, H., Zhang, J., Zhao, M., Qian, C., Wang, K., Wang, Q., Koripella, T.V., Movahedi, M., Li, M., Ji, H., Zhang, H., Zhang, T.: EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In: ICML. vol. 267, pp. 70576–70631 (2025)
52. Zaffar, M., Garg, S., Milford, M., Kooij, J., Flynn, D., McDonald-Maier, K., Ehsan, S.: VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. IJCV **129**, 2136–2174 (2021)
53. Zhang, D., Wang, C., Wang, W., Li, P., Qin, M., Wang, H.: Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. In: ECCV. vol. 15134, pp. 341–359 (2024)
54. Zhang, M., Qu, K., Patil, V., Cadena, C., Hutter, M.: Tag map: A text-based map for spatial reasoning and navigation with large language models. In: CoRL. vol. 270, pp. 2120–2146 (2024)
55. Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al.: WebArena: A realistic web environment for building autonomous agents. In: ICLR (2024)
56. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.W.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL. vol. 229, pp. 2165–2183 (2023)