

Appendix

I. Our Position and Advantages

Position. We position our work as a critical bridge between two well-studied paradigms (Fig. II a): (1) Visual reasoning on 2D images, which lacks any notion of active perception or physical action. (2) Embodied reasoning in 3D space, which requires a complete 3D simulator. This positioning makes our contribution *modular and composable*. Our visual search policy is a direct and natural precursor to a downstream locomotion policy. The agent can first use our method to decide where to go at a junction, and then execute the movement. *In a word, our work provides a foundational layer for building generalized embodied agents across diverse human-centric environments.*

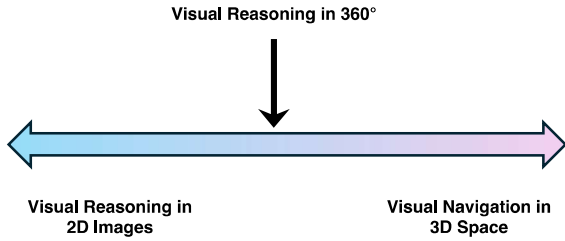


Figure I. Illustration of our position.

Advantages. While prior work typically needs 3D simulators that are hard to scale and limited to household scenes, we bypass the sim-to-real gap via real-world 360 panoramas, and enable the scalable study of visual reasoning across complex in-the-wild scenes. Most importantly, *our work fills the active vision gap for open-world embodied AI, and our dataset also serves as a valuable resource for building in-the-wild 3D simulators.*

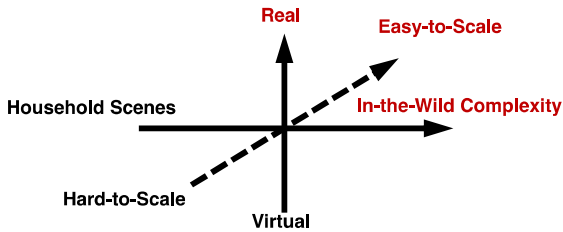


Figure II. Illustration of our advantages.

II. Geographical Distribution of H^* Bench

As shown in Fig. III, our dataset exhibits broad geographical coverage, including thirteen cities across twelve countries and four continents. This diversity is reflected in the wide range of architectural styles, languages and scripts on signage, and environmental conditions.

III. HVS Task Difficulty Visualization

III.1. HOS Difficulty Visualization

Figure V illustrates the three difficulty levels of our *HOS* task with concrete examples. For each keyframe, we overlay the target object’s area on the initial view and provide its visibility ratio d . From left to right, we show one example per level: **Easy** (mostly visible), **Medium** (partially visible), and **Hard** (invisible), demonstrating the model’s initial observation.

III.2. HPS Difficulty Visualization

We define the four difficulty levels of the *HPS* task based on two criteria: whether the relevant cue aligns with the navigable path and whether textual information is provided. Examples are shown in Fig. VI-Fig. IX.

IV. Training and Inference Prompts

The natural-language prompts used for training and inference of the *HOS* and *HPS* tasks are shown in Fig. IV.

V. Objective Functions

SFT stage. The SFT objective function is the expected negative log-likelihood (cross-entropy loss) over the dataset \mathcal{D}^{SFT} which consists of task input x and labeled trajectory \mathcal{H}_T :

$$\min_{\theta} \mathbb{E}_{(x, \mathcal{H}_T) \sim \mathcal{D}^{SFT}} \left[- \sum_{i=0}^{T-1} \log \pi_{\theta}(y_i, a_i \mid o_i, x, \mathcal{H}_i) \right].$$

RL stage. For each task, we sample G times and get outputs $\{\omega_1, \omega_2, \dots, \omega_G\}$ where ω_i includes all the output tokens in the output sequence $\{y_0, a_0, y_1, a_1, \dots, y_{T-1}, a_{T-1}\}$, then calculate the GRPO advantage to update the parameters. The GRPO objective function is:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = & \mathbb{E}[(s_o, x, y) \sim \mathcal{D}^{RL}, \{\omega_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\Omega \mid s_o, x)] \\ & \frac{1}{G} \sum_{i=1}^G \frac{1}{|\omega_i|} \sum_{t=1}^{|\omega_i|} \\ & \left\{ \min \left[\frac{\pi_{\theta}(\omega_{i,t} \mid s_o, x, \omega_{i,<t})}{\pi_{\theta_{\text{old}}}(\omega_{i,t} \mid s_o, x, \omega_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(\omega_{i,t} \mid s_o, x, \omega_{i,<t})}{\pi_{\theta_{\text{old}}}(\omega_{i,t} \mid s_o, x, \omega_{i,<t})} \right. \right. \right. \\ & \left. \left. \left. , 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{KL}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right\}, \end{aligned}$$

where $\hat{A}_{i,t}$ denotes the group relative advantage at $\omega_{i,t}$:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(r)}{\text{std}(r)}.$$

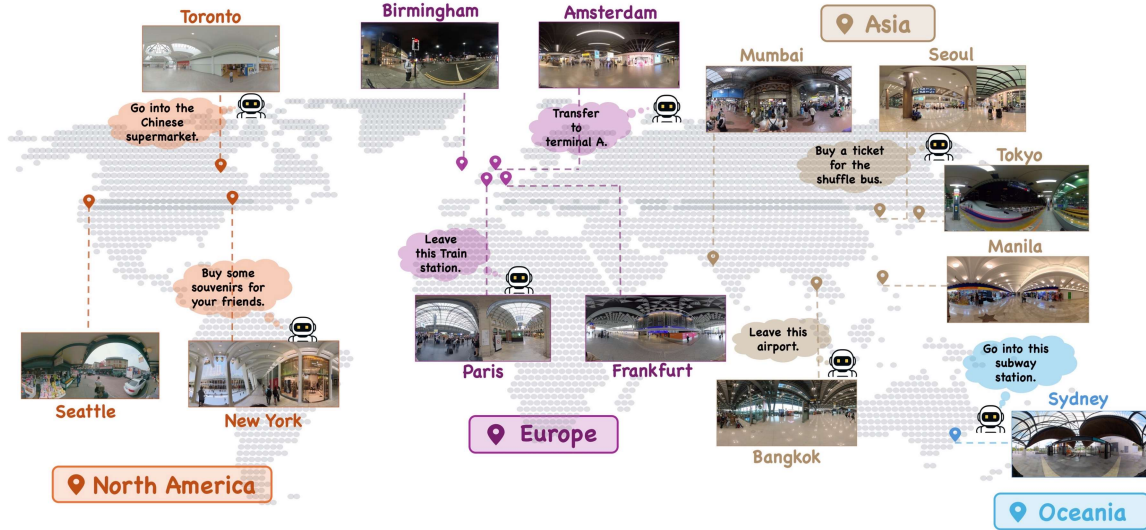


Figure III. H^* *Bench* aggregates panoramic videos from diverse global locations, featuring visually cluttered environments.

VI. Reward Shaping

We use rule-based reward function to calculate the reward of the trajectory, which includes correctness reward, format reward.

$$r = r_{corr} + r_{form},$$

where:

$$r_{corr} = \begin{cases} 0.5, & \text{if the submitted answer satisfies the} \\ & \text{completion condition,} \\ 0, & \text{otherwise,} \end{cases}$$

$$r_{form} = \begin{cases} 0.5, & \text{if the response is in } \langle \text{think} \rangle \langle / \text{think} \rangle \\ & \langle \text{answer} \rangle \langle \text{answer} \rangle \text{format,} \\ 0, & \text{otherwise.} \end{cases}$$

Specially, we add a distance-to-goal reward for *HPS*. It is calculated by the distance of the final direction to the target bounding box.

$$r_{dist} = \frac{\pi - d(\phi_T, \phi^*) + \pi - d(\gamma_T, \gamma^*)}{2\pi}.$$

Distance to bounding box is calculated by:

$$d(\alpha, \alpha^*) = |\alpha - (\alpha^* - \tau_\alpha)| + |\alpha - (\alpha^* + \tau_\alpha)|$$

which remains a constant minimum value when the direction is in the bounding box.

VII. Experimental Setup

Training Details. For the SFT stage, we use a learning rate of $1e-5$. For the RL stage, we apply the GRPO algorithm with a batch size of 32, an actor learning rate of

1×10^{-7} , and a KL penalty coefficient $\beta = 0.01$. Rollouts are conducted under the H^* *Bench* prompts with a temperature of 0.7, a maximum of 8 trajectories, and a dynamic turn limit (5 or 10) based on computational resources. Both stages use an input resolution of 1280×720 and are run on 8 NVIDIA H100 GPUs.

Benchmark Setting. The maximum number of inference turns is set to 10, as the step-cumulative success rate converges before this limit. At each step, the model processes up to five perspective images and uses the latest five dialogue turns as context. The image resolution is 1920×1080 , and the sampling temperature is 0. Due to computational constraints, each episode is capped at 10 steps, with unfinished episodes counted as failures.

Train-Test Split. We annotated $\sim 3,000$ task instances. These were divided into three mutually exclusive splits per task: a benchmark split, an SFT split, and an RL training split. Specifically, we reserved 1,000 instances (600 *HOS* and 400 *HPS*) as the H^* *Bench*, resulting in 4,000 evaluation episodes. From the remaining data, we constructed the SFT dataset by randomly sampling 250 instances from both the *HOS* and *HPS* pools. All leftover instances were allocated exclusively for RL training.

VIII. Additional Qualitative Error Analysis

In Section 5.2, we identified two common-sense reasoning errors in the *HPS* task: (1) *lack of physical commonsense* and (2) *lack of socio-spatial commonsense*. This section provides a detailed explanation of these error types, supported by additional qualitative examples.

Inference Prompt

SYSTEM PROMPT - HOS

You are a robot and perform object searching tasks according to instructions. Your goal is to rotate the camera to center the target object in the camera view, and then submit the task. The camera center is presented as a green cross in the picture.

SYSTEM PROMPT - HPS

You are a robot and perform navigation tasks according to instructions. Your goal is to turn your camera center to the target direction you need to move towards to reach the target location. The camera center is presented as a green cross in the picture. Don't move in the unavailable direction, such as obstacles or gaps.

Actions you can take: rotate(yaw:int,pitch:int), submit(yaw:int,pitch:int)

rotate(yaw:int,pitch:int): rotate the camera in the yaw and pitch direction relative to the current direction. Yaw is the rotation angle in the x-y plane, pitch is the rotation angle in the y-z plane. Yaw angle < 0 means rotate to the right, yaw angle > 0 means rotate to the left. Pitch angle > 0 means look up, pitch angle < 0 means look down.

submit(yaw:int,pitch:int): submit the task with the current camera view with the target object at the center, yaw and pitch are the angles of the current camera view, which is reported by the environment. You can only take one action at a time. The instruction will be provided with each observation. Look at the image carefully to complete the instruction.

FEW SHOT EXAMPLE (disabled after training)

Example:

Round 1: image_1

<think>I need to find the coffee machine. I can see a table on my left, a couch in front of me, and a door to the right. The coffee machine is likely on the table, which is to my left.</think><answer>rotate(-45,0)</answer>

Round 2: Env_feedback: Last action is executed successfully, your current direction (yaw,pitch) is (315,0).

image_2

<think>From the scene, I see that by turning left 45 degrees, a kitchen table is in front of me. The coffee machine is on the left of the table and slightly lower than the camera center. I need to turn leftward and downward a little bit.</think> <answer>rotate(-30,-5)</answer>

Round 3: Env_feedback: Last action is executed successfully, your current direction (yaw,pitch) is (285,-5).

image_3

<think>The coffee machine is right now at the center of my camera, I think I can submit the task.</think>
<answer>submit(285,-5)</answer>

Round 4: Env_feedback: Success

You can take 1 action(s) at a time. You should first give your thought process, and then your answer.

Your response should be in the format of: <think>...</think><answer>...</answer>

e.g. <think>I need to find the coffee machine. I can see a table on my left, a couch in front of me, and a door to the right. The coffee machine is likely on the table, which is to my left.</think><answer>rotate(-45,0)</answer>

USER PROMPT

After your answer, the extracted valid action is {valid_action}.

The environment feedback is: {env_feedback}

done: {done}

After that, the observation is: {observation}

Human Instruction: {instruction}

Decide your next action.

You can take 1 action(s) at a time. You should first give your thought process, and then your answer.

Your response should be in the format of: <think>...</think><answer>...</answer>

Figure IV. Prompt used for inference and rollout.

Examples of HOS Tasks

Find the **Hello Kitty** on the shelf.



$$d = \frac{\text{area}(P \cap T)}{\text{area}(P)}$$

P - area of **target object**
 T - area of perspective observation



Easy: $d \geq 0.5$

Medium: $0 \leq d < 0.5$

Hard: $d = 0$

Find and Buy a **black backpack**



$$d = \frac{\text{area}(P \cap T)}{\text{area}(P)}$$

P - area of **target object**
 T - area of perspective observation

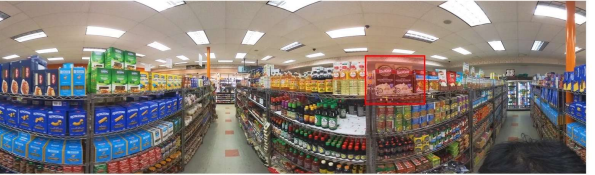


Easy: $d \geq 0.5$

Medium: $0 \leq d < 0.5$

Hard: $d = 0$

Search for a **box of mashed potato**.



$$d = \frac{\text{area}(P \cap T)}{\text{area}(P)}$$

P - area of **target object**
 T - area of perspective observation



Easy: $d \geq 0.5$

Medium: $0 \leq d < 0.5$

Hard: $d = 0$

Look for the shelf selling **K-POP Album**



$$d = \frac{\text{area}(P \cap T)}{\text{area}(P)}$$

P - area of **target object**
 T - area of perspective observation



Easy: $d \geq 0.5$

Medium: $0 \leq d < 0.5$

Hard: $d = 0$

Look for the **fridge containing cold drinks**.



$$d = \frac{\text{area}(P \cap T)}{\text{area}(P)}$$

P - area of **target object**
 T - area of perspective observation



Easy: $d \geq 0.5$

Medium: $0 \leq d < 0.5$

Hard: $d = 0$

Find the **zero sugar**.



$$d = \frac{\text{area}(P \cap T)}{\text{area}(P)}$$

P - area of **target object**
 T - area of perspective observation



Easy: $d \geq 0.5$

Medium: $0 \leq d < 0.5$

Hard: $d = 0$

Figure V. Visualizations of *HOS* task instances.

Easy-Level Examples of HPS Tasks

Easy-level Task: You want to go into a store EYEWORLD OPTICAL, which direction are you going to move?

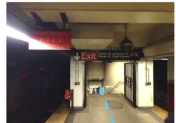
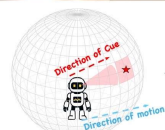
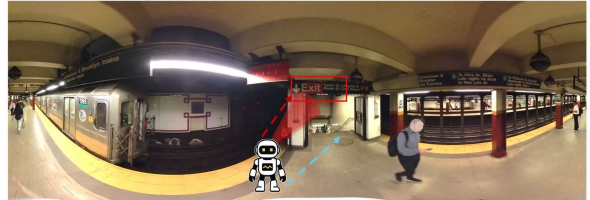


Aligned directions of Cue and Motion with textual instruction.

Direction of **Motion**

Direction of **Cue**

Easy-level Task: You want to exit this platform, which direction are you going to move?



Aligned directions of Cue and Motion with textual instruction.

Direction of **Motion**

Direction of **Cue**

Easy-level Task: You want to go to 9th Avenue, which direction are you going to move?

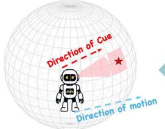


Aligned directions of Cue and Motion with textual instruction.

Direction of **Motion**

Direction of **Cue**

Easy-level Task: You want to buy some jewelry, which direction are you going to move?



Aligned directions of Cue and Motion with textual instruction.

Direction of **Motion**

Direction of **Cue**

Easy-level Task: You want to enter the building called ONE FIVE ONE, which direction are you going to move?

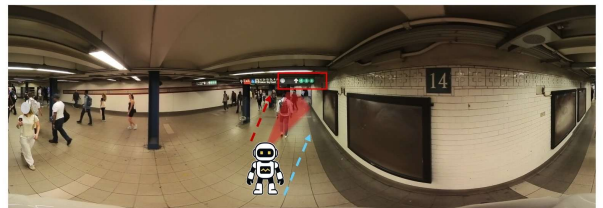


Aligned directions of Cue and Motion with textual instruction.

Direction of **Motion**

Direction of **Cue**

Easy-level Task: You want to take the subway line 4, which direction are you going to move?



Aligned directions of Cue and Motion with textual instruction.

Direction of **Motion**

Direction of **Cue**

Figure VI. Visualizations of easy-level HPS task instances.

Lack of Physical Commonsense. This error type denotes a failure in applying intuitive knowledge about 3D geometry and basic physics. Potential failures caused by the lack of physical commonsense include:

- **Ignoring Permanent Obstacles:** Attempting to move through non-traversable objects such as solid walls, glass barriers, or furniture.
- **Misjudging Vertical Connections:** Failing to understand how different floors are connected, *e.g.*, not recognizing that a staircase or elevator is required to change levels.
- **Direct Path Fallacy:** Heading in a straight-line towards the target without first identifying a feasible path, thereby ignoring the need to follow corridors, detour around obstacles, or use doorways.
- **Overlooking Drop-offs:** Proposing a path that would lead to falling from a significant height, such as walking off a balcony or over a ledge.

Lack of Socio-Spatial Commonsense. This error refers to the agent’s inability to grasp the implicit social norms and functional roles of different areas in public spaces, leading to potential failures as follows:

- **Violating Traffic Norms:** Jaywalking across a busy driveway or road instead of using a nearby crosswalk or pedestrian lane.
- **Trespassing Restricted Zones:** Attempting to cut through behind a retail counter, through a staff-only area, across a floor hazard warning sign (*e.g.*, "Wet Floor"), or through a private property shortcut.
- **Disregarding Spatial Etiquette:** Violating implicit social norms that govern public behavior, *i.e.*, actions that are physically possible but socially disruptive, such as interrupting a queue or invading personal space.
- **Ignoring Functional Layouts:** Violating explicit architectural constraints and intended circulation paths, *i.e.*, attempting to navigate through physical obstructions (like tables or seats) rather than designated aisles.
- **Misusing Spaces:** Proposing to walk through a decorative fountain or flowerbed, or using an emergency exit as a routine shortcut.

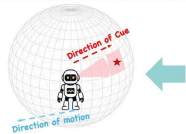
Additional qualitative examples for all five error types (two for *HOS* and three for *HPS*) are provided in Figures X and XIV.

IX. Case Study

In this section, we provide qualitative case studies (See Figs. XV-XVII) comparing the behavior of the model before and after post-training. For each case, we focus on one of the three key capabilities introduced by post-training, as described in Sec. 5.3.

Medium-Level Examples of HPS Tasks

Medium-level Task: You want to exit, which direction are you going to move?



Misaligned directions of Cue and Motion with textual instruction.

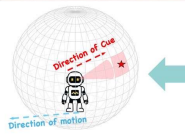


Direction of **Motion**



Direction of **Cue**

Medium-level Task: You want to go to Gate E, which direction are you going to move?



Misaligned directions of Cue and Motion with textual instruction.

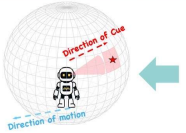


Direction of **Motion**



Direction of **Cue**

Medium-level Task: You want to go to Fulton Center, which direction are you going to move?



Misaligned directions of Cue and Motion with textual instruction.

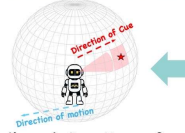


Direction of **Motion**



Direction of **Cue**

Medium-level Task: You want to play PHOENIX Roller Coaster, which direction are you going to move?



Misaligned directions of Cue and Motion with textual instruction.

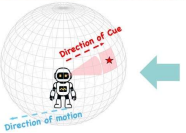


Direction of **Motion**



Direction of **Cue**

Medium-level Task: You want to go Room 104, which direction are you going to move?



Misaligned directions of Cue and Motion with textual instruction.

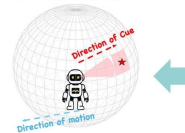


Direction of **Motion**



Direction of **Cue**

Medium-level Task: You want to take the subway to the brooklyn, which direction are you going to move?



Misaligned directions of Cue and Motion with textual instruction.



Direction of **Motion**

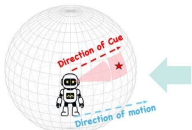


Direction of **Cue**

Figure VII. Visualizations of medium-level *HPS* task instances.

Hard-Level Examples of HPS Tasks

Hard-level Task: You need to leave this airport, which direction are you going to move?



Aligned directions of Cue and Motion **without** textual instruction.

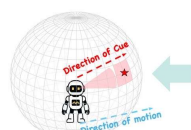
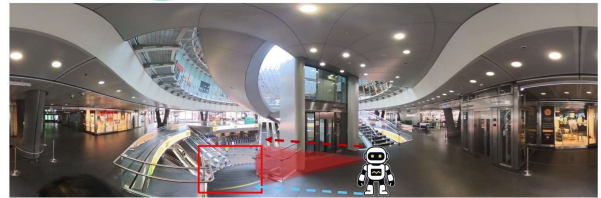


Direction of **Motion**



Direction of **Cue**

Hard-level Task: You want to take the stairs to the subway station, which direction are you going to move?



Aligned directions of Cue and Motion **without** textual instruction.

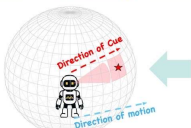
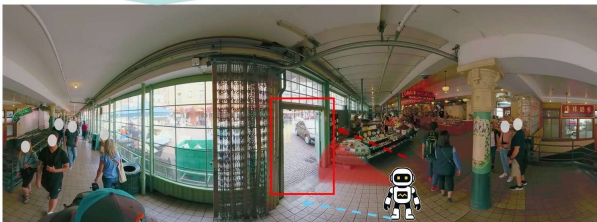


Direction of **Motion**

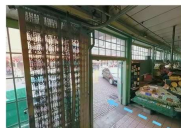


Direction of **Cue**

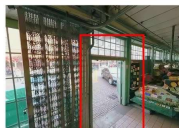
Hard-level Task: You want to take a taxi, which direction are you going to move?



Aligned directions of Cue and Motion **without** textual instruction.

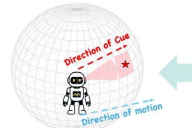


Direction of **Motion**



Direction of **Cue**

Hard-level Task: You want to go into the shopping mall, which direction are you going to move?



Aligned directions of Cue and Motion **without** textual instruction.

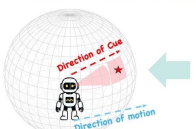
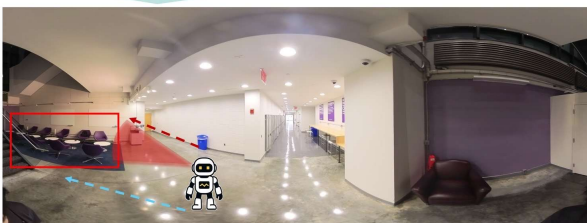


Direction of **Motion**



Direction of **Cue**

Hard-level Task: You want to have a meeting with your teammates, which direction are you going to move?



Aligned directions of Cue and Motion **without** textual instruction.

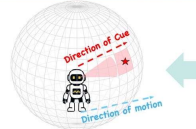
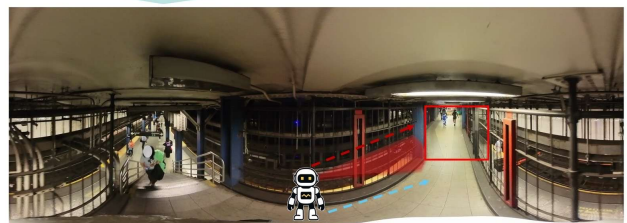


Direction of **Motion**



Direction of **Cue**

Hard-level Task: You want to leave this subway station, which direction are you going to move?



Aligned directions of Cue and Motion **without** textual instruction.



Direction of **Motion**

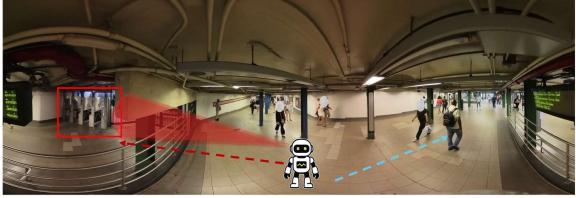


Direction of **Cue**

Figure VIII. Visualizations of hard-level *HPS* task instances.

Extreme-Level Examples of HPS Tasks

Extreme-level Task: You need to leave this subway station, which direction are you going to move?



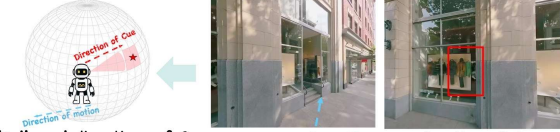
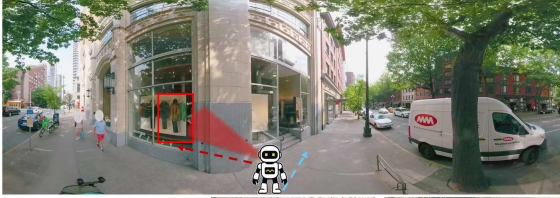
Misaligned directions of Cue and Motion **without** textual instruction.



Direction of **Motion**

Direction of **Cue**

Extreme-level Task: You would like to inquire about the price of the green sweaters on display, where are you going to move?



Misaligned directions of Cue and Motion **without** textual instruction.



Direction of **Motion**

Direction of **Cue**

Extreme-level Task: You want to take escalator to go upstairs, which direction are you going to move?



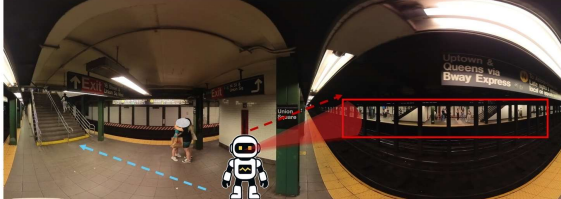
Misaligned directions of Cue and Motion **without** textual instruction.



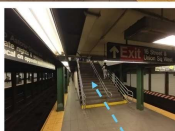
Direction of **Motion**

Direction of **Cue**

Extreme-level Task: You want to go to the opposite platform, which direction are you going to move?



Misaligned directions of Cue and Motion **without** textual instruction.



Direction of **Motion**

Direction of **Cue**

Extreme-level Task: You want to go the platform below, which direction are you going to move?



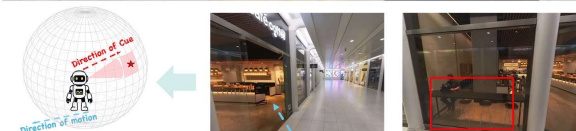
Misaligned directions of Cue and Motion **without** textual instruction.



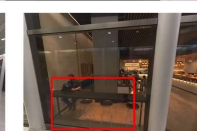
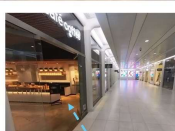
Direction of **Motion**

Direction of **Cue**

Extreme-level Task: You want to have a seat, which direction are you going to move?



Misaligned directions of Cue and Motion **without** textual instruction.



Direction of **Motion**

Direction of **Cue**

Figure IX. Visualizations of extreme-level *HPS* task instances.

🔍 Limited Visual Grounding Capabilities

	<p>Human Instruction: Look for the goods with panda-pattern on it.</p>			
<p>MLLM: <think>I need to find the pattern of a panda. In the current view, I see various products on the shelves. The panda pattern might be on a product package. I should start by looking to the left where there are many packages with different designs. I will rotate the camera to the left to get a better view of the packages on the left shelf.</think><answer>rotate(-75,0)</answer></p>		<p>Human Instruction: Find the Biography, poetry & journals sign.</p>		
	<p>Wrong: There is not any product with panda-pattern on it on the second shelf. You should rotate to explore more area.</p>	<p>MLLM: <think>...The area seems to be part of a larger retail space, but without clear directional signs, it's challenging to determine the exact location of the desired section. Since the goal is to locate the specific sign and there are no visible indicators here, rotating 90 degrees might reveal more of the store layout and potentially the target sign. Therefore, the next action is rotate(90,0).</think><answer>rotate(90,0)</answer></p>		<p>Wrong: The target is not visible in the current view, you should continue to explore this area.</p>
<p>MLLM: <think>In the current view, the panda pattern is now centered in the camera view. It is clearly visible on the package on the second shelf from the top. I have successfully located the panda pattern, and I can now submit the task.</think><answer>submit(290,0)</answer></p>	<p>MLLM: <think>... There is no visible signage indicating "Biography, poetry & journals," but the arrangement suggests this is a section dedicated to literature. Since the goal is to locate the specific sign and the current view aligns with the assumption that this is the correct area, it makes sense to submit the current position. Therefore, the next action is submit(180,0).</think><answer>submit(180,0)</answer></p>			

Figure X. Qualitative Examples of Limited Visual Grounding Capabilities in HOS.

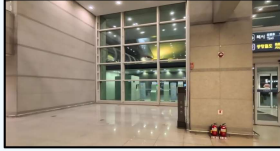
🎯 Perception-Action Gap

	<p>System:... Your goal is to rotate the camera to center the target object in the camera view, and then submit the task.</p> <p>Human Instruction: Find a white t-shirt with a boy and girl on it.</p>		<p>System:... Your goal is to rotate the camera to center the target object in the camera view, and then submit the task.</p> <p>Human Instruction: Look for blue packaged snacks named crisps on the shelf</p>
<p>MLLM: <think>... Since the current view does not show any relevant items, it makes sense to explore further. Rotating 180 degrees will allow me to view the opposite direction, potentially revealing more clothing options or signage that could guide me to the target item. Therefore, the next action is rotate(180,0).</think><answer>rotate(180,0)</answer></p>	<p>MLLM: <think>Since the blue boxes are not visible, I need to rotate the camera to find the blue packaged snacks named crisps. I should rotate the camera to the left to look for the blue boxes on the shelves.</think><answer>rotate(-45,0)</answer></p>		
	<p>Wrong: Target is not foveated in the final submitted view. You should slightly rotate to the left to foveate it.</p>		<p>Wrong: Target is not foveated in the final submitted view. You should slightly rotate to the left to foveate it.</p>
<p>MLLM: <think>I notice a section with various t-shirts, including a white one with images of a boy and girl on it. This matches the task requirement. The target item is clearly visible and centered in the current view. Since the goal is to find and buy this specific t-shirt, it makes sense to submit the current view as the final action. Therefore, the next action is submit(90,0).</think><answer>submit(90,0)</answer></p>	<p>MLLM: <think>Since the blue boxes are visible, I have found the blue packaged snacks named crisps on the shelf. I should submit the task with the current camera view.</think><answer>submit(0,0)</answer></p>		

Figure XI. Qualitative Examples of Perception-Action Gap in HOS.



Vision-Action Mismatch

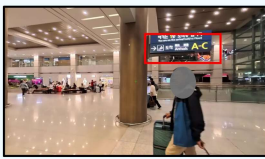


System: ... your goal is to rotate the camera to face the direction you **need to move forward**, and then submit the task.

Human Instruction: you want to go to arrival lounge B, where are you going to move?

MLLM: <think>...Since there are no clear indicators for the arrival lounge B in this view, rotating 180° might reveal additional signage or pathways leading to the arrival lounge. Therefore, the next action is rotate(180,0). </think><answer>rotate(180,0)</answer>

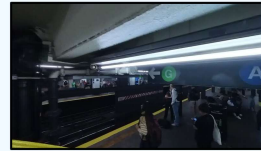
Rotate (180,0)



Wrong: You should turn right to face the direction of arrival lounge B according to the sign.



MLLM: <think>I notice a sign indicating directions to arrivals A-C, with an arrow pointing straight ahead. This suggests that the arrival lounge B is likely in the same direction. The area is spacious with people moving towards the indicated path. **Since the goal is to reach arrival lounge B, following the direction of the sign is logical. Therefore, the next action is submit(270,0).** </think><answer>submit(270,0)</answer>



System: ... your goal is to rotate the camera to face the direction you **need to move forward**, and then submit the task

Human Instruction: You want **move to the direction of A C**, where are you going to move?

MLLM: <think>In the image, the sign indicating the direction of the A and C lines is to the right of the camera center. To align the camera center with the direction of the A and C lines, I need to rotate the camera to the right. </think><answer>rotate(45,0)</answer>

Rotate (45,0)



Wrong: You should turn right to face the direction of the A&C according to the sign.



MLLM: <think>In the current image, the sign indicating the direction of the A and C lines is now directly in front of the camera center. **This means the camera is aligned with the direction I need to move towards. I can now submit the task with the current direction.** </think><answer>submit(45,0)</answer>

Figure XII. Qualitative Examples of Vision-Action Mismatch in HPS.



Lack of Physical Commonsense

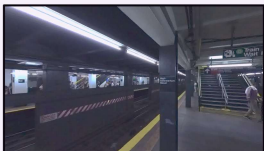


System: ... Don't move in the unavailable direction, such as obstacles or gaps.

Human Instruction: You want to go to the opposite platform, where are you going to move?

MLLM: <think>The previous action rotated my camera to the right. Now I can see the other platform. I need to rotate my camera further to the right to move closer to the other platform. I should rotate my camera to the right a little bit more. </think><answer>rotate(45,0)</answer>

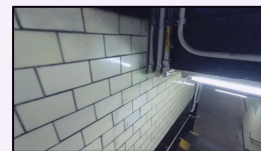
Rotate (45,0)



Wrong: You should first go upstairs to leave this platform, therefore you should continue to turn right to face the staircase.



MLLM: <think>The previous rotation brought me closer to the other platform. **Now my camera is facing the other platform. I can submit the task now.** </think><answer>submit(0,0)</answer>

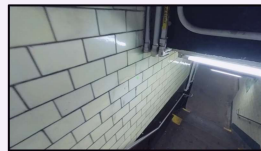


System: ... Don't move in the unavailable direction, such as obstacles or gaps.

Human Instruction: You want to go downstairs, where are you going to move?

MLLM: <think>I am looking downwards now. I can see a dark area below, which is likely the stairs. I need to turn further downwards to get a better view of the stairs and confirm the direction. I should continue to look down. </think><answer>rotate(0,-10)</answer>

Rotate (0,-10)



Wrong: Your final submitted direction is not navigable, you should turn right to face the direction of staircase.

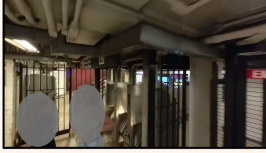


MLLM: <think>I can now clearly see the stairs. I need to submit the task now. </think><answer>submit(90,-30)</answer>

Figure XIII. Qualitative Examples of Lack of Physical Commonsense in HPS.



Lack of Socio-Spatial Commonsense



Human Instruction: You want to leave this subway station, where are you going to move?

...
✦ MLLM: <think>I am still moving towards the right. The current direction is 315. I need to continue rotating to the right to reach the exit.</think><answer>rotate(45,0)</answer>

Rotate (45,0)



Wrong: You are now facing the Emergency Exit, you should leave this subway station through the subway gate, that you have already seen in previous view.

✦ MLLM: <think>The exit is in front of me. I need to submit the task.</think><answer>submit(0,0)</answer>



Human Instruction: you want to take the subway, where are you going to move?

✦ MLLM: <think>I am looking at a wall with a metal railing. I need to move towards the subway entrance, which is likely to the right. I should rotate my camera to the right to get a better view of the direction I need to move.</think><answer>rotate(90,0)</answer>

Rotate (90,0)



Wrong: You are now on a staircase in a subway station. The subway platform is typically underground, so you should rotate 180° to face the stairs going down to the platform in order to take the subway.

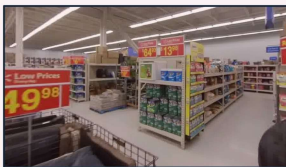
✦ MLLM: <think>The subway entrance is now in front of me. I should submit the task now.</think><answer>submit(270,0)</answer>

Figure XIV. Qualitative Examples of Lack of Socio-Spatial Commonsense in HPS.

✗ Failure in Foveating the Target

Post-Training

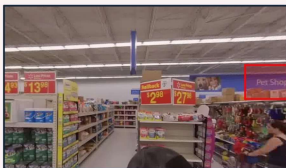
✓ Precise Angle-Control for Target Foveation



Human Instruction: Look for the sign Pet Shop.

✦ Qwen2.5-VL-3B: <think>... Since the sign is not immediately visible, I'll try rotating the camera slightly to the right to see if it appears in the new view.</think><answer>rotate(30,0)</answer>

Rotate (30,0)



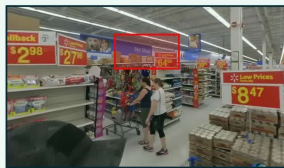
✦ Qwen2.5-VL-3B: <think>Since the 'Pet Shop' sign was found in the previous observation, the task is completed successfully. No further actions are needed.</think><answer>submit(210,0)</answer>



Human Instruction: Look for the sign Pet Shop.

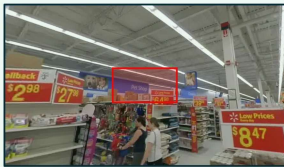
★ HVS-3B: <think>... Since the goal is to find the "Pet Shop" sign, it makes sense to explore more of the store by rotating. Therefore, the next action is rotate(90,0) to potentially reveal new areas and signs.</think><answer>rotate(90,0)</answer>

Rotate (90,0)



★ HVS-3B: <think>I notice the "Pet Shop" sign is clearly visible in the center of the current view. The task is to find this sign, and since it is prominently displayed, I should slightly rotate to center it.</think><answer>rotate(0,15)</answer>

Rotate (0,15)



★ HVS-3B: <think>The "Pet Shop" sign is clearly visible and centered in the current view. The task is to find this sign, and since it is prominently displayed, the objective is achieved. Therefore, the next action is to submit(270,15) as the target is in the center of the view.</think><answer>submit(270,15)</answer>

Figure XV. Case study of target foveation. **Left:** Pre-trained model submits prematurely after coarse rotation toward "Pet Shop" sign, leaving target off-center. **Right:** Post-trained model performs corrective turn with fine-grained rotation to accurately center target before submission.

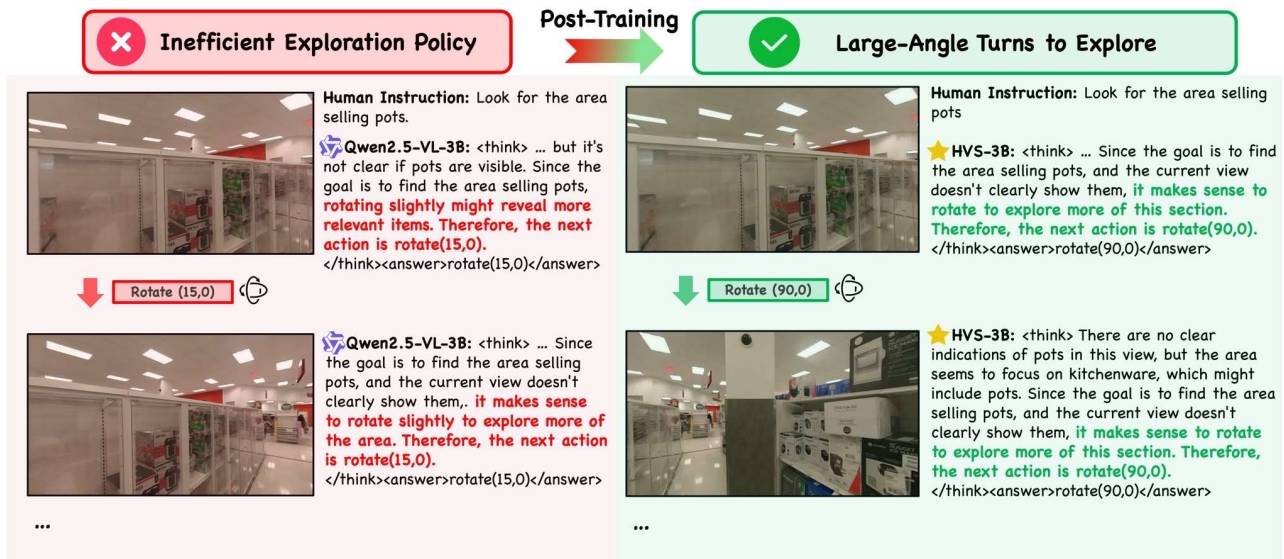


Figure XVI. Case study of exploration behavior. **Left:** The pre-trained model employs an inefficient exploration strategy, making repeated small-angle turns that provide little new information about the location of pots. **Right:** The post-trained model learns to perform large-angle exploratory turns, which quickly reveal new aisles and identify the target section, resulting in a more efficient search trajectory.

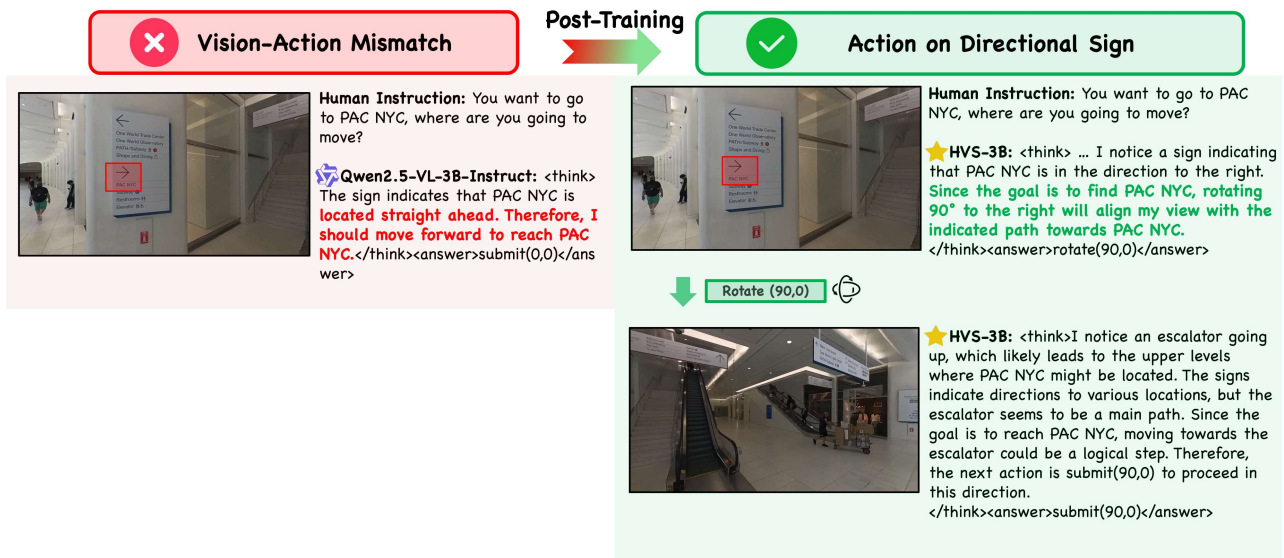


Figure XVII. Case study of action selection based on directional signs. **Left (Pre-training):** The model misinterprets the sign, selecting an action that contradicts the indicated route and resulting in a vision-action mismatch. **Right (Post-training):** After training, the model correctly follows the sign's instruction, rotates 90° to align with the target direction, and proceeds towards the goal.