
A Benchmark Platform for Embodied Agent in Urban Environment

Chen Gao*, Baining Zhao*, Weichen Zhang, Jinzhu Mao,
Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui,
Xinlei Chen, Yong Li
Tsinghua University
Beijing, China
chgao96@gmail.com
liyong07@tsinghua.edu.cn
chen.xinlei@sz.tsinghua.edu.cn

Abstract

Recent embodied intelligence research aims to build up models to possess perceiving, planning, and acting abilities, thereby enabling real-time interaction with the world. However, most works focus on bounded indoor environments, with limited literature on open-world scenarios. Urban environments, being the most complex open-world settings, contain rich semantic information and high-density structures. To address this, we propose a comprehensive benchmark platform for embodied tasks in urban environments tailored to multi-modal large language models (LLMs) and agents empowered by LLMs. This platform includes a simulator and datasets on five representative tasks for embodied intelligence evaluation in an urban environment. We highlight three representative aspects based on these tasks: scene understanding, reasoning, and decision-making. Based on the benchmark, we evaluate the abilities of those popular large language models, and the results show large differences among different LLMs. This benchmark serves as a basic test tool for embodied intelligence levels in urban environments.

1 Introduction

Embodied artificial intelligence [1] serves as the recent advance of artificial intelligence, presenting an emerging paradigm shift from the traditional Internet artificial intelligence. Specifically, embodied artificial intelligence is more like the real human, which is able to learn from the environment and dynamically interact with the world, considered an essential approach to Artificial General Intelligence (AGI) [2]. Various tasks for embodied intelligence have been established in different domains, including robotics [3, 4, 5], game AI [6, 7], unmanned vehicles/aerial drones [8], etc. Embodied intelligence requires an agent to accurately understand the environment, perform high-level reasoning, and effectively choose appropriate actions to execute tasks, which is quite challenging for traditional artificial intelligence models. Therefore, although the existing works have made some achievements, especially for low-level control, there are still critical challenges for high-level reasoning and decision-making in an open and dynamic environment. Recently, large language models have shown astonishing human-like abilities, not only in natural language processing tasks

*The first two authors contribute to this work equally.

28 but also in general tasks and problems [9]. Quickly, large language models are considered one critical
29 component in embodied artificial intelligence tasks [10, 11, 12, 13].

30 The definition of systematic tasks and evaluation standards is crucial to supporting research on
31 embodied intelligence. The most important aspect of this process is providing an environment where
32 embodied agents can obtain observations in real-time from a first-person perspective, generate actions,
33 and receive feedback [14]. This environment should further support the practical implementation and
34 application of embodied agents.

35 However, existing research mainly focuses on relatively limited environments [1]. For instance,
36 many environments [15, 16, 17] emphasize indoor embodied intelligence scenarios, considering
37 tasks within only a few rooms. These tasks include visual QA tasks targeting certain objects in
38 the room or simple task decomposition [18, 19]. Such benchmarks actually restrict the validation
39 of embodied agents’ capabilities within a very narrow boundary, with low task difficulty and no
40 comparison to general artificial intelligence. Therefore, this paper will, for the first time, consider
41 extending embodied agents from indoor rooms to outdoor cities, expanding tasks beyond indoor
42 spaces to a broad urban environment. We will define relevant evaluation tasks in this urban space,
43 provide task labels through a combination of large models and manual annotation, and define a series
44 of quantifiable metrics.

45 It is worth mentioning that artificial intelligence research in the open world is not a new topic [20, 21,
46 22]. However, embodied intelligence’s uniqueness lies in the need for a learning process that starts
47 almost from scratch, involving many reasoning tasks. This differs from previous open-world tasks,
48 presenting higher difficulty, greater research value, and a level closer to human intelligence.

49 In this work, we first constructed a city-embodied environment simulator. This platform is developed
50 based on a city simulator, providing 3D environments and interactions. The environment is built
51 in one of the China’s largest cities, Beijing. In a commercial area nearby, we established realistic
52 and detailed city-building models as the foundation for the entire city’s open environment embodied
53 intelligence platform. Furthermore, we defined a system of five tasks, including embodied scene
54 description, embodied question answering, embodied dialogue, embodied visual language navigation,
55 and embodied task planning. For each task, we carefully and manually set up the input/output, and
56 construct the ground truth data combined with large language models and human labor. We also
57 provide the interface for the platform through which the agents can obtain the embodied observations
58 and take actions in real-time simulation, after which the agent can be evaluated. Moreover, we
59 deploy the most famous and widely used large language models to construct the embodied agents, the
60 intelligence level of which is evaluated on five tasks. The contribution of this work can be summarized
61 as follows.

- 62 • To the best of our knowledge, we take the first step to construct a benchmark platform for embodied
63 intelligence in an urban environment, in which we define five important embodied tasks.
- 64 • For each task, we set up the ground truth and evaluation metrics for these tasks. The test samples
65 are carefully designed to fully reflect the intelligence level and the requirements in the real-world
66 urban environment. We then construct the ground-truth labels with large language models and
67 human refinement.
- 68 • For the benchmark platform, we build the interface for embodied agents to observe, take action,
69 and receive feedback. We further conduct evaluations on those popular large language models to
70 verify the usability of our benchmark and have a quick look at the embodied intelligence level of
71 these large language models.

72 The organization of this paper is as follows. We will first present the simulator in Section 2 and
73 introduce the five embodied tasks in Section 3. We then conduct evaluations on those popular large
74 language models with the constructed benchmark in Section 4. At last, we discuss the usage and
75 limitations of the benchmark in Section 5 and conclude our work in Section 6. The framework of the
76 benchmark² is illustrated in Figure 1.

²The open platform of this benchmark is at this link: <https://opencity.fiblab.net/>

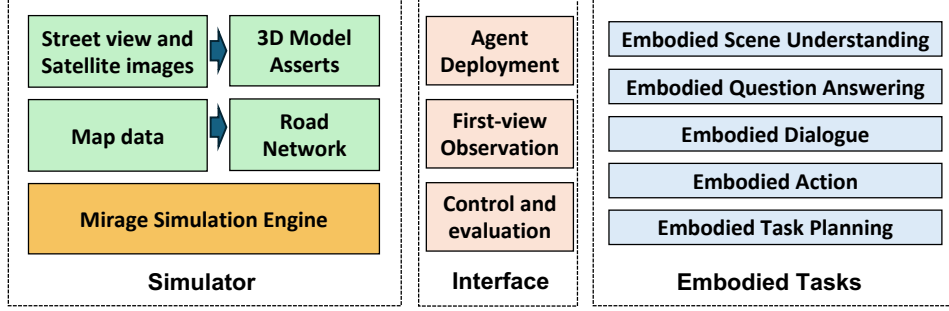


Figure 1: The framework of our constructed benchmark, including the simulator, interface, and embodied tasks.

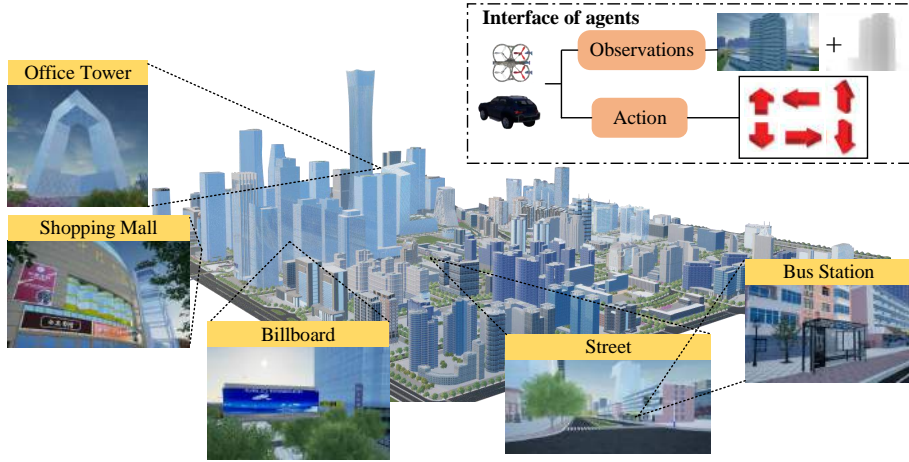


Figure 2: Illustration of the benchmark platform, in which there is a simulated city environment and the interface for agent-environment interaction.

2 Benchmark: Simulator

The simulator is one of the most important modules of the benchmark, for which we try to construct an environment where the agent can perceive, reason, and take actions, which is expected to be similar to the real world. In the following, we present the constructed simulator, which includes the simulation environment, the interface for embodied agents, and the open platform. We provide an illustration of the whole system in Figure 2.

2.1 Simulation Environment

The basic environment of the simulator includes a large business district in Beijing, one of the biggest city in China, in which we build 3D model for buildings, streets, and other elements, hosted by Unreal Engine 4.17³.

- Buildings. We first manually use Blender⁴ to create the 3D asserts of the buildings, for which we use the streetview services of Baidu Map⁵ and Amap⁶. The city level detail includes a variety of building types such as office towers, shopping malls, residential complexes, and public facilities.

³<https://www.unrealengine.com/>

⁴<https://www.blender.org/>

⁵<https://map.baidu.com/>

⁶<https://amap.com/>

90 These models are textured and detailed to closely resemble their real-world counterparts to enhance
91 realism in the simulation.

92 • Streets. The streets are modeled to include all necessary components such as lanes, intersections,
93 traffic signals, and road markings. We also incorporate pedestrian pathways, cycling lanes, and
94 parking areas. Data from traffic monitoring systems and mapping services help ensure that the
95 street layout and traffic flow patterns are accurate and realistic.

96 • Other Elements. Besides streets and buildings, other elements include street furniture (benches,
97 streetlights, signs), vegetation (trees, shrubs, lawns), and urban amenities (bus stops, metro en-
98 trances, public restrooms). These are also created using Blender, based on real-world references
99 from the streetview services mentioned above. Additionally, dynamic elements like vehicles and
100 pedestrians are simulated to move realistically within the environment, contributing to the liveliness
101 and accuracy of the urban simulation. The simulation algorithms of vehicles and pedestrians are
102 based on Mirage Simulation System [23].

103 2.2 Interface of embodied agents

104 With the simulation environment of unreal engine, we further build the interface of embodied agents
105 to ensure the agents can indeed embody themselves in the system. To implement it, we use the AirSim
106 plugin provided by Microsoft. Specifically, AirSim is originally designed for airdrones, for which
107 the observations are conducted through a first-view manner, and the control for airdrones includes
108 motion, velocity, accelerated velocity, etc.

109 • Observations. The observations for the embodied agents are designed to replicate the sensory
110 inputs that would be available to real-world agents. This includes first-person view (FPV) cameras
111 to simulate visual input, capturing RGB images, depth information, and segmentation images.
112 Additionally, we can incorporate other optional sensors such as LiDAR for spatial awareness, and
113 GPS for positional data. These inputs allow the agents to perceive their environment accurately and
114 make informed decisions based on the simulated sensory data.

115 • Actions. The actions for the embodied agents are designed to mimic realistic controls similar
116 to those used by air drones. This includes commands for movement in three-dimensional space,
117 such as changes in position (x, y, z coordinates), orientation (pitch, yaw, roll), and velocity. The
118 agents can also execute more complex maneuvers by adjusting their acceleration and applying
119 force vectors. These action commands enable the agents to navigate the simulated environment
120 effectively, avoiding obstacles, reaching targets, and performing tasks as required by the simulation
121 objectives.

122 3 Benchmark: Embodied Tasks

123 With the constructed environment, based on the different aspects of embodied intelligence abilities,
124 we further define five important embodied tasks. Although it is not a huge number, the tasks are
125 representative. Specifically, the intelligent agents in the open world are expected to have three kinds
126 of human-like abilities: perception, reasoning, and decision-making. For perception, we consider the
127 task of embodied first-view scene understanding; for reasoning, we consider the task of embodied
128 question answering and dialogue; for decision-making, we consider the task of embodied action
129 (visual-language navigation) and embodied task planning. For a better understanding, we present five
130 tasks in Figure 3.

131 3.1 Embodied first-view scene understanding

132 The first-view scene understanding requires the agent able to well observe the environment, and
133 give the accurate description, which could be considered as a basic ability for the further tasks. In our
134 benchmark, we observe from different perspectives at the same location, generating a set of RGB
135 images, *i.e.*, the input of scene understanding. The output is the textual description for the given
136 scene images.

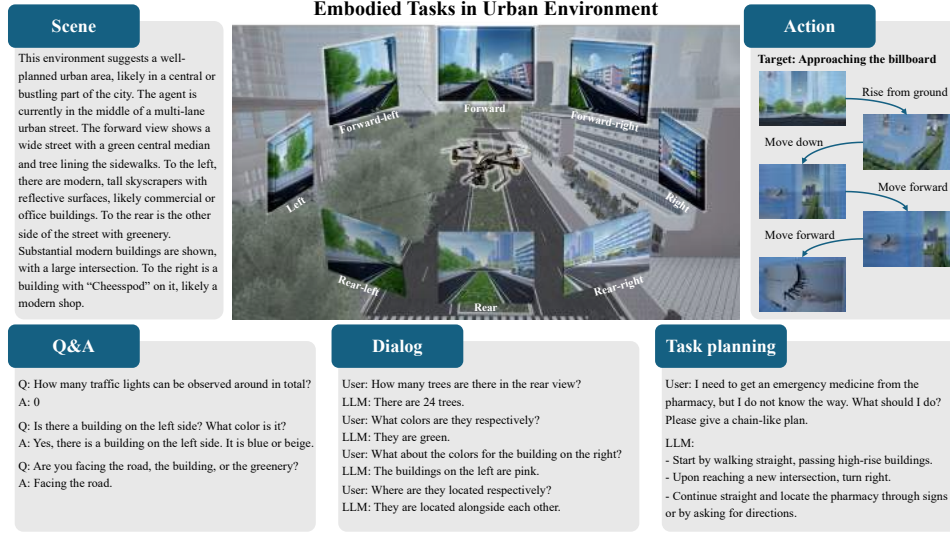


Figure 3: Embodied tasks in urban environment.

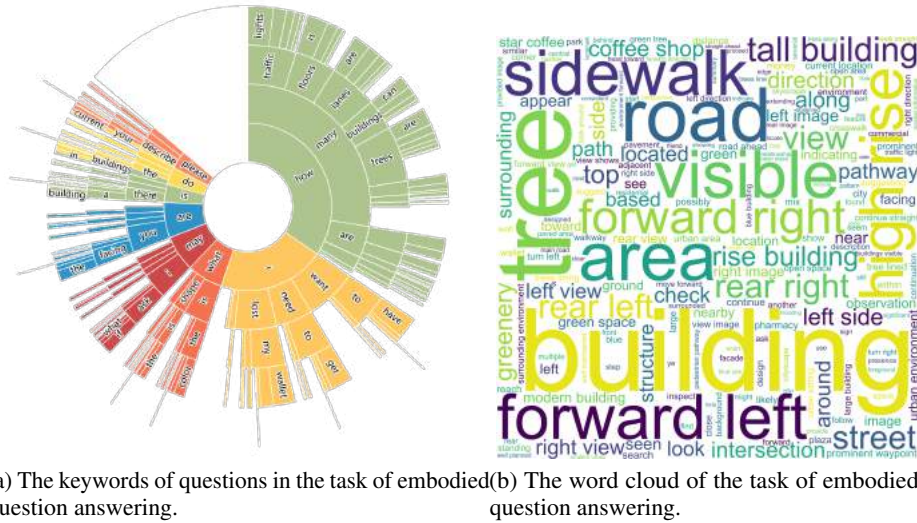


Figure 4: Illustration of the involved topics and keywords in the task of embodied question answering in our benchmark.

3.2 Embodied Question Answering

With the first-view observation, the embodied agent could be further fed with a query posed in natural language about the environment. For example, "What is the color of the building?" or "How many buildings are in the street?" Therefore, the input includes both the first-view RGB images and a query about the environment. The output should be the direct textual responses to the question.

3.3 Embodied dialogue

Despite the task of embodied question answering, a more complex embodied task close to the reasoning ability is embodied dialogue. Specifically, embodied dialogue involves ongoing interactions where the agent engages in a back-and-forth conversation with the user. This requires maintaining context and understanding the flow of dialogue. Therefore, the input includes embodied observations and multi-round queries, and the output is the multi-round responses.

Refinement Procedures	
Raw Responses	Refined Responses
1. Wrong Counting: There are two traffic lights around.	1. Wrong Answers There are two curtains in the room.
2. Wrong Existence: A cafe is on the left.	2. Removing Non-Existence response. A cafe is on the left.
3. Wrong Position: A pink tree stands on the rear left.	3. Position Refinement: A pink tree stands on the rear right.
4. Negative Responses: Question: What shape is the building in the right front? Answer: Unknown	4. Genenerate Correct Responses: Answer: The building is a columnar building covered with transparent glass.
5. Unnecessary Contents Certainly, I would like to help ...	5. Removing Unnecessary Contents Certainly, I would like to help ...

Figure 5: The refinement procedures when constructing the benchmark, which aims to address the errors in raw responses, inspired by [24].

3.4 Embodied Action (VLN)

Embodied Action, often referred to as Vision-and-Language Navigation (VLN), is a research area in artificial intelligence that focuses on enabling an agent to navigate an environment based on natural language instructions. The input combines visual perception and natural language instructions to guide the agent through complex environments. The output is the action sequences following the language instructions.

3.5 Embodied Task Planning

Most times, the decision-making in the real world does not have explicit instructions; otherwise, there is only a task goal. Thus, it is significant for the embodied agents to be able to compose the complex and long-term task goals into several sub-tasks, which we refer to as embodied task planning. The input is the first-view observations and a given natural language described task goal, and the output should be a series of sub-tasks that the agent plans to execute.

Human refinement More details about collecting the labels can be found in the supplemental material. During the construction process, human refinement plays an important role, which is illustrated in Figure 5, inspired by [24]. There are five types of scenarios that require human refinement: wrong counting, wrong existence, wrong position, negative responses, and unnecessary content.

4 Evaluation of Large Language Model Agents

We select popular and representative large language models for evaluation to verify the application value of our benchmark and test their intelligence level for embodied tasks in the urban environment. The considered large language models include fuyu-8b [25], Qwen-VL [26], Claude 3 [27], GPT-4 Turbo [28].

4.1 Task I: Embodied first-view scene understanding

The results of the performance evaluation are presented in Table 1, from which we have the following observations:

- Claude 3 has shown the best performance on the task of embodied scene understanding, with the best performance on almost all metrics. Actually, in this task, the different metrics have similar

Table 1: Results of embodied first-view scene understanding, including typical evaluation metrics: BLEU [29], ROUGE [30], METEOR [31], and CIDEr [32].

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
fuyu-8B	40.25	20.26	8.40	1.57	17.29	15.80	21.55
Qwen-VL	40.57	17.59	5.90	0.98	14.61	19.13	18.40
Claude 3	57.38	31.73	16.83	7.19	21.60	29.00	29.20
GPT-4 Turbo	54.01	27.63	12.73	4.53	21.99	28.48	22.39

Table 2: Results of embodied question answering. The Counting task involves querying the number of a specific object within the field of view. The Property task entails inquiring about the attributes of spatial entities such as city buildings or objects within the field of view, including aspects like shape and color. The Position task concerns querying the spatial relationships between different urban elements within the field of view.

Type	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
Counting	fuyu-8B	12.00	7.15	1.07	0.40	16.45	15.41	8.87
	Qwen-VL	5.49	1.19	0.10	0	11.46	17.89	3.58
	Claude 3	6.08	4.33	2.79	2.13	10.54	16.82	7.95
	GPT-4 Turbo	12.84	8.81	4.33	2.78	19.26	20.18	11.56
Property	fuyu-8B	20.19	18.36	16.39	14.64	31.55	20.34	22.56
	Qwen-VL	55.77	48.43	40.90	31.94	65.33	61.73	33.30
	Claude 3	49.34	41.88	34.10	23.44	60.51	55.29	29.84
	GPT-4 Turbo	76.63	72.17	68.57	65.51	80.16	77.10	61.44
Position	fuyu-8B	7.46	0.15	0	0	18.94	4.40	12.86
	Qwen-VL	7.88	4.63	3.81	0.83	18.03	22.00	16.62
	Claude 3	7.57	5.85	4.37	1.56	19.04	34.28	18.82
	GPT-4 Turbo	64.54	61.85	59.44	55.31	70.72	68.87	58.45

distinguishing abilities, *i.e.*, a more with better performance on one metric is likely to be better on another metric.

- Larger scale models steadily outperform those smaller ones. As we can observe, fuyu-8B and Qwen-VL have similar parameter sizes (7B-8B), which are far smaller than Claude 3 and GPT-4 Turbo.

4.2 Task II: Embodied question answering

The results of the performance evaluation are presented in Table 2, from which we have the following observations:

- GPT-4 Turbo achieves a very significant performance improvement against all other models, of which the average improvement is larger than 100%. This may be explained by the GPT-4’s stronger ability to handle textual data.
- Smaller models are very unsteady on three types of tasks, counting, property, and position, for which some metrics are 0.

4.3 Task III: Embodied dialogue

The results of the performance evaluation are presented in Table 3, from which we have the following observations:

- GPT-4 Turbo shows the best performance with significant gain, which could be explained by the long-context abilities, which is the major requirement of multi-round conversions.
- The poor performance of Qwen-VL provides insights that it may be a promising solution to combine large language models that do not support vision input, such as QWen.

Table 3: Results of embodied dialogue.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
fuyu-8B	29.05	16.73	8.24	4.30	28.53	30.12	14.47
Qwen-VL	17.91	9.54	3.90	2.03	19.33	19.65	10.30
Claude 3	24.86	18.02	13.14	9.70	29.06	38.56	28.62
GPT-4 Turbo	41.77	34.27	27.82	23.26	42.29	51.72	35.64

Table 4: Results of embodied vision-and-navigation.

Model	Short			Long			Mean		
	SR/%	SPL/%	NE/m	SR/%	SPL/%	NE/m	SR/%	SPL/%	NE/m
Qwen-VL	33.33	29.60	67.30	8.33	6.67	145.3	22.22	19.33	120.44
Claude 3	76.92	75.60	139.11	20.00	19.65	185.48	34.90	34.25	162.35
GPT-4 Turbo	60.90	55.21	95.93	15.62	14.16	127.87	27.71	25.12	111.92
GPT-4O	76.92	75.60	77.23	20.00	19.65	102.98	34.90	34.25	90.11

Table 5: Results of embodied task planning.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr
fuyu-8B	15.11	6.37	1.71	0.45	14.72	19.11	16.84
Qwen-VL	20.28	9.10	3.75	1.44	19.42	17.90	11.36
Claude 3	29.21	16.22	9.17	4.40	22.85	31.58	21.78
GPT-4 Turbo	28.23	13.72	6.26	2.82	21.61	28.47	16.41

194 4.4 Task IV: Embodied VLN

195 The results of the performance evaluation are presented in Table 4, from which we have the following
 196 observations:

- 197 • Both GPT-4o and Claude 3 achieve the best performance on SR and SPL. And GPT-4o also has the
 198 lowest NE compared to other models. This implies that GPT-4o has the strongest spatial reasoning
 199 capacity that always navigates the drone in the correct direction.
- 200 • Chinese LLM (Qwen-VL) has a significant performance drop against English LLM. Qwen-VL is
 201 12% and 15% lower than the best-performing model in SR and SPL metrics, respectively. This
 202 result can be attributed to the superior performance of the English LLM in understanding English
 203 task descriptions and applying them to action reasoning.
- 204 • All models perform better on short navigation tasks than long navigation tasks which involve longer
 205 reasoning chains and more dramatic scene changes, causing higher failure rates.

206 4.5 Task V: Embodied task planning

207 The results of the performance evaluation are presented in Table 5, from which we have the following
 208 observations:

- 209 • Claude-3 achieves the best performance on embodied task planning. Actually, task planning relies
 210 more on decision-making ability with common sense and contextual information. Therefore, it pay
 211 less attention to the multi-modal understanding ability.
- 212 • Smaller LLMs show poorer performance, but the performance gap is acceptable, which inspires us
 213 to deploy mixture-architecture agents, combining the strengths of larger and smaller LLMs.

214 5 Discussions and limitations of the benchmark

215 5.1 Application of the benchmark

216 The usage of the benchmark does not only serve as the pure evaluation of the large language model
 217 or LLM agents but also could be a sim2real tool that supports the pre-training or pre-testing before
 218 deployed to the real-world city environment. From the type of agents, the benchmark does not set
 219 constraints. That is, the agent deployed could be a robot or air drone. The differences among different
 220 types of real world agents are the input or output. For example, the input of a robot may only include

221 the RGB images, and for air drones, the input can also contain the radar signals. The degree of
222 freedom of different agents could also be different.

223 5.2 About human refinement

224 When constructing the benchmark, we spent a lot of effort in using human refinement steps to filter
225 out low-quality responses or revise incorrect answers provided by GPT. It is worth noticing that the
226 paradigm of combining large language models and human crafts is recently widely used since large
227 language models accurately and skillfully generate various responses (but may be totally wrong).
228 The key challenge here is the accuracy rather than diversity, and thus the human efforts to refine the
229 answers are quite essential and useful. On the other hand, the cost of collecting all the responses
230 with human labor only is not affordable. Therefore, the usage of the large language models does not
231 introduce a large bias. In other words, using GPT as the first-step large language model does not
232 bring unfairness for other large language models when evaluating the refined answers.

233 5.3 Limitations of task evaluation

234 In our constructed benchmark, we consider five types of embodied tasks, scene description, embodied
235 question answering, embodied dialogue, visual-language navigation, and embodied task planning.
236 From a perspective of human-like critical abilities, these tasks well cover the three most significant
237 aspects: perception, reasoning, and decision-making. The follow-up work, based on the simulation
238 environment, promises to extend to more tasks, of which the potential tasks could be as follows. (1)
239 Multi-agent Collaboration: Introducing tasks that require coordination and communication between
240 multiple agents to achieve common goals. (2) Long-term Planning and Memory: Developing tasks
241 that require agents to remember past interactions and plan over extended time periods. (3) Human-
242 Agent Interaction: Creating scenarios where human users interact with agents necessitates a more
243 sophisticated understanding of human behavior and natural language. (4) Adaptability and Learning:
244 Implementing tasks that test an agent’s ability to learn from its environment and adapt to new,
245 unforeseen challenges, such as the fire scenario. (5) Complex Manipulation: This includes tasks that
246 involve manipulating objects in the environment, requiring the combination of the results of the large
247 language model agent and control module.

248 5.4 Online platform for AI embodied agents and MetaCity

249 The platform in the benchmark may be further developed into an online multi-user system, similar
250 to a multi-user online game, where each user can deploy their own AI embodied agent. This
251 platform allows users to create, train, and interact with their agents in a dynamic and engaging virtual
252 environment. These agents can perform various tasks, learn from interactions, and improve their
253 abilities over time. Users can observe their agents’ behaviors, provide feedback, and even program
254 specific functionalities to tailor the agents’ performance to their preferences and objectives. The
255 collaborative and competitive elements of the platform enhance user engagement and provide a rich
256 context for agent development and testing. Looking ahead, the platform holds the potential to evolve
257 into a concept of *metacity*, akin to the concept of the metaverse.

258 6 Conclusion and Future Work

259 In this work, we take a pioneering step by building a systematic benchmark for embodied intelligence
260 in an open city environment. The benchmark contains a 3D city simulator, five kinds of embodied
261 tasks, and human-refined truth labels. We further evaluate the intelligence level of the mainstream
262 large language models, which verify the rationality of the constructed benchmark. For future work,
263 we plan to extend the scale of the simulation environment, add more embodied tasks and labeled
264 data, and add more action space for large language model agents. We also plan to evaluate large
265 language model agents’ performance and embodied intelligence level in the real city environment via
266 a Sim2Real paradigm, which can further validate the application value of the benchmark.

References

- [1] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.
- [2] Edgar A Duéñez-Guzmán, Suzanne Sadedin, Jane X Wang, Kevin R McKee, and Joel Z Leibo. A social path to human-like artificial intelligence. *Nature Machine Intelligence*, 5(11):1181–1188, 2023.
- [3] Qiguang He, Rui Yin, Yucong Hua, Weijian Jiao, Chengyang Mo, Hang Shu, and Jordan R Raney. A modular strategy for distributed, embodied control of electronics-free soft robots. *Science Advances*, 9(27):eade9247, 2023.
- [4] Jose A Barreiros, Artemis Xu, Sofya Pugach, Narahari Iyengar, Graeme Troxell, Alexander Cornwell, Samantha Hong, Bart Selman, and Robert F Shepherd. Haptic perception using optoelectronic robotic flesh for embodied artificially intelligent agents. *Science Robotics*, 7(67):eabi6745, 2022.
- [5] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [6] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.
- [7] Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling. In *International Conference on Machine Learning*, pages 26311–26325. PMLR, 2023.
- [8] Xin Zhou, Xiangyong Wen, Zhepei Wang, Yuman Gao, Haojia Li, Qianhao Wang, Tiankai Yang, Haojian Lu, Yanjun Cao, Chao Xu, et al. Swarm of micro flying robots in the wild. *Science Robotics*, 7(66):eabm5954, 2022.
- [9] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [10] Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazouze, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*, 2023.
- [11] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.
- [12] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.
- [13] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024.

- [14] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.
- [15] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.
- [16] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Proctor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022.
- [17] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073, 2021.
- [18] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.
- [19] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research*, 74:459–515, 2022.
- [20] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.
- [21] Jitendra Parmar, Satyendra Chouhan, Vaskar Raychoudhury, and Santosh Rathore. Open-world machine learning: applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37, 2023.
- [22] Vikash Sehwal, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 105–116, 2019.
- [23] Jun Zhang, Depeng Jin, and Yong Li. Mirage: an efficient and extensible city simulation framework (systems paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–4, 2022.
- [24] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.
- [25] Fuyu-8B: A Multimodal Architecture for AI Agents — adept.ai. <https://www.adept.ai/blog/fuyu-8b/>. [Accessed 06-06-2024].
- [26] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [27] Introducing the next generation of Claude — anthropic.com. <https://www.anthropic.com/news/claude-3-family>. [Accessed 06-06-2024].
- [28] OpenAI Team. GPT-4. <https://openai.com/index/gpt-4/>. [Accessed 06-06-2024].
- [29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

- 356 [30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization*
357 *branches out*, pages 74–81, 2004.
- 358 [31] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with
359 improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic*
360 *and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72,
361 2005.
- 362 [32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image
363 description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern*
364 *recognition*, pages 4566–4575, 2015.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? **[Yes]**
- Did you include the license to the code and datasets? **[No]**
- Did you include the license to the code and datasets? **[N/A]**

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...

- (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[Yes]**
- (b) Did you describe the limitations of your work? **[Yes]**
- (c) Did you discuss any potential negative societal impacts of your work? **[Yes]**
- (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[Yes]**

2. If you are including theoretical results...

- (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
- (b) Did you include complete proofs of all theoretical results? **[N/A]**

3. If you ran experiments (e.g. for benchmarks)...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[Yes]**
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[Yes]**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[Yes]**
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

- (a) If your work uses existing assets, did you cite the creators? **[N/A]**
- (b) Did you mention the license of the assets? **[N/A]**
- (c) Did you include any new assets either in the supplemental material or as a URL? **[Yes]**
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[Yes]**
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[Yes]**

5. If you used crowdsourcing or conducted research with human subjects...

- (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[N/A]**
- (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[N/A]**
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[N/A]**

410 A Supplementary Materials

411 A.1 Simulator

412 A.1.1 Simulation Environment Detail

413 Besides Figure 2, we further provide detailed information on the simulation environment, as shown
 414 in Figure 6. Our city simulator is a tool designed for urban planning, analysis, and autonomous
 415 vehicle simulation. It offers superior capabilities compared to other available simulators, featuring
 416 high-resolution 3D models and real-time data integration for an exceptionally realistic and dynamic
 417 representation of urban environments. The simulator’s customization options allow users to model
 418 diverse scenarios and explore various urban elements, from detailed building features to specific
 419 street-level details. Specifically, it supports simulations for drones and unmanned vehicles, making it
 420 an invaluable resource for testing and optimizing autonomous sensing, navigation, and planning in
 421 urban settings.

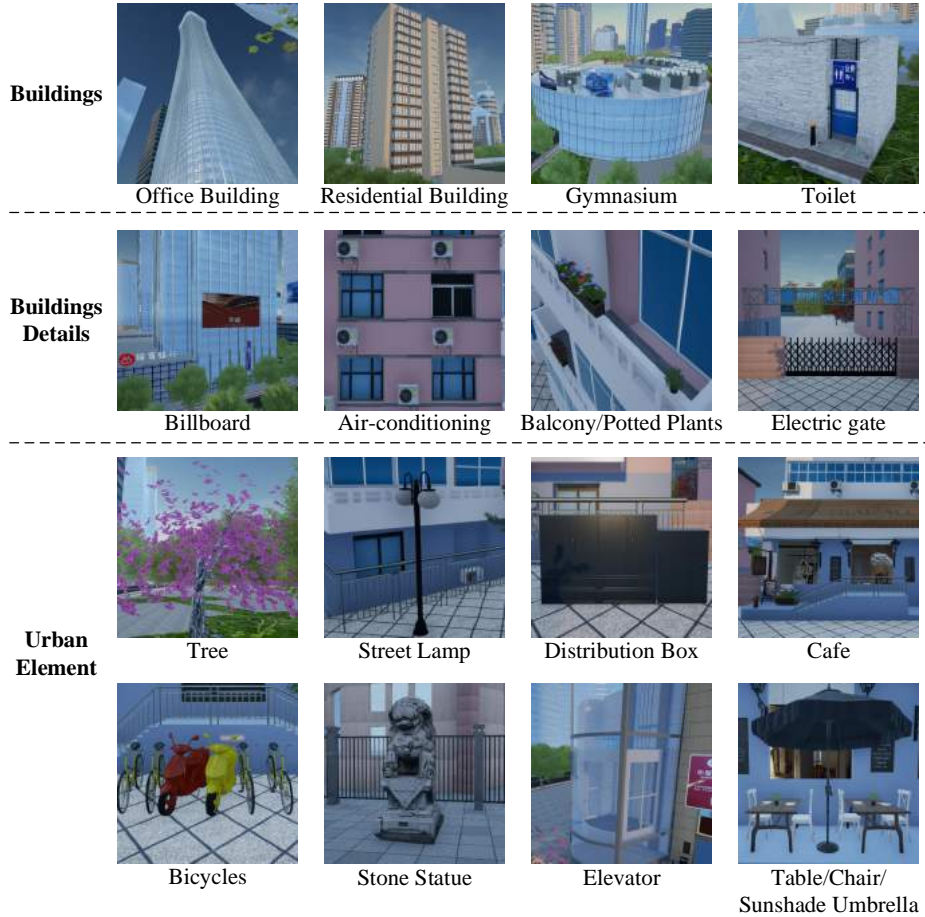


Figure 6: The image showcases various components of our city simulator, which includes a diverse array of buildings such as office buildings, residential buildings, gymnasiums, and public toilets. Additionally, the image highlights detailed structural elements like billboards, air-conditioning units, balconies with potted plants, and electric gates. Urban elements include trees, street lamps, distribution boxes, cafes, bicycles, stone statues, elevators, and outdoor seating arrangements with tables, chairs, and sunshade umbrellas. These components collectively demonstrate the simulator’s comprehensive and detailed approach to urban modeling.



Figure 7: Integration of AirSim within our Unreal Engine city simulator environment, showcasing the simulation of both drones and unmanned vehicles (cars). The left side features a drone navigating an urban setting, while the right side displays an unmanned vehicle on a city street. Both simulations demonstrate the observation and action capabilities of autonomous systems within the city environment.

A.1.2 Interface Detail of Embodied Agents

In our enhanced city simulator, AirSim serves as a powerful plugin to facilitate realistic simulations of drones and unmanned vehicles. These autonomous systems leverage AirSim’s robust observation and action mechanisms to navigate and interact with the urban environment.

For drones, the observation process involves capturing high-resolution images and sensor data from multiple perspectives, including RGB, depth, and segmentation views. These observations enable the drone to perceive its surroundings accurately, identify obstacles, and navigate complex urban landscapes. The action space for drones includes vertical movements (up and down), horizontal movements (forward, backward, left, and right), and rotational adjustments (yaw, pitch, and roll). This comprehensive action space allows drones to maneuver precisely and efficiently in three-dimensional urban environments.

Similarly, for unmanned vehicles, observation is achieved through an array of sensors that provide comprehensive environmental data, including visual feeds and depth information. This allows the vehicle to detect road features, other vehicles, pedestrians, and potential hazards. The action space for unmanned vehicles includes steering (left and right), acceleration (forward movement), and braking (deceleration and stopping). These actions ensure that the vehicle can navigate urban streets safely and efficiently by making real-time adjustments based on its observations.

By integrating AirSim into our city simulator, we provide a detailed and realistic platform for testing and optimizing the performance of autonomous drones and vehicles in urban settings.

A.2 Open Interface

Our city simulator features an open API interface. This API will provide users with the ability to programmatically access and manipulate various aspects of the simulator. Through this interface, users can control camera perspectives, navigate virtual characters, retrieve environmental data, and perform other interactive tasks. The API will be designed with robust measures to ensure safe and authorized access, thereby making our simulator a versatile tool for both research and practical applications.

A.3 Benchmark and Dataset

Based on the urban simulator, we collected 2.1k cases, comprising five types of embodied tasks, as shown in Table 6. Each case in the dataset includes inputs and outputs. It is crucial to generate as diverse inputs as possible and quickly obtain ground truth. Therefore, we introduce GPT-4o, one of the most powerful vision-language models (VLMs), as an aid.

Table 6: **Datasets statistics.** The dataset includes details on how it was collected, how the ground truth was obtained, the number of cases, and the token count of the dataset’s text portion.

Task	Prompt Collection	Ground Truth	Case	Token
Embodied first-view scene understanding	Human	Generation+Refinement	0.2k	117.4k
Embodied question answering	Generation+HumanEdit	Generation+Refinement	0.6k	4.8k
Embodied dialogue	Generation+HumanEdit	Generation+Refinement	0.6k	39.3k
Embodied VLN	Human	Human	0.1k	28.6k
Embodied task planning	Generation+HumanEdit	Generation+Refinement	0.6k	67.3k

Table 7: Examples of first-view scene understanding refinement.

Types	Raw Responses	Refined Responses
Wrong Counting	In front of us, there’s a large building There seem to be three such buildings visible within this frame.	There are several tall buildings made up of glass windows. The surroundings include several large, architecturally modern buildings.
Wrong Existence	Multiple cars parked along the roadside, with varying sizes indicating depth perception Above them, the sky appears clear blue with white clouds scattered throughout.	<i>The unnecessary contents will be removed.</i>
Wrong Position	The scene shows an urban road perspective view in daylight conditions. On both sides of the road stand two-story high walls made of dark-colored stone blocks.	You are in a cityscape with modern and tall buildings. The view shows a tall, modern building made of concrete or stone on the right.
Negative Responses	As an AI language model, I do not have physical senses or locations in the real world. The user is currently standing in an urban area at night time.	Based on the observations from the eight directions, it seems you are in an urban environment surrounded by tall modern buildings, likely in a city center. The user is currently in an urban area at daytime, standing near a road intersection.
Unnecessary Contents	The scene shows an urban street viewed from above at an angle of approximately 45 degrees. The scene shows an urban street viewed from above at a slight angle.	<i>The unnecessary contents will be removed.</i>

• **Embodied first-view scene understanding.** We randomly walk around the city and record the surrounding RGB observations upon reaching a location. For each case, the prompts are fixed and can therefore be designed manually. For the ground truth, we first generate embodied descriptions using the VLM. Then we manually review and correct each response, as shown in Table 7. The refinement process involves five categories of raw responses:

1. Object Counting: The question involves counting a specified object.
2. Object Existence: The response asserts the presence of objects, which may or may not actually exist.
3. Object Position: The response describes the spatial relationship between buildings or objects.
4. Negative Response: Indicates that the question cannot be answered and will be discarded.
5. Unnecessary Content: The response includes redundant information that could impact the calculation of evaluation metrics.

• **Embodied question answering.** Similarly, upon randomly arriving at a certain location, we record the surrounding RGB observations and specifically inquire about details of the current situation, such as the color of buildings in a particular direction or the number of trees nearby. To generate questions with urban characteristics, we have GPT-4o select questions that match the current scene based on the aforementioned images, in conjunction with a pre-generated question bank created manually. The refinement examples are listed in Table 8.

Table 8: Examples of question answering refinement.

Types	Raw Responses	Refined Responses
Wrong Counting	Q: How many traffic lights can be observed around in total? A: None.	Q: How many traffic lights can be observed around in total? A: 1 traffic light can be observed.
Wrong Existence	Q: Is there a building on the left side? A: There is no building visible in any of the provided inputs.	Q: Is there a building on the left side? A: Yes, there is a building on the left side.
Wrong Position	Q: Are you facing the road, the building, or the greenery? A: Road.	Q: Are you facing the road, the building, or the greenery? A: You are facing the greenery.
Negative Responses	Q: How many lanes are there on the road behind? A: There are no lanes on the road behind.	Q: How many lanes are there on the road behind? A: Two lanes.
Unnecessary Contents	Q: How many floors is the tallest building around? A: The spiral shaped skyscraper is clearly the tallest structure in the cityscape.	<i>The unnecessary contents will be removed.</i>

- 471 • **Embodied dialogue.** This task is an enhanced version of question answering, requiring contin-
472 uous question and dialogue responses. It further tests the logical reasoning and vision-language
473 comprehension capabilities of large models. The processes for prompt collection and ground truth
acquisition are similar. Examples of dialogue refinement are shown in Table 9

Table 9: Examples of dialogue refinement.

Types	Raw Responses	Refined Responses
Wrong Counting	There are 3 trees beside the road. I see a tall building straight ahead.	There are four trees at the junction. I see three high ups in front of me.
Wrong Existence	Yes, there is a crosswalk near the road. Yes, there is an intersection ahead which may require turning or stopping for traffic lights.	There are no prominent waypoints. No, there are no prominent waypoints visible in the observations.
Wrong Position	The building is situated next to the road in front of the viewer. It's situated beside the road in front of me.	The building on the left is located on the left side of the road ahead. Adjacent to the road ahead on the left side.
Negative Responses	No, there is no mention of waypoints in the given information about the environment. The locations cannot be determined based solely on the provided information.	Yes, there is a prominent intersection with traffic lights, which can be considered a waypoint. All images depict various angles of one major intersection with traffic lights.
Unnecessary Contents	Certainly, I would like to help ... I'm happy to judge the scene for you, and my verdict is....	<i>The unnecessary contents will be removed.</i>

474

- 475 • **Embodied VLN.** In navigation tasks, it is crucial to reasonably select the agent's starting and target
476 points within the city simulator. The navigation difficulty increases with the distance between the
477 starting point and the target point. Additionally, the target point must be distinctive to ensure the
478 uniqueness of the spatial location referred to by the textual description. The process of Vision-and-
479 Language Navigation (VLN) is dynamic, requiring continuous interaction with the simulator. Each
480 decision at every step influences the subsequent observation, thereby affecting the next decision.
481 Consequently, the input and ground truth for each case are obtained through human annotation.
482 The input consists of the agent's starting coordinates and textual instructions, while the ground
483 truth comprises the route trajectory and the target coordinates.
- 484 • **Embodied task planning.** Compared to the question-answering task, this task presents an increased
485 difficulty level. It requires the output of a step-by-step plan to achieve a complex objective. The

processes of prompt collection and ground truth acquisition are similar to those in question answering. The refinement types consist of abnormal logic, negative responses, non-planned, and unnecessary contents, as shown in Table 10.

Table 10: Examples of task planning refinement.

Types	Raw Responses	Refined Responses
Abnormal logic	Q: I want to have a cup of coffee at ALL-Star coffee shop, but I have not brought any money. What should I do? Please give a chain-like plan. A: 1. Look for any loose change or money on the ground or in the vicinity. 2. If you find some money, use it to purchase a coffee at the ALL-Star coffee shop. 3. If you don't find any money, look for any odd jobs or tasks you can perform around the area in exchange for a cup of coffee, such as helping someone carry their groceries or cleaning up litter.	<i>The abnormal logic will be removed.</i>
Negative Responses (Wrong Position) (Wrong Existence)	Q: I need to get an emergency medicine from the pharmacy, but I do not know the way. What should I do? Please give a chain-like plan. A: 1. Go forward towards the blue building in the center of the image. 2. Turn left and follow the road. 3. Continue straight, crossing the pedestrian crossing. 4. Turn right and follow the road. 5. The pharmacy should be visible on your right. 6. Proceed to the pharmacy and obtain the emergency medicine.	<i>You cannot find a pharmacy along this step. The wrong plan will be removed.</i>
Non-planned	Q: I lost my wallet nearby, and now I need to find it. What should I do? Please give a chain-like plan. A: A good way is that we can use the camera in reverse to see if our wallet appears in this picture.	<i>That's not the plan. It will be removed.</i>
Unnecessary Contents	A: Head forward along the road. There are traffic lights and crosswalks.	<i>The unnecessary contents will be removed.</i>

488

489 A.4 Experimental Setup

490 A.4.1 Evaluation Metrics

491 To evaluate the performance of embodied first-view scene understanding, question answering, di-
 492 alogue, and task planning, we use the following metrics: BLEU-1, BLEU-2, BLEU-3, BLEU-4,
 493 ROUGE, METEOR, and CIDEr. BLEU [29] scores measure n-gram precision at different lengths (1
 494 to 4). ROUGE [30] focuses on recall, capturing how much of the reference text is covered by the
 495 generated text. METEOR [31] combines precision and recall while accounting for synonyms and
 496 word forms. CIDEr [32] evaluates the similarity of generated descriptions to reference descriptions,
 497 particularly in image description tasks, using TF-IDF weighting. These metrics collectively provide a
 498 comprehensive assessment of model performance.

499 A.5 Baselines

500 We utilize powerful large models in multimodal AI, each bringing unique strengths and capabilities
 501 to the embodied tasks.

502 **fuyu-8B** [25], developed by Adept AI, is a multimodal model designed to assist knowledge workers.
503 Its strengths include a simple architecture and training process that facilitate understanding, scalability,
504 and deployment, as well as the ability to handle arbitrary image resolutions, charts, and screen images
505 with fine-grained localization. It offers fast response times (under 100 milliseconds for large images)
506 and performs well on standard image understanding benchmarks such as visual question answering
507 and natural image captions. This model is used to evaluate the capabilities of open-source multimodal
508 large models.

509 **Qwen-VL** [26], proposed by Alibaba Cloud, is a large-scale vision-language model that supports
510 image, text, and detection boxes as input, and produces text and detection boxes as output. It excels
511 in multilingual dialogue, interleaved multi-image dialogue, Chinese open-domain localization, and
512 fine-grained image recognition.

513 **Claude 3** [27], introduced by Anthropic, is a large-scale language model designed for enterprise
514 use, balancing speed and performance. It offers three levels—Haiku, Sonnet, and Opus—catering
515 to different tasks and is known for being secure, reliable, and customizable. This paper primarily
516 utilizes Claude 3 Haiku.

517 **GPT-4 Turbo** [28], released by OpenAI, is a highly intelligent model that supports both image
518 and text inputs, generating text outputs. As one of the most powerful multimodal large models, its
519 performance delineates the optimal embodied capability boundaries of current models.

520 **A.6 Case Study**

521 **A.6.1 Task I: Embodied first-view scene understanding**

522 As shown in Figure 8, fuyu-8B provides a detailed scene description, focusing on visual elements such
523 as the floor, wall, bench, and potted plants. It mentions geometric shapes and artistic perspectives but
524 does not explicitly state the general location or surrounding environment. The model’s strengths lie
525 in its detailed visual analysis and focus on specific elements. However, it lacks a holistic description
526 of the location and environment, which limits its overall effectiveness in this task.

527 Qwen-VL emphasizes its inability to sense physical locations but infers that the scene could be an
528 urban setting with buildings, roads, sidewalks, and possibly parks or plazas. Although it provides a
529 general guess, it lacks specific details and a comprehensive scene description. The model’s strengths
530 include acknowledging its limitations and providing a broad inference, but its weaknesses are evident
531 in the lack of detailed specifics and a thorough scene description.

532 Claude 3 identifies the scene as a large open plaza in an urban environment, surrounded by tall
533 buildings and various types of trees. It notes the tiled ground, benches, and potential for a gathering
534 space. The description is coherent and aligns well with the observed images. Claude 3’s strengths are
535 its comprehensive and coherent description, identification of key features, and correct inference of
536 the environment.

537 GPT-4 Turbo describes the scene as a large urban plaza surrounded by high-rise buildings with a
538 tiled ground and arranged trees. It mentions the presence of a pedestrian area and suggests the scene
539 is a public space within a city, characterized by a blend of natural elements and urban architecture.
540 The model’s strengths include its detailed and accurate description, along with information about the
541 environment and its potential uses.

542 Claude 3 and GPT-4 Turbo excel in providing detailed, accurate, and coherent descriptions, closely
543 aligning with the ground truth. Their responses demonstrate a strong understanding of the scene,
544 balancing specific visual elements with broader contextual insights. Fuyu-8B and Qwen-VL offer
545 valuable observations but fall short of delivering comprehensive descriptions. This analysis highlights
546 the importance of contextual understanding in multimodal models, as demonstrated by Claude 3 and
547 GPT-4 Turbo.

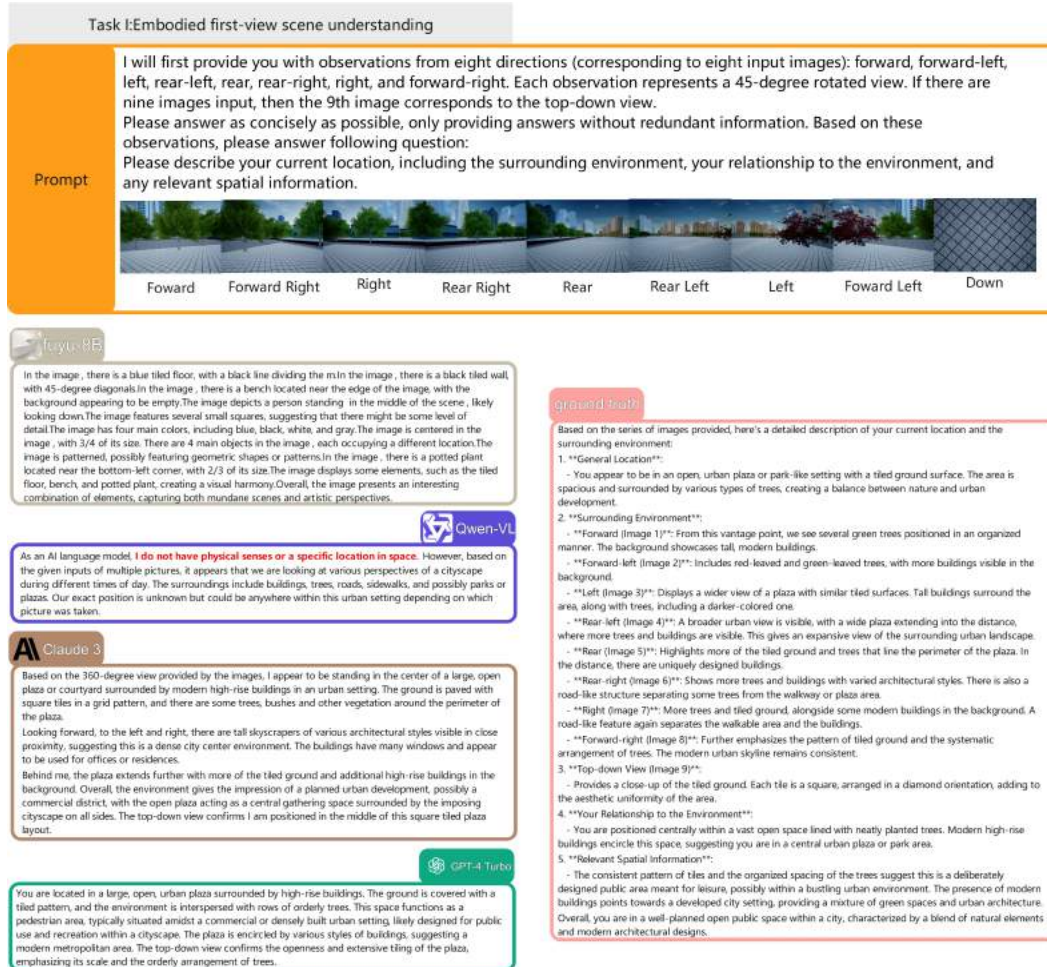


Figure 8: Embodied first-view scene understanding task involves describing one’s current location, surrounding environment, relationship to the environment, and any relevant spatial information based on observations from eight directions (forward, forward-left, left, rear-left, rear, rear-right, right, and forward-right) and one top-down view image. The specific outputs of different methods are listed separately.

548 A.6.2 Task II: Embodied question answering

549 As presented in Figure 9, fuyu-8B responded by identifying three traffic lights visible in the images.
550 However, this response is inaccurate according to the ground truth, which states that only two traffic
551 lights are present. This overestimation indicates a potential issue with embodied recognition or
552 differentiation in Fuyu-8B.

553 Qwen-VL asserted that there are no traffic lights visible in any of the provided images. This response
554 is also incorrect, as it fails to recognize the two traffic lights that are present. This suggests a limitation
555 in Qwen-VL’s ability to detect specific objects accurately in a multimodal context.

556 Claude 3 similarly concluded that there are no traffic lights observable in the images. This response,
557 like that of Qwen-VL, indicates a failure in object detection capabilities, as it overlooks the traffic
558 lights that are present.

559 GPT-4 Turbo, on the other hand, correctly identified that there are two traffic lights in total. This
560 response aligns with the ground truth, demonstrating GPT-4 Turbo’s superior ability to accurately
561 recognize and count specific objects within the provided visual context.

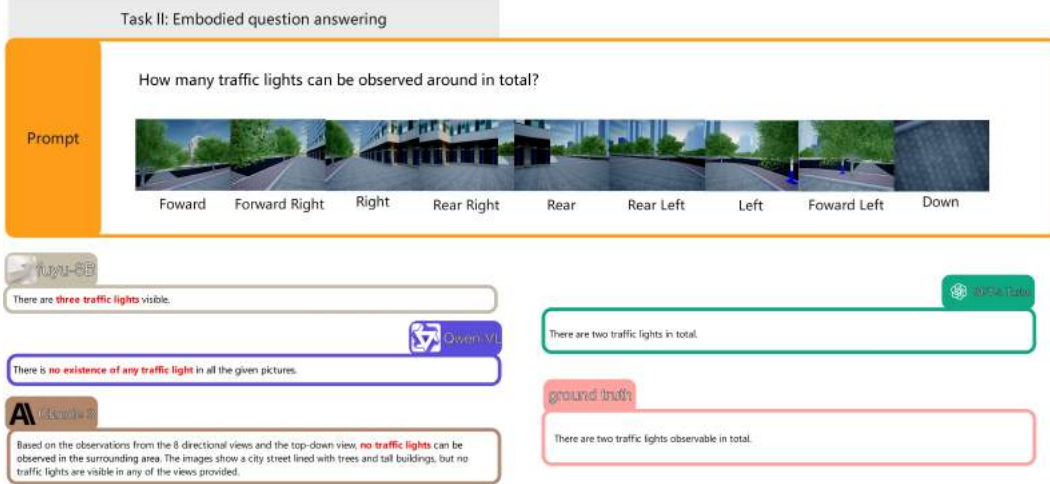


Figure 9: This case of embodied question answering task involves answering the question "How many traffic lights can be observed around in total?" based on images from eight directions (forward, forward-right, right, rear-right, rear, rear-left, left, forward-left) and one top-down view. The original outputs of different models are listed separately.

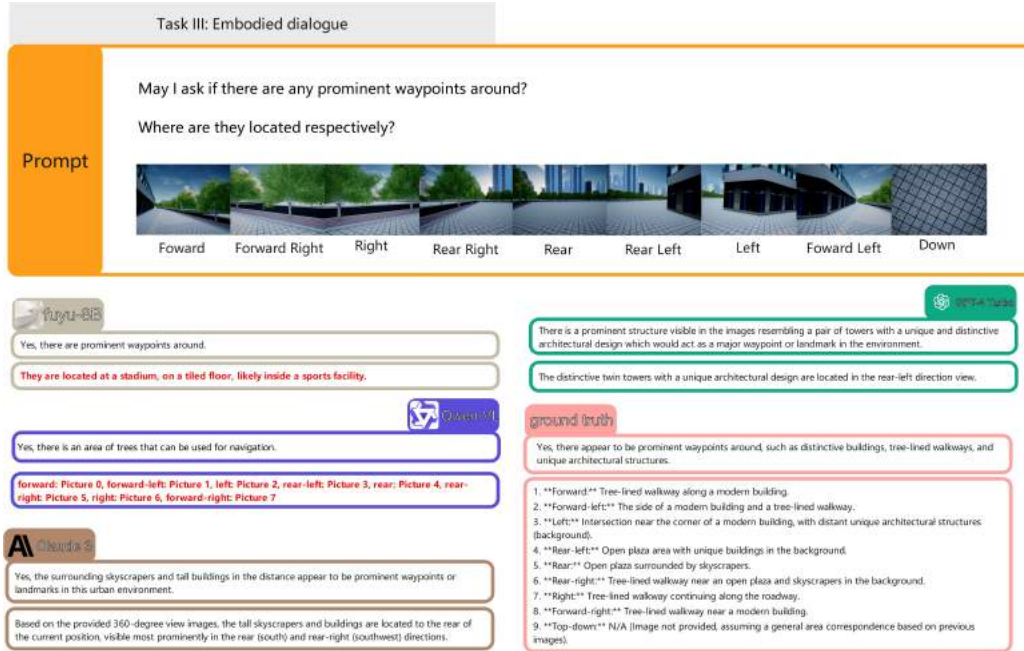


Figure 10: This embodied dialogue case involves a dialogue where the models are asked to identify any prominent waypoints around and specify their locations based on a series of images taken from eight different directions (forward, forward-right, right, rear-right, rear, rear-left, left, forward-left) and one top-down view.

562 The accuracy of the responses varies significantly among the models. GPT-4 Turbo stands out as
 563 the only model to provide the correct answer, reflecting its strong performance in visual recognition
 564 and comprehension tasks. In contrast, Fuyu-8B overestimates the number of traffic lights, while
 565 Qwen-VL and Claude 3 both fail to recognize any traffic lights, highlighting potential limitations in
 566 their object detection capabilities.

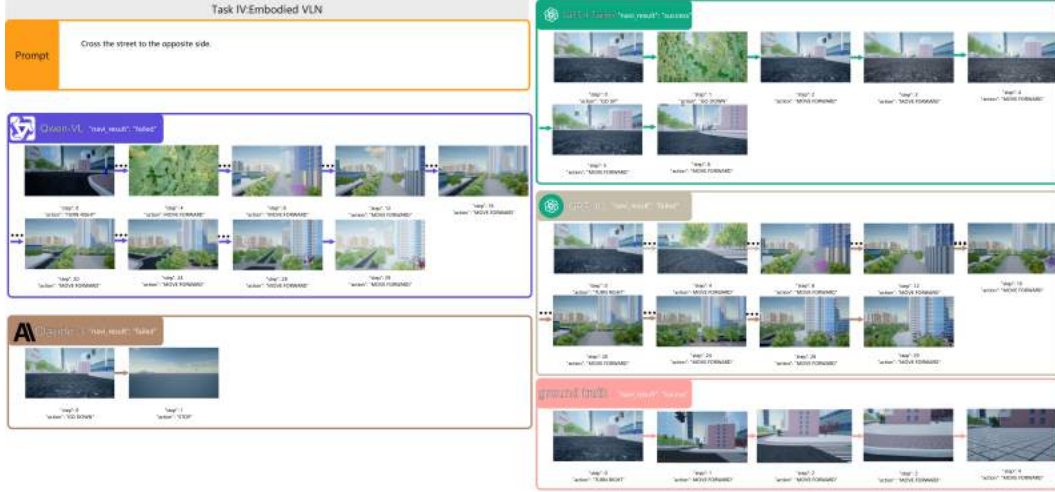


Figure 11: The agent should decide the action according to the observations until arriving the goal.

A.6.3 Task III: Embodied dialogue

The models' performances are compared to the ground truth to evaluate their accuracy and descriptive capabilities, as shown in Figure 10. Fuyu-8B acknowledges the presence of prominent waypoints but incorrectly identifies their nature and location. It states that the waypoints are located at a stadium on a tiled floor, likely inside a sports facility. This response is inaccurate as it fails to mention the actual prominent waypoints, such as distinctive buildings or structures, and does not align with the ground truth. Qwen-VL identifies an area of trees that can be used for navigation but fails to recognize the prominent architectural structures. It provides a detailed list of all image directions but does not accurately describe the waypoints. This response is partially correct in identifying natural features but misses the crucial architectural landmarks highlighted in the ground truth. Claude 3 asserts that surrounding skyscrapers and tall buildings serve as prominent waypoints or landmarks in the urban environment. It accurately pinpoints the locations of these landmarks, noting their visibility in the south and southwest directions relative to the current position. This response closely aligns with the ground truth, demonstrating a good understanding of the urban landscape and identifying the correct waypoints. GPT-4 Turbo provides a precise and accurate description, identifying a prominent structure resembling twin towers with unique architectural design as major waypoints. It specifies that these towers are visible in the rear-left direction. This response most accurately reflects the ground truth, which mentions distinctive buildings and tree-lined walkways as prominent waypoints.

Among the models, GPT-4 Turbo provides the most accurate and descriptive response, closely aligning with the ground truth by identifying the twin towers as prominent waypoints. Claude 3 also offers a strong response by correctly identifying the surrounding skyscrapers and their specific locations. In contrast, Fuyu-8B and Qwen-VL fail to accurately identify the architectural landmarks, highlighting the need for improvement in their embodied ability to recognize and describe complex urban environments.

A.6.4 Task IV: Embodied VLN

In order to compare different models on the VLN task, we give a detail case in Figure 11. The analysis reveals that only GPT-4 Turbo successfully completes the task, suggesting it has a superior capability in interpreting and navigating based on RGB observations. Both Qwen-VL and GPT-4o show similar patterns of failure, indicating potential areas for improvement in their navigation algorithms. Claude 3's failure highlights a critical need for enhancement in its initial perception and decision-making processes. The ground truth provides a clear and effective navigation path, demonstrating the importance of precise and context-aware actions in achieving the objective.

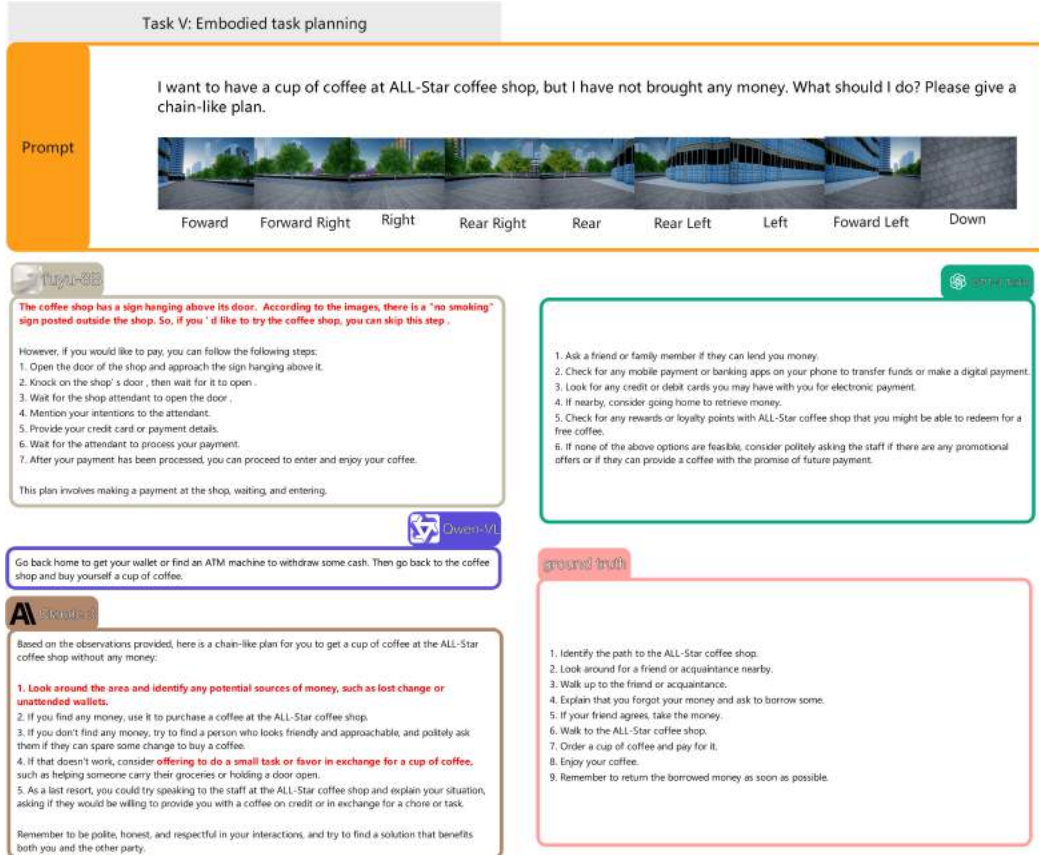


Figure 12: This case of embodied task planning involves creating a chain-link plan to get a cup of coffee from the ALL-Star coffee shop without having brought any money. The AI models are asked to provide a step-by-step plan based on a series of images taken from eight different directions (forward, forward-right, right, rear-right, rear, rear-left, left, forward-left) and one top-down view.

599 A.6.5 Task V: Embodied task planning

600 As shown in Figure 12, Fuyu-8B's response focuses on a detailed description of the coffee shop,
 601 mentioning a "no smoking" sign. It then provides a procedure involving opening the door, waiting for
 602 the shop to open, mentioning intentions to the attendant, and providing payment details. This plan is
 603 not practical as it assumes the user has money or a payment method, which contradicts the prompt's
 604 condition of not having brought any money. Qwen-VL suggests going back home to get money or
 605 finding an ATM to withdraw cash before returning to the coffee shop. While this response is practical,
 606 it lacks creativity and does not explore alternative solutions available in the immediate environment,
 607 making it less optimal than the ground truth. Claude 3 provides a detailed and creative plan, which is
 608 practical, creative, and aligns well with the ground truth, addressing the situation effectively without
 609 requiring the user to leave the area. Similar to Claude 3, GPT-4 Turbo's response is practical and
 610 creative, providing several feasible options without needing to leave the vicinity, and aligns well with
 611 the ground truth.

612 Claude 3 and GPT-4 Turbo provide the most practical and creative solutions, closely aligning with the
 613 ground truth. They explore multiple options to solve the problem without requiring the user to leave
 614 the immediate area. Fuyu-8B's response is less practical as it does not address the lack of money, and
 615 Qwen-VL's solution, while practical, lacks creativity and does not leverage immediate resources.