BEAR: Benchmarking and Enhancing Multimodal Language Models for Atomic Embodied Reasoning Abilities

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Embodied reasoning abilities refer to the capabilities for agents to perceive, comprehend, and interact effectively with the physical world. While multimodal large language models (MLLMs) show promise as embodied agents, a thorough and systematic evaluation of their embodied reasoning capabilities remains underexplored, as existing benchmarks primarily focus on isolated domains such as planning or spatial understanding. To bridge this gap, we propose BEAR, a comprehensive and fine-grained benchmark designed to evaluate MLLM's atomic embodied reasoning abilities. BEAR comprises 4,469 interleaved video–image–text entries across 14 skills in 6 categories, including tasks from low-level pointing, trajectory understanding, spatial reasoning, to high-level planning. Evaluation results of 20 state-of-the-art MLLMs reveal their persistent limitations across all categories of embodied reasoning. Moreover, our failure analysis indicates that fine-grained visual reasoning and spatial reasoning remain major bottlenecks, underscoring key directions for future improvement in MLLMs.

1 Introduction

In artificial intelligence, embodied agents are systems that perceive and interact meaningfully with environments through grounded understandings of the physical world [8]. To accomplish a task, an agent must perform a systematic set of visual reasoning skills: from low-level perception and localization, such as pointing to recognize objects, through trajectory reasoning to predict dynamic motion, 3D spatial reasoning for navigation, and ultimately high-level planning to decompose a task into structured steps. Together, these hierarchical skills constitute the foundation of embodied reasoning, which enables agents to act robustly in physical environments [9, 7].

Multimodal large language models (MLLM) [11, 1] have emerged as promising solutions to build embodied agents, and many benchmarks are proposed to evaluate their potential. These fall into two main categories. The first uses offline VQA-style inputs but focuses narrowly on isolated abilities, such as pointing [19, 20], spatial reasoning [17, 14], planning [16]. The second evaluates MLLMs in simulation [18, 12] and measures the overall task success rate without skill-level decomposition, making it unclear which reasoning skills drive performance. Both categories lack holistic evaluation of fine-grained categories of different embodied reasoning skills.

These limitations motivate two fundamental questions: *(1) To what extent do current MLLMs possess embodied reasoning abilities (2) what factors constrain their performance?*
To address these questions, we propose BEAR, short for Benchmarking Embodied Atomic Reasoning, the first benchmark to unify embodied reasoning into 6 categories and 14 atomic skills, all framed under a consistent VQA-style format. It comprises 4,469 unique interleaved image–video–text entries, providing a comprehensive and systematic evaluation of embodied reasoning. Additionally,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

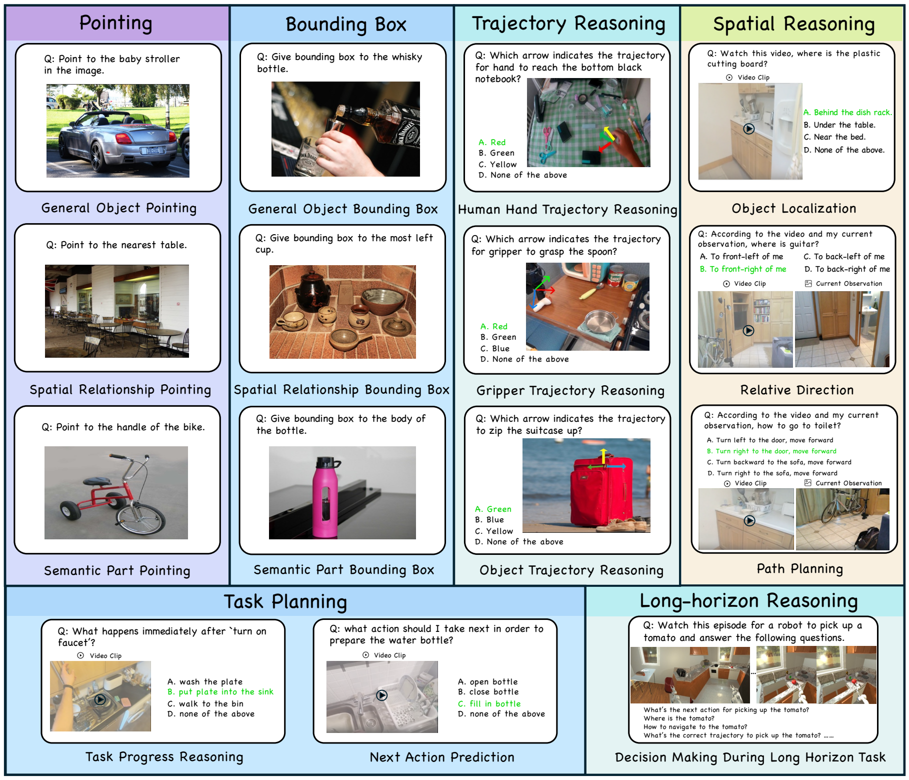


Figure 1: Overview of the BEAR Benchmark.

we introduce a long-horizon category including episodes from simulation where an agent completes a full task (e.g., setting a table). Each episode is decomposed into atomic reasoning steps aligned with our taxonomy, demonstrating that our taxonomy is both cognitively motivated and grounded in embodied task execution. We evaluate 15 representative MLLMs on BEAR, as shown in Table 1, and conduct a thorough failure analysis. The results reveal two key findings: (1) Most current MLLMs exhibit weak embodied reasoning abilities, ranging from low-level pointing to high-level planning, with closed-source models generally outperforming open-source ones. (2) Fine-grained visual reasoning and 3D reasoning abilities remain major bottlenecks—models struggle to perceive subtle visual details, translate visual inputs into dynamic motions or human activities, and understand 3D spatial layout based on 2D observations.

In summary, our contributions are listed as follows:

1.We introduce BEAR, the first comprehensive benchmark that unifies embodied reasoning into 6 categories and 14 atomic skills, with 4,469 image–video–text entries.

2. Our evaluation and error analysis reveal key failure modes in MLLMs and highlight directions for improving MLLMs on embodied reasoning abilities.

2 The BEAR Benchmark

2.1 Overview of BEAR

We introduce BEAR, the first unified fine-grained embodied reasoning benchmark with 4,469 image, video, and text VQA entries spanning 6 categories and 14 atomic skills, as shown in Fig. 1. Detailed statistics and category distribution are reported in Fig. 2 and Fig. 3.

2.2 Data Collection and Curation Process

Statistic	Number
Total questions	4,469
- with only one image	2,886 (64.6%)
- with only one video	995 (22.2%)
- with interleaved data	588 (13.2%)
Number of multiple-choice questions	2,563 (57.4%)
Number of free-form questions	1,906 (42.6%)
Unique number of images	2,079
Unique number of videos	918
Category number	6
Subtype number	15
Maximum question word count	82
Maximum choice word count	15.9
Average question word count	20
Average choice word count	3.7

Figure 2: Key statistics.

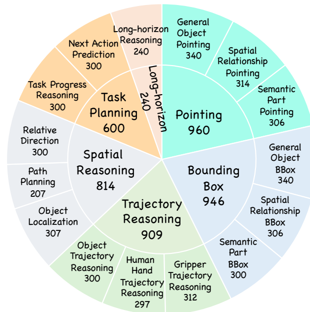


Figure 3: Category distribution.

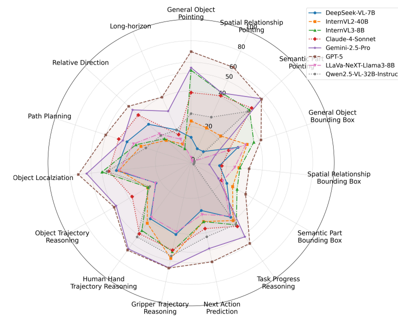


Figure 4: Evaluation on Radar Map.

	Pointing				Bounding Box				Task Planning		
	GEN	SPA	PRT	Avg	GEN	SRA	PRT	Avg	PRG	PRD	Avg
Random Choice	-	-	-	-	-	-	-	-	25	25	25
Open-source Models											
DeepSeek-VL-7B [13]	14.12	8.50	9.24	10.62	0.276	0.160	0.231	0.222	37.67	27.33	32.50
InternVL2-4B [4]	18.53	10.78	12.42	13.91	0.117	0.082	0.107	0.102	37.33	32.33	34.83
InternVL2-8B [4]	21.18	21.90	21.97	21.68	0.294	0.194	0.179	0.222	**44.00**	31.67	37.84
InternVL2-26B [4]	21.18	15.36	18.79	18.44	0.201	0.202	0.147	0.183	41.33	34.33	37.83
InternVL2-40B [4]	23.24	21.24	22.29	22.25	0.329	0.269	0.268	0.289	40.00	33.67	36.84
InternVL3-8B [21]	**52.65**	**42.48**	**43.95**	**46.36**	**0.369**	**0.275**	**0.297**	**0.314**	43.00	33.67	38.34
InternVL3-14B [21]	37.94	27.78	32.80	32.84	0.304	0.258	0.276	0.279	41.00	33.00	37.00
LLaVa-NeXT-Llama3-8B [10]	2.94	1.31	0.96	1.73	0.320	0.246	0.205	0.257	36.67	29.67	33.17
Qwen2.5-VL-7B-Instruct [3]	6.18	1.63	0.96	2.92	0.007	0.003	0.009	0.007	40.67	32.33	36.50
Qwen2.5-VL-32B-Instruct [3]	27.35	27.78	42.68	32.60	0.020	0.018	0.017	0.018	42.67	**42.33**	**42.50**
Proprietary Models											
Claude-3.7-Sonnet [2]	47.94	36.27	37.58	40.60	0.195	0.132	0.187	0.171	32.67	44.33	38.50
Claude-4-Sonnet [2]	39.12	40.86	45.54	41.84	0.221	0.173	0.197	0.197	44.00	37.67	40.84
Gemini-2.5-Flash [5]	46.76	33.33	39.49	39.86	0.183	0.145	0.156	0.161	48.33	43.67	46.00
Gemini-2.5-Pro [5]	55.00	42.48	**55.41**	50.96	0.144	0.103	0.177	0.141	52.00	49.00	50.50
GPT-5 [15]	**70.00**	**63.69**	54.90	**62.86**	**0.411**	**0.326**	**0.352**	**0.363**	**59.67**	**61.00**	**60.34**

	Trajectory				Spatial Reasoning				Long-horizon
	GPR	HND	OBJ	Avg	LOC	PTH	DIR	Avg	-
Random Choice	25	25	25	25	25	50	25	25	25
Open-source Models									
DeepSeek-VL-7B [13]	41.03	38.72	22.67	34.14	42.02	**37.68**	**32.00**	**37.23**	20.00
InternVL2-4B [4]	44.55	34.01	25.67	34.74	40.07	33.82	26.33	33.41	8.57
InternVL2-8B [4]	41.67	38.38	22.33	34.13	39.41	29.95	25.33	31.56	11.49
InternVL2-26B [4]	53.21	43.77	**30.33**	42.44	26.06	26.57	22.00	24.88	11.29
InternVL2-40B [4]	**57.69**	41.75	28.00	42.48	40.39	29.47	18.67	29.51	11.43
InternVL3-8B [21]	51.28	46.80	27.67	41.92	**50.16**	32.37	20.00	34.18	8.57
InternVL3-14B [21]	51.28	49.49	31.43	43.36	43.00	28.02	21.33	30.78	**28.57**
LLaVa-NeXT-Llama3-8B [10]	39.42	37.71	23.00	33.38	40.39	33.82	24.00	32.74	14.29
Qwen2.5-VL-7B-Instruct [3]	54.49	48.15	30.00	44.21	38.44	31.40	21.00	30.28	22.86
Qwen2.5-VL-32B-Instruct [3]	55.45	**52.19**	26.67	**44.77**	47.23	26.57	22.67	32.16	20.00
Proprietary Models									
Claude-3.7-Sonnet [2]	52.88	48.82	31.33	44.34	38.76	33.33	34.67	35.59	20.00
Claude-4-Sonnet [2]	50.00	49.16	38.00	45.72	46.25	42.51	39.67	42.81	17.14
Gemini-2.5-Flash [5]	64.42	63.97	45.00	57.80	61.24	43.00	44.67	49.64	31.43
Gemini-2.5-Pro [5]	66.67	65.99	48.33	60.33	64.50	40.10	44.00	49.53	31.43
GPT-5 [15]	**66.99**	**67.34**	**49.67**	**61.33**	**72.31**	**50.24**	**47.00**	**51.52**	**40.00**

Table 1: **Evaluation results on BEAR**. We evaluate 15 MLLMs on BEAR using direct prompting format without reasoning chains. GEN = General Object (Pointing/Box); SPA = Spatial Object (Pointing/Box); PRT = Semantic Part (Pointing/Box); PRG = Task Progress Reasoning; PRD = Next Action Prediction; GPR = Gripper Trajectory Reasoning; HND = Human Hand Trajectory Reasoning; OBJ = Object Trajectory Reasoning; LOC = Object Localization; PTH = Path Planning; DIR = Relative Direction.

Categorization in BEAR is thoughtfully designed. To evaluate MLLMs on embodied reasoning, we define five core categories: Pointing, Bounding Box Localization, Trajectory Reasoning, Spatial Reasoning, and Task Planning, which align with both human cognition process and task structures in robotics. In addition, the Long-horizon category verifies the soundness of our benchmark by decomposing each task into structured reasoning steps, with each step mapped to a reasoning skill in other categories.

Curation and VQA Generation Process. We adopt a category-specific data generation process, combining automated scripts with human annotation. This hybrid strategy also incorporates manual difficulty control to ensure qualified, balanced and reliable evaluation.

3 Experiments

Experiment setup and experiment result. Our evaluation includes 15 distinct MLLMs, as shown in Table 1. For most models, we follow the standard evaluation protocol outlined by the VLMEvalKit [6] contributors. We adopt a direct prompting strategy, where the MLLM is asked to produce an answer directly without intermediate reasoning steps.

MLLMs remain limited across all embodied reasoning categories. Figure 5 shows that most MLLMs achieve only 20% to 40% average performance. Even the strongest model, GPT-5 [15], reaches only 55.52%, indicating substantial space for improvement in MLLMs on embodied reasoning tasks.

Proprietary models generally outperforms open-sourced models As shown in Figure 5, proprietary models achieve significantly higher overall performance than open-source ones, with an average score of 40.48% compared to 27.17%. GPT-5 [15] leads with 52.06%, followed by Gemini-2.5-Pro and Gemini-2.5-Flash at 42.81% and 40.14%, respectively. In contrast, most open-source models remain below 35%, underscoring the performance gap between the two groups and highlighting substantial room for further advancement in embodied reasoning.

Fine-grained visual reasoning abilities is the major bottle neck for perception and trajectory reasoning tasks. As illustrated in Figure 6, models are often able to reason about and localize the approximate region of the target object, yet they frequently fail to pinpoint the exact location. This limitation becomes even more pronounced in trajectory reasoning, where the inability to reliably identify the precise target object and to infer the correct direction of motion severely constrains model performance. These challenges suggest that improving fine-grained visual reasoning abilities is critical for advancing perception and trajectory reasoning capabilities.

3D spatial reasoning is the major bottleneck for spatial reasoning tasks. As shown in Figure 7, most path planning errors arise from 3D and direction reasoning, showing that MLLMs struggle to estimate scene geometry and perceive their own orientation. While models can detect relevant objects, they often misjudge depth, spatial layout, or directional relations, underscoring that robust spatial grounding remains a major challenge for embodied reasoning.

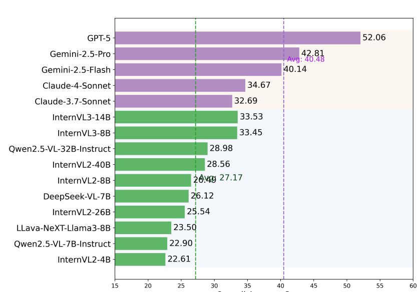


Figure 5: Open-sourced v.s. Proprietary Models

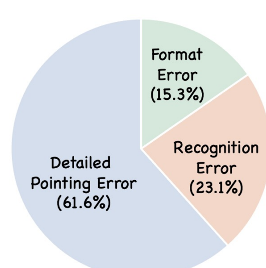


Figure 6: Pointing error analysis.

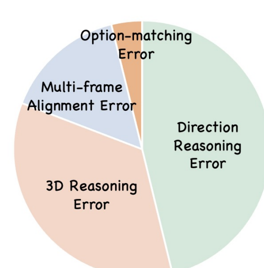


Figure 7: Path Planning error analysis.

References

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

[2] Anthropic. Claude 3 Model Card. `https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf`, 2024. Accessed: 2025-08-23.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

[6] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. *URL https://arxiv. org/abs/2407.11691*, 7.

[7] 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.

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

[9] Li Kang, Xiufeng Song, Heng Zhou, Yiran Qin, Jie Yang, Xiaohong Liu, Philip Torr, Lei Bai, and Zhenfei Yin. Viki-r: Coordinating embodied multi-agent cooperation via reinforcement learning. *arXiv preprint arXiv:2506.09049*, 2025.

[10] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. URL `https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/`.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[12] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024.

[13] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[14] Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025.

[15] OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025.

[16] Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.

[17] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[18] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

[19] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

[20] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

[21] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Figure 8: **Unified Benchmark Data Format.** All our data adheres to a consistent format across tasks. For example, in an object localization instance, fields that are not applicable are left blank.

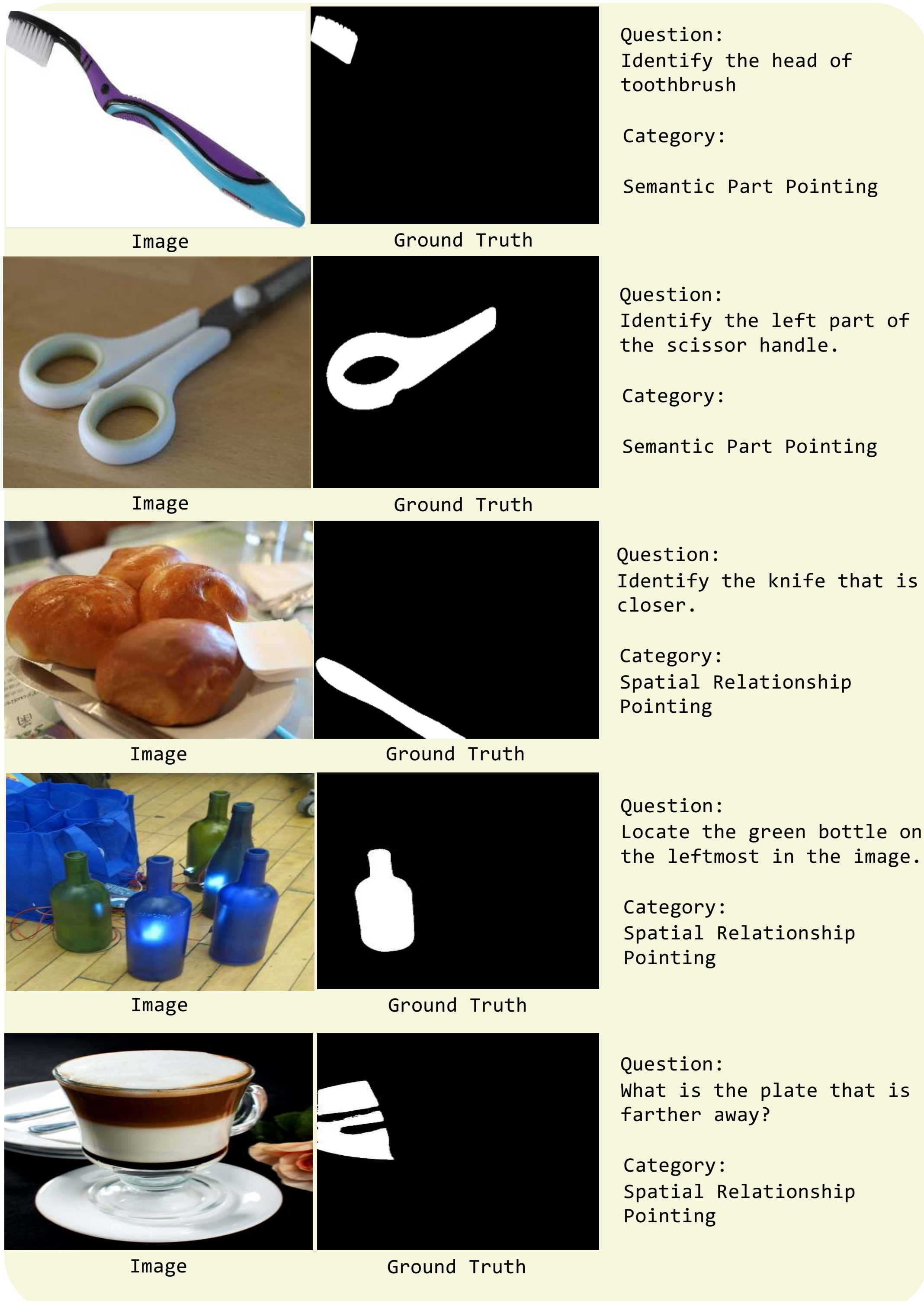


Figure 9: **Benchmark Examples**

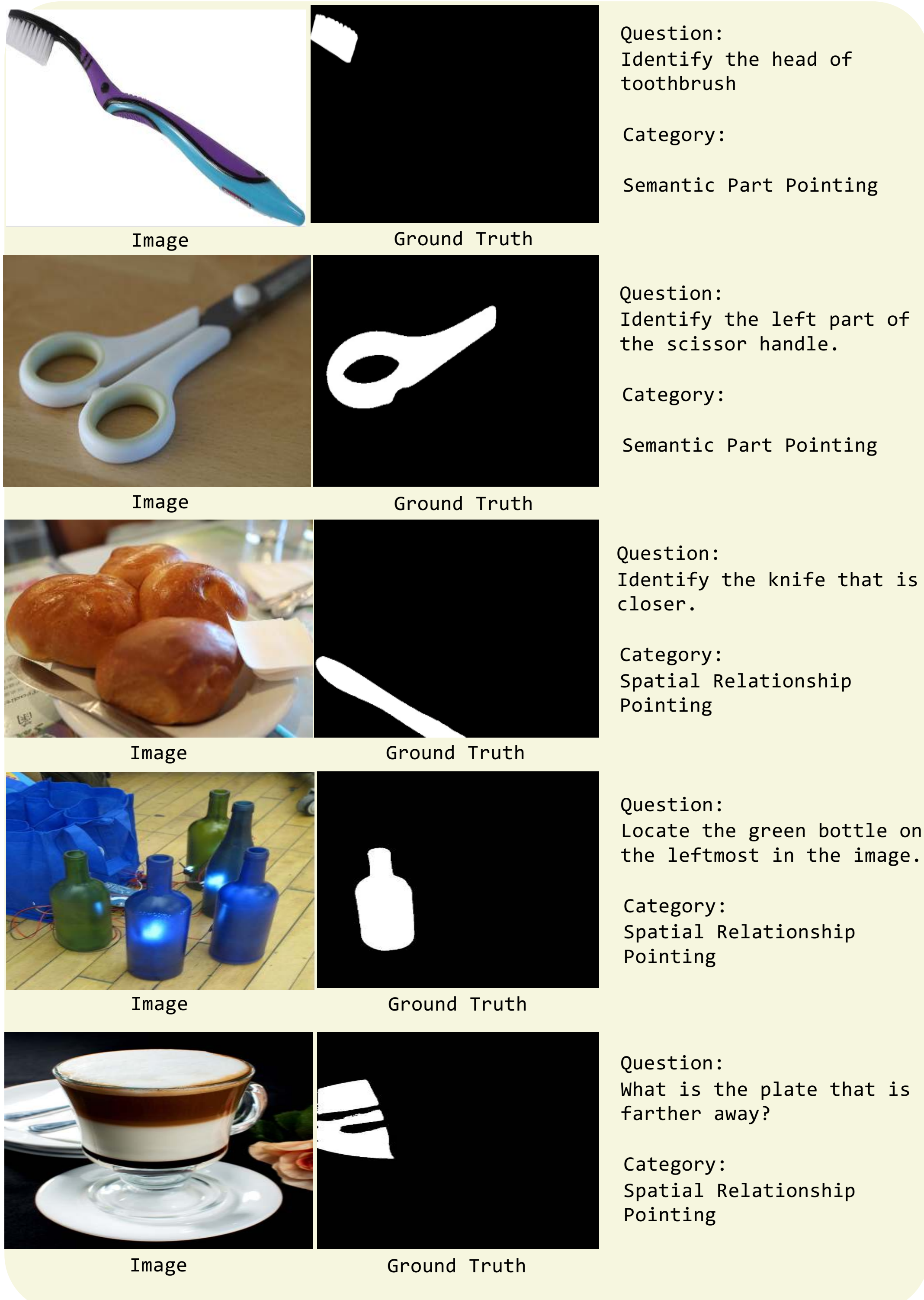
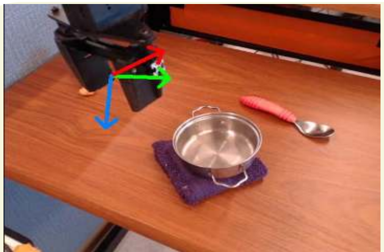


Figure 10: **Benchmark Examples**



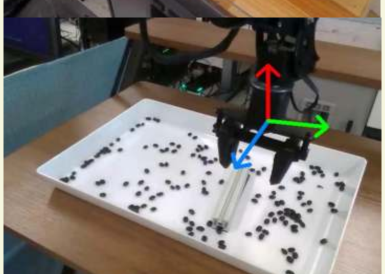
```
Question:
which arrow should the robot follow to move
toward the **spatula**?
A. Green
B. Blue
C. Red
D. None of the above       Ground Truth: A
```

```
Question:
which arrow should the robot follow to move
toward the **vessel**?
A. Green
B. Blue
C. Red
D. None of the above       Ground Truth: A
```

```
Question:
which arrow should the robot follow to
move toward the **fork**?
A. Green
B. Blue
C. Red
D. None of the above       Ground Truth: B
```

```
Question:
which arrow should the robot follow to
move toward the **yellow cloth**?
A. Green
B. Blue
C. Red
D. None of the above       Ground Truth: A
```

```
Question:
which arrow should the robot follow to
move toward the **blue brick**?
A. Green
B. Blue
C. Red
D. None of the above       Ground Truth: D
```



```
Question:
which arrow should the robot follow to
move toward the **sweep**?
A. Green
B. Blue
C. Red
D. None of the above       Ground Truth: B
```

Figure 11: **Benchmark Examples**

Question:
which arrow should the hand follow to move toward the **watering can**?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: C

Question:
Which direction should you move in to close the cabinet?
A. Red
B. Green
C. Yellow
D. None of the above

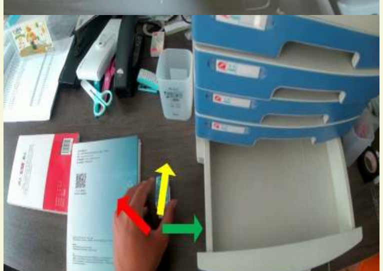
Ground Truth: A

Question:
which direction is the hand most likely to place the dish cloth on the black rack?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: C

Question:
which arrow indicates the correct direction to clean the surface of this soap box?
A. Green
B. Blue
C. Red
D. None of the above

Ground Truth: A



Question:
which direction is the hand most likely to place the blue stapler inside the open drawer on the right of the hand?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: B

Question:
which direction is the hand most likely to move if you want to use the knife to stab the small white plate?
A. Green
B. Blue
C. Red
D. None of the above

Ground Truth: C

Figure 12: **Benchmark Examples**

Question:
which arrow indicates the direction in which the hand will be moved to pull out the drawer?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: A

Question:
Which arrow best represents the hand's movement to rotate the handle downwards?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: B

Question:
Which arrow indicates the direction the hand will take to take the milk bottle out?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: B

Question:
Identify the arrow that indicates the direction the hand will rotate to unlock the pump
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: A

Question:
Which arrow indicates the direction the hand should move to lift the cap of the bottle?
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: A

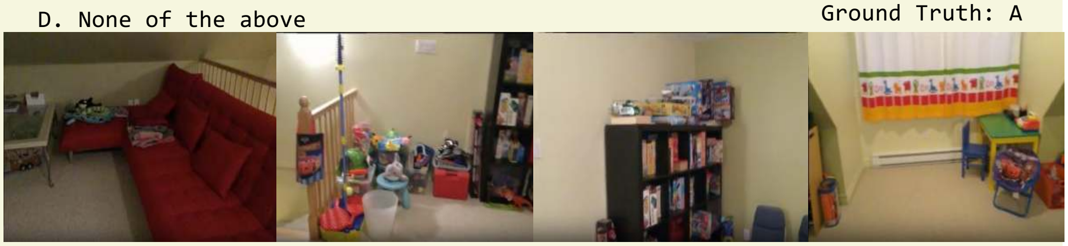
Question:
Identify the arrow that indicates the direction the hand will move to open the microwave door.
A. Red
B. Green
C. Yellow
D. None of the above

Ground Truth: C

Figure 13: **Benchmark Examples**

Which description of following about the white plastic cutting board is true according to the video given?
A. Behind the dish rack near the sink.
B. On the stove beside the pots
C. Hanging on the wall above the counter
D. None of the above
Ground Truth: A

Which description of following about the mini soccer ball toy is true according to the video given?
A. On the top left shelf inside the yellow bin
B. On the floor near the white trash bin
C. On the blue stool next to the table
D. None of the above
Ground Truth: A



Which description of following about the large blue bag is true according to the video given?
A. Next to the television stand against the wall
B. On top of the glass coffee table
C. Beside the red sofa
D. None of the above
Ground Truth: A

Which description of following about the book next to the plant is true according to the video given?
A. On the floor near the gray carpet
B. On the sofa near the yellow cushion
C. On the black shelf
D. None of the above
Ground Truth: C

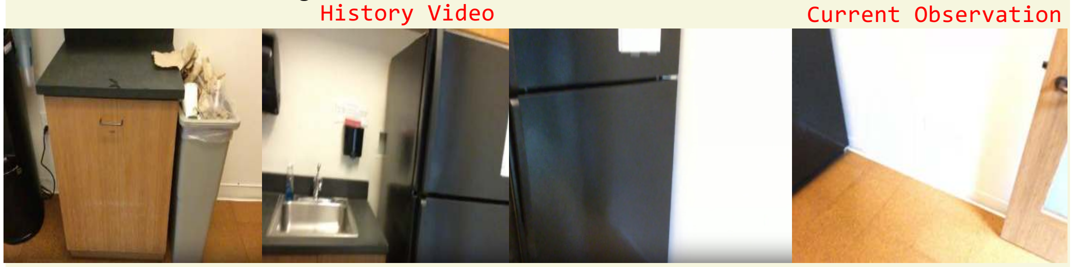
Figure 14: **Benchmark Examples**

According to the current observation, where is the kitchen counter?
A. To the front-right of me.
B. To the front-left of me.
C. To the back-left of me.
D. To the back-right of me.

Ground Truth: B

History Video

Current Observation



Where is the coffee table?
A. To the front-right of me.
B. To the front-left of me.
C. To the back-left of me.
D. To the back-right of me.

Ground Truth: C

History Video

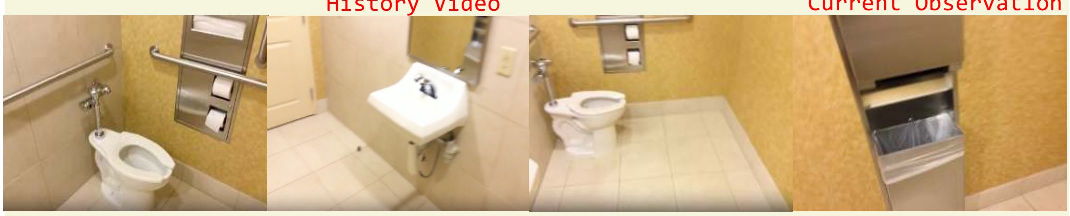
Current Observation

Where is the toilet?
A. To the front-right of me.
B. To the front-left of me.
C. To the back-left of me.
D. To the back-right of me.

Ground Truth: D

History Video

Current Observation



Where is the blue box?
A. To the front-right of me.
B. To the front-left of me.
C. To the back-left of me.
D. To the back-right of me.

Ground Truth: B

History Video

Current Observation

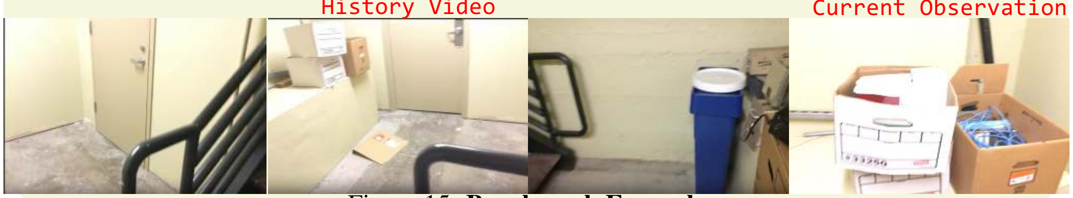


Figure 15: **Benchmark Examples**

You want to navigate to the toilet. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. Go forward until the TV 2. [please fill in] 3. Go forward until the shower 4. [please fill in] 5. Go forward until the toilet. You have reached the final destination.

A. Turn Back, Turn Left
B. Turn Left, Turn Left
C. Turn Left, Turn Right
D. Turn Right, Turn Right

Ground Truth: C

You want to navigate to the trash bin. You will perform the following actions (Note: for each [please fill in], choose either 'turn back,' 'turn left,' or 'turn right.'): 1. [please fill in] 2. Go forward until the cabinet 3. [please fill in] 4. Go forward until the trash bin is on your right. You have reached the final destination.

A. Turn Left, Turn Left
B. Turn Right, Turn Left
C. Turn Back, Turn Left
D. Turn Right, Turn Right

Ground Truth: B

Figure 16: **Benchmark Examples**

Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to prepare meat for cooking?

A. cut meat

B. throw cover

C. walk to the trash bin

D. none of the above

Ground Truth: A

Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to fold and put away bag?

A. close drawer

B. pick up bag

C. walk to the drawer

D. none of the above

Ground Truth: A

Considering the progress shown in the video and my current observation in the last frame, what action should I take next in order to wash and rinse various kitchen utensils and dishes?

A. wash spoon

B. walk to the measuring cup

C. put down measuring cup

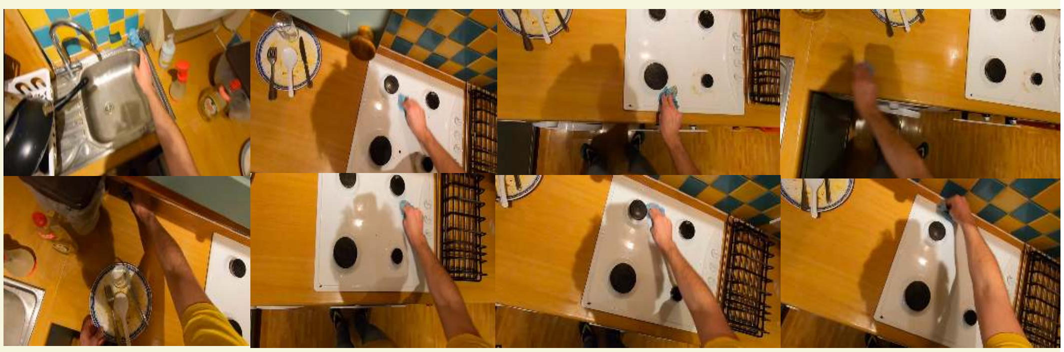
D. none of the above

Ground Truth: D

Figure 17: **Benchmark Examples**

Which action does not happen before 'put away raisins'
A. open drawer
B. pour cereal
C. open fridge
D. none of the above
Ground Truth: C

Which of the following actions is not performed after 'pick up plate’?
A. wipe hob
B. put down plate
C. turn off tap
D. none of the above
Ground Truth: C



What action occurs immediately after drying the pot?
A. put down cloth
B. pick up pot
C. open drawer
D. none of the above
Ground Truth: A

Figure 18: **Benchmark Examples**