EmbRACE-3K: Embodied Reasoning and Action in Complex Environments (Supplementary Material)

1 More Details on Data Collection and Benchmark Construction

1.1 Task Instruction Generation

Figure 3 shows the prompt template used to generate natural language navigation tasks with Gemini. The prompt encodes task types, visibility constraints, and spatial-semantic reasoning requirements, guiding the generation of grounded instructions aligned with the agent's egocentric observation. Each task is conditioned on a sampled agent pose, surrounding object metadata, and a list of reasoning constraints, ensuring task diversity and consistency.

Standard navigation tasks. For Type 0–3 tasks, we provide Gemini-2.5-Pro [2] with the current egocentric image and a list of candidate objects within a 1000-meter radius. Based on this input, Gemini generates instructions falling into four predefined categories:

- Type 0 (Basic): The target is clearly visible and reachable without exploration.
- **Type 1 (Exploration):** The target is not initially visible and must be discovered through search.
- **Type 2 (Dynamic Spatial Semantic):** The target is visible but referred to via relative or ordinal spatial language (e.g., "the second chair," "the one on the right").
- **Type 3 (Multi-stage):** Two visible targets must be remembered across steps and approached in sequence.

Interaction-based tasks. For Type 4 tasks involving object interaction, we follow a hybrid generation process that combines manual layout curation with template-guided instruction creation. Interaction tasks are divided into:

- **Type 4.1 (Open Door):** We manually place the agent near an interactive door. If the door is clearly visible from the agent's initial view, a visible-door template is selected. Representative templates include:
 - "Walk to the door ahead and open it.";
 - "Step forward and open the door visible in the scene.";
 - "Move toward the obvious door in front of you and open it.";
 - "You're facing a door. Approach and open it."

When the door is not visible initially, we sample from an exploratory search template set designed to encourage navigation and search behavior:

- "Look around and locate a door to open.";
- "Search the area and open a nearby door once you find it.";
- "Explore the environment and open the first door you find.";
- "Walk around until a door is in view, then go open it."
- **Type 4.2 (Pick & Drop):** Graspable objects (e.g., trash bags, candle holders, burlap sacks) are placed near the agent, and distinct landmarks are selected as drop targets. Instructions emphasize both object manipulation and relative spatial reasoning. Representative templates include:
 - "Pick up the trash bag and place it near the red car.";
 - "Bring the candle holder over and nestle it into the far-right stone corner.";
 - "Carry the burlap sack and drop it next to the table on the far-left side closest to you."

Final instruction curation. All generated instructions are manually reviewed for scene consistency, target uniqueness, and spatial clarity. This ensures that each command remains executable and unambiguous given the visual field of the agent. The final instruction set used for training and evaluation supports realistic yet controlled embodied tasks, with clear grounding in egocentric perception and task semantics.

1.2 Step-wise Reasoning Annotation

To provide interpretable supervision aligned with agent behavior, we annotate each step of the demonstration trajectory with natural language reasoning that explains the agent's decision at that moment. These *thinking* annotations simulate the internal deliberation process of an embodied agent, grounded in egocentric perception and constrained by task semantics.

As illustrated in Figure 4, we construct a prompt that includes the task instruction, the agent's complete egocentric image sequence, and the full history of executed actions. This prompt is provided to Gemini-2.5-Pro, which returns a sentence describing why the current action is appropriate given what the agent perceives and remembers. Importantly, the model is explicitly instructed to avoid using any future information and to write in the first person, as if it were the agent itself.

To ensure geometric consistency and spatial reasoning fidelity, the prompt imposes constraints on spatial language: object locations must be described using viewpoint-aligned phrases such as "on the left-hand side" or "in the center," and references like "to the left" or "right beside me" are disallowed. For multi-stage instructions, the reasoning must focus exclusively on the first subgoal until the designated MidwayTarget action appears in the trajectory.

This step-wise annotation process results in a rich set of perception-action-thought triplets that reflect how embodied agents might reason in closed-loop settings. These annotations serve as fine-grained supervision for training and offer a transparent lens for evaluating spatial-semantic alignment during embodied decision-making.

1.3 Data Prompts

We construct three types of data prompts for model training: (i) supervised fine-tuning (SFT) with reasoning, (ii) SFT without reasoning, and (iii) reinforcement learning (RL). All prompts simulate the embodied decision-making process based on task objectives, egocentric visual observations, and action histories.

SFT with reasoning. For step-wise reasoning supervision, we provide the model with the full task instruction, a time-ordered sequence of first-person egocentric views up to the current step, and the full trajectory of past actions. The model is expected to produce both a natural language explanation (<think>) and the corresponding action (<action>) at each step. This setup encourages learning spatially grounded, interpretable decision policies aligned with visual context and task semantics.

SFT without reasoning. For instruction-following SFT without reasoning, the prompt retains the same structure but omits the reasoning component. The model directly predicts the next action given the full observation-action history and task instruction. This setting isolates the effect of action-only learning, enabling ablation comparisons with reasoning-augmented variants.

RL training. During reinforcement learning, the same prompt format as the reasoning-free SFT setup is used. The model interacts with the environment under a policy trained via reward feedback and generates the next action at each step without explicit language supervision. The reward signal supervises the selected action and its format, encouraging alignment with the target behavior defined by the dataset.

Figures 5 and 6 illustrate examples of the two prompt formats used in SFT and RL training.

2 More Visual Details of EmbRACE-3K

To better illustrate the structure and coverage of EmbRACE-3K, we provide visualizations of representative human-annotated trajectories across all six task types. Each example includes the full



Figure 2: Training curves of three key metrics during reinforcement learning with GRPO. "Completion Length" indicates the output length of the policy model during rollout process. The policy model is initialized from the SFT checkpoint.

instruction, a sequence of egocentric first-person views, the corresponding human actions, and (where available) step-wise reasoning annotations. These demonstrations reflect the diversity of embodied behaviors and instruction formats captured in the dataset.

Figures 7 through 12 showcase how different task types correspond to distinct reasoning challenges. In particular, EmbRACE-3K is designed to probe three core limitations in current video-trained vision-language models: (i) short-sighted exploration, where agents lack persistence in visually-guided search; (ii) dynamic spatial-semantic drift, where spatial references change meaning under egocentric motion; and (iii) target forgetting, where objects that briefly leave the visual field are not recalled later in the task. The examples emphasize how task instructions, visual cues, and behavior sequences interact to reveal these difficulties.

Rather than prescribing a unique solution, the visualizations present one plausible human trajectory per task. These examples serve to concretely illustrate the design intent and diagnostic scope of EmbRACE-3K, providing interpretable reference patterns for supervised or policy-based training.

3 More Details of Experiments

3.1 SFT Training Curve

Figure 1 presents the SFT loss curve of Qwen2.5-VL-7B [1] on the EmbRACE-3K dataset. The training loss decreases steadily as the model is optimized to align egocentric observations with instruction-conditioned action decisions. The smoothed trajectory demonstrates stable convergence across training iterations, indicating the learnability of our structured prompts and the coherence of human-annotated trajectories and reasoning signals.

Over the course of 200 training steps, the loss drops from an initial value above 1.5 to below 0.2. The sharp decline in the early phase (0–50 steps) reflects rapid adaptation to the format and structure of instruction-action pairs. Between steps 50 and 150, the curve flattens into a gradual descent, suggesting progressive refinement of action



Figure 1: SFT training curve of Qwen2.5-VL-7B on EmbRACE-3K. The smoothed curve shows consistent convergence.

alignment and reasoning consistency. After step 150, the model enters a stable convergence phase, where improvements in loss are marginal, and most trajectories are predicted with high confidence. This progression illustrates the effectiveness of EmbRACE-3K in providing dense and learnable supervision for embodied instruction following.

3.2 RL Training Curve

Figure 2 illustrates the training curves of the Qwen2.5-VL-7B model after undergoing EmbRACE-3K SFT and further optimization through the GRPO algorithm during the reinforcement learning process. In Figure 2(a), we present the defined format training curve, where the reasoning process is encapsulated within the think></think> tags and the final action decision is enclosed within the <action></action> tags. Due to the model's strong inherent instruction-following capabilities,

the format reward consistently remains close to a high score of 1. Figure 2(b) demonstrates the accuracy of the model's action decisions during training. Over the course of 500 steps, the average accuracy reward improves from below 0.5 to approximately 0.6, showcasing the effectiveness of the reinforcement learning algorithm in enhancing the model's reasoning and decision-making abilities in embodied environments beyond its SFT baseline. Finally, Figure 2(c) depicts the average output length of the model, which remains relatively stable throughout training. This stability reflects the nature of the agent's reasoning, which primarily involves spatial and planning-related tasks, where concise and purposeful reasoning suffices to support effective decision-making.

3.3 Visualization Results of Different Models on EmbRACE-3K

Exploration Tasks. Figures 13 through 15 illustrate model behaviors on the exploration task: "Go find the refrigerator." In this scenario, the target object is not initially visible and requires the agent to actively search by leveraging spatial cues and egocentric observations. Gemini-2.5-Pro and GPT-40 both begin with reasonable initial moves, identifying nearby doors and exploring adjacent rooms. However, both models exhibit repetitive turning patterns after failing to locate the refrigerator within a few steps. Gemini falls into prolonged oscillation, executing many redundant turns without converging toward new areas. GPT-40 performs slightly better by opening a door and entering another room, but ultimately fails to locate the refrigerator and times out without completing the task. In contrast, the Qwen2.5-VL-7B model fine-tuned on EmbRACE-3K demonstrates a more structured and consistent search behavior. It executes a smooth scan of the environment, identifying the refrigerator as soon as it becomes visible. Once detected, it adjusts its heading to center the target and proceeds directly toward it. This behavior reflects improved visual persistence, spatial awareness, and efficient exploration planning—key benefits introduced by the structured task supervision in our dataset.

Dynamic Spatial-Semantic Tasks. Figures 16 through 18 visualize model responses to the instruction "Go to the second pillar on the left in the initial view," a task requiring the agent to interpret ordinal and directional references grounded in the initial egocentric frame. Both Gemini-2.5-Pro and GPT-40 initially identify the general area of the pillars and begin forward motion. However, as they approach the scene, both models exhibit difficulty maintaining the spatial semantics of the instruction. Gemini becomes disoriented near the first pillar, entering a loop of repeated turning without making progress toward the second. GPT-40 initially moves toward the pillars, but after some turns loses alignment with the target and begins a sequence of corrective but ultimately inefficient reorientations. In both cases, the models fail to preserve the mapping between initial-frame spatial language (e.g., "second on the left") and the dynamically evolving viewpoint. By contrast, the Qwen2.5-VL-7B model fine-tuned on EmbRACE-3K maintains a consistent interpretation of the ordinal reference. It accurately identifies the second pillar from the initial view, tracks it through egocentric movement, and successfully approaches it without losing spatial context. This behavior suggests that our dataset's explicit supervision of target reasoning and spatial grounding helps mitigate semantic drift as the agent's perspective changes.

Multi-stage Tasks. Figures 19 through 21 illustrate the performance of Gemini-2.5-Pro, GPT-40, and Qwen2.5-VL-7B on the same multi-stage task: "First, walk to the left pillar, then go to the right pillar." Both Gemini and GPT-40 are able to complete the first subgoal and reach the left pillar. However, they fail to effectively transition to the second target. Gemini enters prolonged turning loops after reaching the first pillar, issuing multiple orientation commands without successfully identifying or approaching the right pillar. GPT-40 similarly fails to relocate the second target and issues repeated MidwayTarget signals even after subgoal completion, indicating confusion or redundancy in goal grounding. In both cases, the task ultimately times out. In contrast, the Qwen2.5-VL-7B model fine-tuned on EmbRACE-3K successfully completes both subgoals in sequence. After reaching the left pillar, it correctly emits a MidwayTarget signal and initiates a turn toward the right pillar. Although the second subgoal is temporarily out of view, the model is able to infer its likely direction based on its position in the initial frame. It then proceeds with decisive and directed movement, completing the task without redundant actions. This indicates that the model not only retains the high-level task structure but also leverages earlier visual context to guide planning beyond the current field of view. Such behavior highlights the benefit of fine-tuning on EmbRACE-3K, which supports step-level supervision and reinforces multi-stage goal reasoning under egocentric constraints.

4 Limitations and Broader Impacts

Limitations. Although EmbRACE-3K covers a wide range of embodied reasoning scenarios, its current interactive capabilities are limited to two action types: opening doors and picking or dropping objects. This constraint stems from the capabilities of the underlying UnrealCV-Zoo platform [3], which currently supports only a restricted set of interactions. Expanding the set of supported object manipulations, tool use, and environment-triggered events is a promising direction for future extensions. Additionally, while our dataset emphasizes step-wise reasoning and egocentric visual grounding, it does not yet include human-agent interaction or multi-agent coordination, which are increasingly important in real-world embodied AI applications.

Broader impacts. EmbRACE-3K contributes to the development of vision-language agents that reason and act in physically grounded, sequential decision-making tasks. By providing a benchmark focused on closed-loop interaction, it encourages research that moves beyond passive scene understanding toward actionable multimodal intelligence. The dataset does not contain real-world images or identifiable human data, minimizing risks related to privacy or fairness. However, as embodied agents become more capable, their deployment in physical or virtual environments raises questions of safety, transparency, and alignment with user intent. We encourage the community to use EmbRACE-3K in ways that promote interpretable, controllable, and socially beneficial AI systems.

Objective: Design navigation tasks for embodied AI training in Unreal Engine. The objects within 1000-meter are: [{', '.join(nearby_object_names)}]

Core Challenges with Examples

Challenge 1: Short-sighted Exploration

- For situations where the agent cannot see the target from the first-person view initially and has no knowledge of its location, the model may fail to produce continuous, sensible decisions.

- Example: "Explore the surroundings and walk to the red car." A conventional model might turn left once, see no red car, then turn back immediately; then turn right once, see no red car, and turn back. Seeing no red car on either side, it simply moves forward. However, the red car might require turning multiple times to discover.

Challenge 2: Dynamic Spatial-Semantic Drift

- Instructions like "Go to the door of the white house in front of you" or "Go to the second intersection" can become invalid as the agent moves, because references such as "in front of you" and "the second intersection" change relative to the agent's new position

Challenge 3: Target Forgetting

- In multi-stage tasks such as "First walk to the trash can, then walk to the red car," the agent may forget the second target after completing the first sub-task.

- Even in single-step tasks like "Walk near the red car," the agent might momentarily lose sight of the car if it takes a large stride and fail to readjust its path, indicating insufficient long-term target binding in a dynamic environment.

Task Design Rules

- 1. Half of the instructions should describe simple, direct paths to clearly visible objects.
- 2. Avoid any references to previously seen objects or prior knowledge (no memory of earlier views)
- 3. For exploration tasks, assume the agent has only its current first-person view with no map priors.
- 4. Each command should be concise and clear (e.g., "Go to the red car").
- 5. All commands must be in English.
- 6. Every command must specify a final destination (e.g., "Go to X"), not single-step actions like "turn left."
- 7. Ensure no ambiguity: there should not be multiple objects in the scene that match the same description.
- For large objects, specify a distinct component (e.g., "go to the door of the house" instead of "go to the house").
 For dynamic spatial semantics tasks, do not include targets that are not visible in the current first-person view (e.g., avoid "Go to the main door behind you").

Output Format Rules

1. Return exactly 10 tasks as a list of dictionaries in JSON format. No other extra information.

- 2. Each dictionary must contain:
- 'Instruction': Full navigation command
- 'Target Object': List of EXACT name(s) from the objects (within 1000-meter) list.
- Single stage: Single-element list (e.g., ["A"])
- Multiple stage: multi-element ordered list (e.g., ["A", "B"])
- 'Reason': Explain why this task is relevant or interesting
- 'Type': Integer code:
- 0: Basic (direct path to visible target)
- 1: Exploration (requires visual search)
- 2: Dynamic spatial semantic
- 3: Multi-stage (only two sequential sub-goals here)

Strict Rule (Must Be Followed)

The input image is the agent's only first-person egocentric view. It is the sole basis for determining what the agent can

Although all candidate objects are within a 1000-meter radius, only those that actually appear in the input image are considered visible to the agent. Any object that does not appear in the image must be treated as invisible, regardless of distance.

Visibility constraints by task type

- Type 0 (Basic): Target must be clearly visible and directly reachable in the input image.

- Type 1 (Exploration): Target is currently invisible (not shown in the input image) and must be discovered through navigation. - Type 2 (Dynamic Spatial Semantic): Target is visible, but requires spatial or ordinal reasoning to identify (e.g., "the second
- chair," "the one on the right," "the front house").

- Type 3 (Multi-stage): Both targets must be visible in the input image. The challenge lies in remembering or reasoning about the second target after it temporarily goes out of view.

Do not use any object that is outside the input image for Type 0, Type 2, or Type 3 tasks. Only visible objects should be used for these task types.

Figure 3: Prompt template used to generate navigation task instructions with Gemini. The prompt defines core challenges (short-sighted exploration, spatial-semantic drift, target forgetting), formalizes instruction types (Basic, Exploration, Spatial Semantic, Multi-stage), and enforces strict visibility constraints based on the agent's egocentric view. Tasks are designed to simulate realistic, instructionguided navigation in interactive embodied settings.

You are an embodied AI agent making real-time decisions during an embodied task.

The following is a sequence of first-person images and actions, recorded in Unreal Engine. Each image represents your visual observation just before you chose an action.

You may access the full trajectory to help disambiguate scenes, but your explanations must be written as if you do not know what happens next. Never refer to future steps.

Your responsibilities for each step:

- 1. Articulate your internal reasoning for the chosen action:
- This is the only section where reasoning and inference are allowed.
- Think and write as yourself the embodied agent deciding how to act based on what you perceive in the scene.
- Your reasoning must be forward-looking: simulate the thought process that leads to an action based on the current view. Use first-person, natural-sounding language like:
 - "I see the barrels on the left-hand side. To center them in my view, I'll turn slightly to the left."

Focus on task-relevant objects you perceive in the scene. Mention all visible objects of the same category as the task target when relevant to your decision-making, including their spatial location relative to your viewpoint (e.g., "on the left-hand side", "in the center", "on the right-hand side"). Base spatial locations on the image frame being divided vertically into three roughly equal sections: left-hand side, center, and right-hand side.

Do not use phrasing like "to the left" or "to the right" that implies endpoint targeting. Always describe spatial alignment with "on the left-hand side", "in front", "on the right-hand side", etc.

Assume your camera and body are aligned (ego-view equals agent orientation), unless the action is LookUp, LookDown, LookHand or LookBack

Crucially, ensure your stated reasoning is geometrically consistent with the spatial locations of objects as observed in the image. For example, if you perceive an object on the left-hand side, turning left should logically follow if your goal is to center it or move towards it. Turning right would move it further left in your view. Use this geometric logic to ensure your reasoning makes physical sense.

If the visual scene seems inconsistent with geometric turning logic based on previous steps, trust the geometric reasoning and assume the visual elements support the action chosen

Always prioritize what makes sense physically.

Language constraints:

- 1. The word "right" may only be used in explicit spatial phrases (e.g., "on the right-hand side", "in the right half of the image"). You must **never** use "right" in a non-spatial way (e.g., "right beside me", "right away", "the right option").
 If the task instruction contains a phrase like "on the right", you must:
- Put that phrase in guotation marks: "on the right".
- Explicitly state that this is a quote from the instruction.
- If the object's position has changed relative to your view, explain this shift:
 - Example: The instruction says "on the right", but the object is now on the left-hand side of my view.

If the task instruction includes multiple stages (e.g., "first go to the barrels, then go to the tree"), follow these strict rules:
 Before the appearance of an action tagged with "MidwayTarget", you must **only consider the first-stage goal** in your

reasoning.

- Do not mention, plan, or hint at the second-stage goal before MidwayTarget.
- When the action is MidwayTarget, only explain why the first-stage goal is satisfied.
- You may begin considering the second-stage goal only in the frame after the MidwayTarget action.

Task instruction: {instruction}

Trajectory steps: {stepwise text}

Return a JSON list, one entry per frame. Each entry must include:

- "image_name": the frame's image filename (e.g., "002.jpg") "action": the action you chose after viewing this frame
- "thinking": your internal reasoning for the action
- This is the only field where inference is allowed
- Think in first-person, based on current and past frames only.

Directly reflect your visual perception of the scene, focusing on task-relevant objects and their spatial locations (left/center/right)

- Ensure reasoning is consistent with perceived spatial location and geometric turning logic.
- Avoid any mention of second-stage goals unless MidwayTarget has occurred
- If MidwayTarget has occurred, you may begin reasoning toward the second-stage goal.

Begin your step-by-step reasoning now.

Figure 4: Prompt template used to generate step-wise reasoning annotations with Gemini-2.5-Pro. The input includes the task instruction, the complete sequence of egocentric visual observations, and the full history of executed actions. At each step, the model generates a first-person explanation for the current action, grounded in spatial layout and task semantics, while explicitly avoiding any reference to future states.

```
"messages": [
```

"role": "user", "content": "<image>\nYou are an autonomous embodied agent operating in Unreal Engine.\nYour goal is to select the next best action that helps efficiently and safely complete the assigned task.\n\nInputs:\n- TASK: \"Survey the environment for any door, then approach and open it.\"\n- FIRST-PERSON VIEW IMAGES: A time-ordered sequence of the agent's first-person views. Images later in the sequence correspond to later time steps.\n- TRAJECTORY: []\n\nAnalysis Criteria:\n1. Trajectory Evaluation:\n- Detect repeated movement patterns such as loops or oscillations\n- Identify inefficiencies like back-and-forth motion\n- Infer how the agent's relative position to the target has evolved\n\n2. Current Visual State:\n- Identify visible obstacles around the agent\n- Estimate the proximity of the target object based on size/position\n- Evaluate navigable movement directions (e.g., clear paths vs. blocked ones)\n/n3. Task-oriented Action Assessment:\n- Consider which next action is most likely to bring the agent closer to completing the task effectively\n- Avoid actions that merely preserve motion without strategic progress\n\nAdditional Objectives:\n- Maximize path efficiency\n- Avoid unproductive patterns: circular loops, repeated back-and-forth movements\n\nOutput MUST be in the following format:\n<think>Your reasoning about the next action</think><action>One of: MoveForward, MoveBackward, TurnLeft, TurnRight, Finish, LookUp, LookDown, LookHand, LookBack, JumpForward, OpenDoor, PickObject, DropObject, MidwayTarget</action>\n\nConstraints:\n- The `action` value must be exactly one of the allowed commands listed above (case-sensitive).\n- Use \"MidwayTarget\" if the agent has just reached a **sub-goal** or **intermediate objective** in a multi-stage task."

```
"role": "assistant",
```

"content": "<think>My task is to find a door, approach it, and open it. I don't see any doors in my current view, which shows shelves on the right-hand side and tables with cabinets on the left-hand side and in the center. I'll turn left to survey more of the environment.</think><action>TurnLeft</action>"

"images": [

"embrace3k_train_image/SuburbNeighborhood_Day/x03703_y01064_z-0132_door0/000.jpg"

```
"messages": [
```

"role": "user"

content": "<image>\n<image>\nYou are an autonomous embodied agent operating in Unreal Engine.\nYour goal is to select the next best action that helps efficiently and safely complete the assigned task. \n\nlnputs:\n-TASK: \"Survey the environment for any door, then approach and open it.\"\n- FIRST-PERSON VIEW IMAGES: A time-ordered sequence of the agent's first-person views. Images later in the sequence correspond to later time steps.\n- TRAJECTORY: ['TurnLeft']\n\nAnalysis Criteria:\n1. Trajectory Evaluation:\n- Detect repeated movement patterns such as loops or oscillations\n- Identify inefficiencies like back-and-forth motion\n- Infer how the agent's relative position to the target has evolved\n\n2. Current Visual State:\n- Identify visible obstacles around the agent\n- Estimate the proximity of the target object based on size/position\n- Evaluate navigable movement directions (e.g., clear paths vs. blocked ones)\n\n3. Task-oriented Action Assessment:\n- Consider which next action is most likely to bring the agent closer to completing the task effectively\n-Avoid actions that merely preserve motion without strategic progress\n\nAdditional Objectives:\n- Maximize path efficiency\n-Avoid unproductive patterns: circular loops, repeated back-and-forth movements\n\nOutput MUST be in the following format:\n<think>Your reasoning about the next action</think><action>One of: MoveForward, MoveBackward, TurnLeft, TurnRight, Finish, LookUp, LookDown, LookHand, LookBack, JumpForward, OpenDoor, PickObject, DropObject, MidwayTarget</action>\n\nConstraints:\n- The `action` value must be exactly one of the allowed commands listed above (casesensitive).\n- Use \"MidwayTarget\" if the agent has just reached a **sub-goal** or **intermediate objective** in a multi-stage task.'

"role": "assistant".

},

"content": "<think>I've turned left. I now see a wooden panel or wall on the left-hand side, and tables and cabinets in the center and on the right-hand side. I still don't see a clear door. I'll continue turning left to scan the area.</think><action>TurnLeft</action>'

"images": ["embrace3k_train_image/SuburbNeighborhood_Day/x03703_y01064_z-0132_door0/000.jpg", "embrace3k_train_image/SuburbNeighborhood_Day/x03703_y01064_z-0132_door0/001.jpg"

Figure 5: Examples of a training prompt used for supervised fine-tuning with reasoning. The input includes the task instruction, a full history of egocentric observations and actions, and the model is expected to output both a natural language explanation (<think>) and the next action (<action>).



Figure 6: Examples of a training prompt used for supervised fine-tuning or reinforcement learning without reasoning. The model receives the task instruction, full visual history, and past actions, and directly predicts the next action in <action> format.

Instruction: Go to the red banner hanging on the wall.



MoveForward I need to go to the red banner. I see a red banner on the right-hand side of my view, hanging on a distant wall. To get closer, I will move forward.



MoveForward The red banner is still visible on the righthand side of my view, a bit closer now. I will continue to move forward to approach it.



MoveForward I am still approaching the red banner, which is on the right-hand side of my view. I will move forward again.



TurnRight The red banner is on the right-hand side of my view. To face it more directly and align my path towards it, I will turn to the righthand side.



Mover-orward After turning, the red banner is now more centered in my view, slightly on the righthand side. I will move forward to get closer to it.



TurnRight Ive moved closer, and now the red banner is on the left-hand side of my view, on a wall that forms a corner. The passage leading to the banner appears to be to my right-hand side. To enter this passage and continue towards the banner, I will turn to the righthand side.



MoveForward After turning, I am in a corridor, and the red banner is on the wall to my left-hand side. I will move forward to get closer to it.



The red banner is still on the wall to my lefthand side. I am getting closer. I will move forward again.



TurnLeft I am now quite close to the red banner, which is prominently on my left-hand side. To face it directly, I will turn to the left-hand side.



IUTLET After the turn, the red banner is now in front of me, slightly on the right-hand side of the center of my view. To center it, I will turn slightly to the left-hand side.



MoveForward The red banner is now mostly in the center of my view, perhaps slightly to the left-hand side. I am very close to the wall it's on. I will move forward to get directly beneath or in front of it.



I've moved as close as I can to the wall. The red banner is hanging above me. To confirm I'm at the target, I need to look up at it.



Figure 7: Visualization of a **Basic** task. The target object is clearly visible from the start, and the agent moves directly toward it with minimal ambiguity. Reasoning focuses on spatial orientation and efficient path alignment.

Instruction: Search the environment for a fire extinguisher and navigate to it.



TurnLeft I need to find a fire extinguisher. Currently, I see a room with a doorway on the left-hand side leading to stairs, and walls on the center and right-hand side. I don't see a fire extinguisher. To explore the room and search for it, I'l turn left to scan my surroundings.



TurnLeft I've turned left. The doorway is now more in the center of my view. I still don't see a fire extinguisher. I'll continue turning left to see more of the room.



TurnLeft After turning left again, the doorway is now on the right-hand side of my view. I see a wall in front of me and on my left-hand side. No fire extinguisher is visible. 'Ill keep turning left to continue my search.



I've turned further left. Now I see a wall with some pipes in the center and on the lefthand side. The doorway is no longer in view. I still haven't spotted a fire extinguisher. I'll turn left again to see what else is in this room.



TurnLeft I've turned left again. A new area with windows is now visible on the left-hand side of my view. The wall with pipes is now on the right-hand side. I don't see a fire extinguisher yet. I'll turn left noce more to get a better look at this new area.



MoveForward After turning. I now see a longer room or corridor in front of me, with windows on the left-hand side. The wall was previously looking at is on the right-hand side. I don't see a fire extinguisher immediately, but this new area needs to be explored. I'll move forward to investigate.



TurnRight I've moved forward into the corridor. There are windows on the left-hand side. On the right-hand side, I see a wall, and importantly, a red object mounted on it. This red object is a fire extinguisher, my target. It's on the right-hand side of my view. To face it more directly, I will turn towards the right-hand side.



I've turned towards the right-hand side, and the fire extinguisher is now more clearly visible on the wall on the right-hand side of my view. To reach it, I need to move forward.



MoveForward I've moved closer. The fire extinguisher is still on the right-hand side of my view, but it's more centered now. I need to get right next to it. I'll move forward again.



MoveBackward Ive moved forward, but now the fire extinguisher is no longer visible. Is see a door and a keypad on the right-hand side. It seems I might have moved too dose to the wall or slightly past the fire extinguisher. TII move backward to try and get it back in view.



I've moved backward. The fire extinguisher is still not visible. I see the door and keypad more clearly on the right-hand side. I recall the fire extinguisher was next to this door. I'll mycht still be too close or misaligned. I'll move backward again to get a better vantage point.



After moving backward again, I can now clearly see the fire extinguisher. It's mounted on the wall in the center of my view, next to the door. I have successfully reached the fire extinguisher. My task is complete.

Figure 8: Visualization of an **Exploration** task. The target is initially out of view, requiring the agent to conduct visual search by turning and scanning the scene before identifying and approaching the goal.

Instruction: First go to the garden lamp closest to you, then go to the red car.



TurnRight My task is to first go to the garden lamp closest to me. I see two garden lamps. One is further away, near the house entrance on the left-hand side. The other is attached to the garage wall on the right-hand side and appears closer. To approach the closer lamp on the right-hand side. I need to turn right to center it in my view.



MoveForward I have turned right, and the garden lamp on the garage wall is now roughly centered in my view. It's the closest one I identified. My goal is to go to it, so I will move forward.



LookUp I've moved closer to the garage wall. The garden lamp, my target, is mounted high on the wall, mostly above my current field of view. To confirm I'm at the correct location and get a better view of the lamp, I need to look up.



MidwayTarget I am now looking up directly at the garden Iamp attached to the garage. This is the closest garden Iamp, as required by the first part of the instruction. I have reached the first target.



TurnLeft I have completed the first stage of the task. Now I need to go to the red car. I am currently looking up at the lamp, the red car started approaching the lamp, the red car vight and moved forward to reach the lamp, the car should still be generally to my left. I need to turn left to locate the red car.



I turned left, and now I can see the red car appearing on the far left-hand side of my view. I need to continue turning left to bring it more towards the center.



TurnLeft I turned left again. The red car is now more visible, positioned on the left-hand side of my view. To face it directly for approach, I need to turn left slightly more.



Mover orward After turning left, the red car is now centered in front of me. My current goal is to go to the red car, so I will move forward towards it.



Figure 9: Visualization of a **Multi-stage** task. The agent must complete a sequence of subgoals in order. The reasoning must shift at the appropriate time, and target persistence is required to recall goals after partial occlusion or disorientation.

Instruction: Go to the leftmost crate visible under the archway.



My goal is to find the leftmost crate under the archway. Currently, I see a wall in front of me and a starcase on the right-hand side. On the left-hand side, I can see part of an archway and the crates, I need to turn towards the left-hand side.



TurnLeft I turned left. Nove I have a clearer view of the area with the archway on the left-hand side and in the central. I can see multiple crates stacked under the archway. The task requires me to go the leftmost crate visible under the archway. To center this area in my view and prepare to approach it, I will turn left again.



MoveForward After turning left again, the archway area is now mostly in the center and on the righthand side of my view. I can see the crates dearly under the arches on the right-hand side of the open space. The task is to go to the leftmost crate visible under the archway. All visible crates are clustered together under the arches further away. To got closer to the crates and identify the specific fetmost one, inced to move forward.



I moved forward. The crates under the archway are closer now and are primarily located in the center and on the right-hand side of my view. I need to reach the leftmost crate within that group. From my current perspective, the cluster containing the leftmost crate is slightly to the right-hand side. To align mysel better with the target crates, i will turn slightly to the right-hand side.



I turned towards the right-hand side. The cluster of crates is now centered in my view. I can distinguish several crates. By target is the leftmost one visible under the archway, which appears to be the reclangular wooden box on the left degle of the main stack. To approach this target crate, I will move forward.



Invoid forward and am now much closer to the crates. The leftmost crate, a rectangular wooden box, is directly in front of me, sightly to the left-hand side of the center of my view. Other crates are visible behind it and on the right-hand side. I need to get right next to the target crate to complete the task. I will move forward again.



I moved forward one last time. I am now standing directly in front of the target crate, which is the leftmost crate that was visible under the archway. It is centered in my view and appears to be within inferencion distance. Since I have reached the specified destination, I will now finish the task.

Figure 10: Visualization of a **Dynamic Spatial-Semantic** task. The instruction involves relational spatial references such as "leftmost" and "under the archway", which require the agent to resolve dynamic spatial semantics as its viewpoint evolves.

Instruction: The door is right there. Head over and open it.



TurnRight My goal is to go to the door and open it. I can see a door in the center-left portion of my view, beyond a couch. The instruction says "the door is right there". To get a better path towards it and explore the space, I'll turn to the right-hand side.



MoveForward Now the door is on the left-hand side of my view. The couch is also on the left-hand side. I see a kitchen area on the right-hand side. To get closer to the door, I need to move forward.



TurnLeft I've moved forward. The door is still on the left-hand side. There's a dining table and chairs in front of me, and the door is beyond them. To align myself better with the door, I will turn to the left-hand side.



MoveForward After turning, the door is now mostly in the center of my wex, slightly to the left. A white cabinet is on the left-hand side of the door. The dining table is now on my right-hand side. I need to move closer to the door, so I'll move forward.



Figure 11: Visualization of an **Interaction: Open Door** task. The agent must navigate to a nearby visible door and explicitly trigger an OpenDoor action. Reasoning must ensure spatial alignment with the door and recognize successful interaction.

Instruction: Pick up the pillow and place it near the yellow ladder.



TurnLeft My task is to pick up the pillow and place it near the yellow ladder. I see the pillow in the center of my view, on the ground. To better align myself to approach it, I will turn slightly to the left, as it appears to be slightly more on the left-hand side of my immediate path forward.



MoveForward The pillow is now more centered in my view slightly on the left-hand side. I need to get closer to it to pick it up. I will move forward.



MoveBackward I seem to have moved too far forward, as the pillow is no longer visible in my current view. I need to see the pillow to pick it up. I will move backward to bring it back into view.



The pillow is now visible again, in the center and lower part of my view, very close to me. To better position myself to pick it up. 1 will turn to the right. It seems to be slightly to my right now, or turning right will give me a better angle.



PickObject The pillow is now directly in front of me, in the center of my view, and within reach. I will pick up the pillow.

MoveForward

The yellow ladder is now closer and still in

the center of my view. I need to get even closer to place the pillow near it.



I have picked up the pillow. I will look at my hand to confirm that I am holding it.

DropObject

I am now very close to the yellow ladder,

which is in the center of my view. This seems like a suitable position to place the pillow near the ladder. I will drop the pillow



TurnRight I have confirmed I am holding the pillow. Now I need to find the yellow ladder. I recall seeing a yellow ladder earlier. I will turn to the right to locate it.



LookDown I have dropped the pillow. I will look down to confirm that the pillow has been placed on the ground near the yellow ladder.



MoveForward I can see the yellow ladder in the center of my view, in the distance. I need to move closer to it to place the pillow.



I can see the pillow on the ground in the center of my view, at the base of the yellow ladder. I have successfully picked up the pillow and placed it near the yellow ladder. The task is complete.

Figure 12: Visualization of an **Interaction: Pick and Drop** task. The agent must locate a manipulable object, perform a pick-up, navigate to a reference location, and execute a drop. The reasoning combines manipulation awareness with spatial goal alignment.

Instruction: Go find the refrigerator.



Figure 13: Exploration task example for Gemini-2.5-Pro. The model fails to locate the refrigerator and enters a loop of repeated turns.



Figure 14: Exploration task example for GPT-40. The model explores multiple rooms but fails to find the refrigerator and times out.

Instruction: Go find the refrigerator.



TurnLeft I need to find the refrigerator. I am currently looking at what appears to be a living room. I see a couch and a TV in the center and on the righthand side, and a door on the lefthand side. I don't see a refrigerator here. I will turn left to scan the environment and see if it's located elsewhere in this area or beyond.



TurnLoft I've turned left. Now I see stairs on the left-hand side. The living room area is more central and slightly on the right-hand side. I still don't see a refrigerator. Til continue turning left to explore further, as refrigerators are often found in kitchens or utility areas.



Iverturned further left. The stairs are now in the center and on the righthand side. I see a window on the left-hand side. I still haven't located the refrigerator. Til continue turning left to explore more of the area.



TurnLeft I am now facing the stairs and a window. There's no refrigerator here. I need to keep turning left to explore the area.



l've turned left again. I see a window on the left-hand side and stairs on the right-hand side. No refrigerator is visible. I need to keep searching, so I'll turn left once more.



TurnLeft I've turned left again. I see a dining table with chairs and a window on the right-hand side. The stairs are now mostly out of view on the righthand side. I still haven't located the refrigerator. I'll continue turning left.



TurnLeft I've turned left again. Now I see the dining table and chairs more centrally, with a kitchen area behind them. The refrigerator is visible on the far left-hand side, past the dining area. I need to turn left further to bring the refrigerator into the center of my view.



TurnLeft I've turned left again. The refrigerator is now clearly visible on the left-hand side. The drining table is on the right-hand side. Since the refrigerator is my target and it's on the left-hand side, I need to turn left slightly more to face it directly before moving towards it.



MoveForward The refrgerator is now mostly centered in my view, slightly on the left-hand side. I have identified my target. Now I need to move forward to approach it.



Finish I have moved forward and am now standing directly in front of the refrigerator. It takes up most of my view. I have successfully found the refrigerator as instructed. The task is complete.

Figure 15: Exploration task example for Qwen2.5-VL-7B fine-tuned on EmbRACE-3K. The model performs a full scan, identifies the refrigerator, and completes the task.

Instruction: Go to the second pillar on the left in the initial view.



TurnLeft The target is to the left and in the distance. I need to turn left to face the pillars and then move towards them.



ALC: N

TurnRight

The agent is too close to a

pillar, blocking the view. Turning right will help to get a clearer view of the

surroundings and identify the

target pillar.

MoveForward MoveForward I have turned left and can see the first pillar clearly. The second pillar is further ahead and slightly to the left. I need to move forward to get closer to it. The agent is at the first pillar on the left. To reach the second pillar on the left, the agent needs to continue moving forward.



I have reached the first pillar. Now I need to turn left to find the second pillar on the left.



MoveForward The agent is facing the side of the first pillar. To get to the second pillar, I need to move past this one first.



TurnLeft I am too close to the first pillar. I need to turn left to see the second pillar and then move towards it.



I am currently facing the first pillar. I need to turn right to move past it and then find the second pillar on the left.



The agent is facing a wall and needs to turn to find the target or a clear path. The trajectory has several turns. trajectory has several turns, suggesting the agent might be disoriented. Turning right is a reasonable next step to explore the surroundings



The agent is currently facing a wall and the trajectory shows a lot of turning. Turning left will help reorient the agent to find the pillars mentioned in the task.



The agent is facing a wall and the trajectory shows a lot of turns, suggesting it might be stuck or lost. I need to turn to reorient and find the target pillars. Turning right seems like a reasonable first step to get a new perspective.



The agent is next to a large pillar, likely the first one encountered. The trajectory shows repeated turning, indicating difficulty navigating. To reach the "second pillar on the left" (which was further away in the initial view), the agent needs to move past this current large pillar. Turning right should allow the agent to see the path forward alongside this pillar.



MoveForward



The agent is on a path leading towards some pillars teacing towards some piliars. The task is to go to the second pillar on the left. The current view shows pillars in the distance, and the most direct way to approach them is to continue moving forward along the visible path.



TurnRight The agent is too close to the first pillar and needs to turn right to move past it and find the second pillar.



disorientation. I need to reorient to find the target Turning left seems like a reasonable first step to get a better view of the area and potentially spot the target pillars.

MoveForward The agent is very close to a large pillar on the left, likely the first pillar. To find the second pillar on the left, the agent needs to move forward along the visible path.



The trajectory has many turns, suggesting disorientation. The current view shows a path towards some pillars; I need to move forward to get a better view and assess which is the target pillar based on the initial view.



The agent is moving towards the pillars. The path ahead is clear. Continuing forward will bring the agent closer to the potential targets.



MoveForward The agent is on a path leading towards the pillars. The task is to go to the "second pillar on the left". The current view shows a pillar on the left and another one further away. Continuing to move forward will bring the agent closer to both, allowing for better identification and approach to the target.



MoveForward The agent is approaching the first pillar on the left. The target is the second pillar on the left, which is further ahead. Continuing to move forward is the correct action.



MoveForward The agent is approaching the first pillar on the left. The task is to reach the second pillar on the left. Therefore, the agent should continue moving forward to find the second pillar



Finish The agent has moved forward multiple times and is now very close to a pillar. Assuming the previous mavigation was towards the correct target, the agent should now be at the "second pillar on the left in the initial view".



Figure 16: Dynamic spatial-semantic task example for Gemini-2.5-Pro. The model fails to maintain reference alignment and becomes disoriented near the first pillar.

Instruction: Go to the second pillar on the left in the initial view.



Mover orward The second pillar on the left is visible in the distance, and there are no immediate obstacles. Moving forward will efficiently advance toward the target.



Plate-

MoveForward

The trajectory shows

The trajectory shows consistent forward movement, but the second pillar on the left is not yet reached. The pillar is visible and close, suggesting continued forward movement is appropriate.

MoveForward

MoveForward The second pillar on the left is visible in the initial view, and I have been moving forward towards it. The trajectory indicates I am close to the larget, so continuing to move forward seems ontimal

optimal.

is still ahead, so continuing forward is optimal.



The images show consistent progress toward the pillars, with no visible obstacles. The second pillar on the left is clearly visible, and moving forward seems efficient.



MoveForward MoveForward The trajectory shows consistent forward movement towards the pillars, and the second pillar on the left is now visible and closer. There are no obstacles directly ahead, so continuing forward is efficient.



MoveForward mover-orward The agent has made consistent progress towards the target, which is the second pillar on the left. The pillar is now clearly pinal is now clearly visible and in close proximity, suggesting that continuing to move forward is likely the most efficient action.



MoveForward The trajectory shows consistent forward movement, indicating progress toward the target pillars. The second pillar on the left is now closer, suggesting that continuing to move forward is efficient for reaching the target.



MoveForward The second pillar on the left is visible and close, indicating that continued forward movement will efficiently reach the target without obstacles.



MoveForward The trajectory shows The trajectory shows repeated 'MoveForward' actions without deviation, indicating progress toward the target. The second pillar on the left is visible and close in the last image. Continuing forward should efficiently complete the task.



CAN'S MoveForward

The images show the agent is close to a

building with pillars. The trajectory

suggests some inefficient turning movements, but the agent is progressing toward the pillars. The

second pillar on the left is visible, so moving forward will likely get closer to the target.

MoveForward MoveForward Mover-orward The trajectory shows inefficient turning movements, and the second pillar on the left is visible in the current view. Moving forward will help odwarde towards the The target pillar is The target pillar is visible in the distance, slightly to the left. The trajectory shows multiple turns, but the current view suggests moving forward is advance towards the target pillar. clear and efficient



Cost Joint

MoveForward

Mover-orward The trajectory shows some redundant turns but overall progress is toward the pillars. The second pillar on the left is visible and slightly ahead; moving forward will bring the

forward will bring the agent closer.

MoveForward

The agent appears to be heading toward the

building with pillars but is currently facing

stairs. The trajectory has inefficient movements; the agent needs to continue moving forward to

reach the target pillar

The trajectory indicates consistent forward movement, and the images show the agent has passed the first pillar and is approaching the second pillar on the left. Continuing forward should reach the target efficiently. Movel-orward actions suggest progress, but the target pillar is not visible in the latest images, indicating a need to reassess direction. Turning might help locate the pillar.



no visible obstacles

Statt.

MoveForward

Mover-orward The trajectory shows inefficient turns, and the second pillar on the left is visible in the initial image. Moving forward towards the visible pillars is likely to advance the back

to advance the task efficiently.

TurnRight

The agent is currently facing stairs and a building, which suggests it may have

moved away from the pillars. The task requires reaching the

second pillar on the left, which is likely left, which is likely back in the initial direction. Turning around would help reassess the surroundings and locate the pillars.

TurnRight The trajectory shows a series of forward movements followed by a left turn. The first person images person images indicate that the agent has moved away from the pillars. To correct the path and head towards the second pillar on the left, a TurnRight action is roaded

needed

MoveForward

mover-orward The second pillar on the left is visible in the initial view, and I am currently facing towards it. The trajectory shows some inefficient movements, but continues to

but continuing to move forward will likely bring me clo to the pillar.

MoveForward

The images show the agent is near a building with pillars, but the trajectory

suggests inefficient movement with multiple turns. The target pillar seems visible in the distance

from the initial view, so moving forward might be the best action to approach it.



TurnLeft The trajectory shows inefficient movement with unnecessary turns. The target pillars are visible in pillars are visible in the first image but not in the others, indicating a wrong direction. I should turn left to reorient towards the pillars.

MoveForward

The second pillar or

the left is visible and close in the current view. Moving forward will efficiently approach the target.

MoveForward

The pillar we need to reach is visible in the

distance, slightly to the right. The

trajectory shows some inefficient movements, but the current view suggests a clear path forward towards the

pillar.



T

TurnLeft The trajectory shows The trajectory shows a repeated pattern of turning left and right, indicating potential confusion or a lack of progress. The second pillar on the left is not unat visible in the yet visible in the current view. Turning left may help align the view towards the target location.



TurnLeft The second pillar on the left is visible in the images, and the trajectory shows inefficient movements like repeated turns. The pillar is slightly to the left of the current

MoveForward



The first-person view images show the agent is near stairs and construction, The trajectory shows inefficient turning, and and construction, indicating proximity to the target pillar. The trajectory shows repeated movements, suggesting inefficient progress. The agent should continue moving forward towards the visible pillar to reach the target efficiently.



TurnLeft The trajectory

The trajectory indicates some inefficient turns. The second pillar on the left is visible in the initial view, suggesting the need to adjust direction. The most recent images show a path leading away from the pillar, indicating a need to orrere course. correct course.



MoveForward Based on the ima Based on the images, the second pillar on the left in the initial view is visible and appears to be straight ahead. The trajectory shows a lot of turning, which may not be necessary. Moving forward seems to be the most direct actio



the recent images indicate the agent is near stairs and scaffolding, possibly off course. I should correct the direction to approach the pillars.

Figure 17: Dynamic spatial-semantic task example for GPT-40. The model initially heads toward the correct region but fails to preserve ordinal reference during task.



Instruction: Go to the second pillar on the left in the initial view.



MoveForward My task is to go to the second pillar on the left in the initial view. I see three pillars in the distance. One pillar is on the left-hand side, and two more pillars are in the center an slightly to the right-hand side. The instruction specifies the 'second pillar on the left, which corresponds to the pillar in the center of my view. To get closer to II, I will move forward.



MoveForward I have moved forward. The first pillar is now more prominent on the lefthand side. The second pillar, my target, is still visible further ahead, slightly to the right of center. I need to continue moving towards it, so I will move forward again.



MoveForward I am getting closer to the pillars. The first pillar is on the left-hand side, the second pillar (my target) is in the center-left, and the third pillar is on the right-hand side. The target remains centered in my view, so I will continue moving forward.



MoveForward I am approaching the target pillar, which is the second one from the left in the initial view. It is still centered in my view. The third pillar is visible on the right-hand side. I need to get closer, so I will move forward.



I am continuing my approach. The first pillar is on the left-hand side, and the second pillar, my target, is slightly left of center. I need to move forward again.



MoveForward I am continuing my approach. The first pillar is on the left-hand side, and the second pillar, my target, is slightly left of center. I need to move forward again.



MoveForward I am very close to the target pillar now. It's large and centered in my view. The third pillar is still visible on the right-hand side. I need to move forward one more step to reach the base of the target pillar.



I an very close to the base of the target pillar now. It's centered in my view. The top part is starting to move out of view. I need to take another step forward to reach it.



Finish I have moved forward and an now positioned directly at the base of the second pillar, which was the target specified in the instruction (second pillar on the left in the initial view). The pillar is large and centered in my view. I have reached my destination. Therefore, I will finish the task.

Figure 18: Dynamic spatial-semantic task example for Qwen2.5-VL-7B fine-tuned on EmbRACE-3K. The model correctly interprets and tracks the second pillar through egocentric perspective changes.

Instruction: First, walk to the left pillar, then go to the right pillar.









TurnLeft

The agent has reached the left pillar and has turned multiple times. It needs to find the right pillar. Turning left will help reorient the agent to find the next target.

TurnRight

The agent has completed the first sub-goal and is now facing away from the pillars. It needs to turn to find the right pillar.

TurnRight I have reached the first pillar. Now I need to find the right pillar. I will turn right to scan for it.

TurnRight I have turned left twice and am now facing a rock. I need to find the left pillar, which was previously to my right. I will turn right to reorient myself towards the pillars.

*





The agent has turned back to face the pillars. The left pillar is the first target. The agent should move forward towards it.



TurnRight I have reached the left pillar and marked it as a midway target. Now I need to find the right pillar. I will turn right to look for it.



The agent has moved forward and is now better aligned with the left pillar, which is the first target. The next logical step is to continue moving towards it.



The agent has reached the left pillar and needs to turn to face the right pillar, which is currently behind it or to its side.



I have reached the left pillar. Now I need to find the right pillar. I will turn right to scan the environment for the right pillar.



MidwayTarget MidwayTarget The agent has reached the left pilar. The next step is to acknowledge this sub-goal and then proceed to the right pillar.



TurnLeft The agent has been turning back, and forth, indicating it might be disortented. The pillars are not visible in the current view. I need to turn to find the tel pillar, which is the first target. Given the recent turning pattern, i will try turning left to re-scan the environment.



The agent has made progress towards the left pillar. The next action should be to continue me

TurnRight

The agent has reached the left pillar and now needs to find and go to the right pillar. It's currently facing the left pillar, so it needs to turn to locate the right pillar.

TurnLeft I have reached the first pillar. Now I need to find the right pillar. The current view is a rock wall, so I need to turn to locate the next target. Turning left is a good start to scan the environment. TurnLeft

TurnRight I have reached the left pillar. Now I need to go to the right pillar. I can see the right pillar to my right. I need to turn right to face it.



TurnRight I have reached the left pillar. Now I need to go to the right pillar, which is currently to my right. I should turn right to face it.

TurnRight

I have reached the left pillar. Now I need to go to the right pillar. I will turn right to orient myself towards the right pillar.

TurnLeft

The agent has been turning a lot and is not facing a pillar. I need to turn to locate the left pillar.

TurnLeft

TurnLeft I have reached the first pillar. Now I need to go to the right pillar. I need to turn to face the right pillar, which was visible in the first image. The last two actions were "TurnRight', so I will try "TurnLeft' to reorient towards the other millar

MoveForward I have reached the left pillar and turned right. The right pillar is now visible and I should move towards it

TurnRight

I have reached the left pillar. Now I need to find and go to the right pillar I will turn to look for it.

TurnLeft

The agent has been turning a lot, suggesting it's disoriented. I need to find the left pillar. Turning left will help me scan the environment for it.

TurnRight The agent has reached the first pillar and is now facing away from the general direction of the pillars. The next step is to find the right pillar, so the agent needs to turn to reorient itself.



TurnRight TurnRight I have reached the first pillar. Now I need to find and go to the right pillar. The current view doesn't show the right pillar, so I need to turn to locate it. The trajectory shows a lot of turning, so I will continue to turn to scan the environment.



TurnLeft The agent has been turning a lot. I need to turn to find the left pillar, which is the first target. The current view does not show the pillar.



TurnLeft The agent is not facing the pillars and the trajectory shows a lot of turning. I need to turn to face the pillars, which were visible in the first image, to proceed with the task. Turning left seems like a reasonable action to receivent



Figure 19: Multi-stage task visualization for Gemini-2.5-Pro. The model reaches the first subgoal but fails to localize the second, exhibiting repeated turning and no successful task completion.

Instruction: First, walk to the left pillar, then go to the right pillar.



iuggests fliciently he first







the agen 'Midway'

The the wal







repe that with





Figure 20: Multi-stage task visualization for GPT-40. The model completes the first-stage goal but fails to relocate the second, with oscillatory behavior and multiple redundant midway signals.

The has

22

Instruction: First, walk to the left pillar, then go to the right pillar.



Wover or ward My task is to first go to the left pillar. I see two pillars in the distance: one on the left-hand side and one on the right-hand side. To get closer to the area where the pillars are located, I will move forward.



MoveForward I am still approaching the pillars. The left pillar is on the left-hand side, and the right pillar is on the right-hand side. I need to continue moving forward to reach the left pillar.



I am getting closer to the pillars. The left pillar is now more prominent in the center-left of my view. I will continue moving forward towards it.



MoveForward I am very close to the base of the left pillar, which is now primarily on the left-hand side and center of my view. I need to take one more step forward to arrive at it.



I have arrived at the base of the left pillar, which is directly in front of me. This completes the first stage of my task.



I have reached the first larget (the left pillar). Now I need to proceed to the second stage: 'go to the right pillar. From my current position, I can see the right pillar on the right-hand side of my view. I need to turn towards it.



TurnRight I am turning towards the 'right pilgar'. I can see the pillar on the right-hand side of my view now, and another one further away on the left-hand side. I need to continue turning right to face the target pillar more directly.



TurnRight I turned right again. I can now see the wooden waikway leading towards the area where the right pillar is located, which is currently on the right-hand side of my view. I need to turn further right to align myself with the walkway and approach the pillar.



MoveForward The 'right pillar' is now mostly centered in my view, slightly towards the right-hand side. I need to move forward to approach it.



I am moving towards the right pillar. It is still in front of me, slightly on the right-hand side. I will continue to move forward.



MoveForward I am getting closer to the right pillar. It is now larger and positioned slightly on the righthand side of my view. I will continue moving forward.



I am very close to the base of the right pillar, which is on the right-hand side of my view. To face it more directly and confirm I've reached it, I will turn towards the right-hand side.



MoveForward After turning slightly, the right pillar is now centered in my view again. I will move forward to approach it.



MoveForward I am getting very close to the right pillar. It is now large and primarily on the right-hand side of my view. I will move forward again to reach it.



I am very close to the right pillar, which is now slightly on the lefthand side of my view. I need to turn right to face it more directly.



MoveForward After turning right, the right pillar is now centered in my view. I will move forward to reach it.



I am very close to the right pillar, which is now slightly on the lefthand side of my view. I'll turn slightly to the right to face it more directly before finishing.



MoveForward After turning right, the 'right pillar is now centered in my view. I will move forward to reach it.



TurnRight I am very close to the right pillar, which is now slightly on the lefthand side of my view. I'll turn slightly to the right to face it more directly.



The "right pillar" is now centered in my view. I have reached the second target. My task is complete.

Figure 21: Multi-stage task visualization for Qwen2.5-VL-7B fine-tuned on EmbRACE-3K. The model successfully completes both subgoals with coherent orientation and target tracking.

References

- [1] 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.
- [2] DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens. *arXiv preprint arXiv:2403.05530*, 2024.
- [3] Fangwei Zhong, Kui Wu, Churan Wang, Hao Chen, Hai Ci, Zhoujun Li, and Yizhou Wang. Unrealzoo: Enriching photo-realistic virtual worlds for embodied ai. *arXiv preprint arXiv:2412.20977*, 2024.