

A. Details of THEORY OF SPACE

A.1. Formal Definition

Definition: THEORY OF SPACE

Ability to **construct**, **revise**, and **exploit** an internal spatial belief.

We formalize THEORY OF SPACE within a partially observable framework over a spatial structure $S \in \mathcal{S}$. The agent interacts with S to generate a history $h_t = (o_{0:t}, a_{0:t})$, where o and a denote observations and actions. We define THEORY OF SPACE as the capacity to manipulate a probabilistic belief B_t through three core operations:

1. **Construct:** To form a globally consistent internal spatial belief by actively seeking out and integrating partial observations. Formally, the agent integrates h_t to approximate the true posterior, denoted as $B_t(S) \approx P(S | h_t)$.
2. **Revise:** To dynamically update the internal belief by using new information acquired through further exploration to resolve conflicts with prior beliefs. Upon an environmental shift $S \rightarrow S'$, the agent utilizes exploratory actions Δh to minimize the divergence from the new ground truth, i.e., $B_{t+\Delta t} \rightarrow P(S' | h_{t+\Delta t})$.
3. **Exploit:** To utilize the current belief to support spatial tasks. The agent utilizes a policy π conditioned on the belief, $\pi(a_t | B_t)$, to perform a downstream task \mathcal{T} . In a benchmark context, we measure the *value of belief* by the performance metric \mathcal{J} achieved by this policy: $\mathcal{J}(\pi(\cdot | B_t), \mathcal{T})$.

A.2. Downstream spatial tasks

Drawing on prior work [36, 47], we define tasks to evaluate an agent’s **Route** and **Survey** knowledge. An overview of the tasks is present in Figure 4 and Table 5.

A.3. Assessment Dimensions

Here we show the detailed definition of five assessment dimensions.

(D1) Belief Construction Efficiency. *Measures how efficiently the agent collapses spatial uncertainty during exploration.* We quantify this using a normalized information gain metric, \mathcal{E} . Let M be the number of possible positions for any object at the start of exploration (a uniform prior), and let C_i be the number of positions for object i that remain consistent with all observations gathered by the agent (calculated by AC-3 algorithm, check Appendix B.2). The efficiency is calculated as $\mathcal{E} = 1 - \frac{\sum_{i=1}^N \log_2 \max(1, C_i)}{N \log_2 M}$. This score ranges from 0 (no information gained, $C_i = M$) to 1 (all objects perfectly localized, $C_i = 1$). Note that it can also be used to calculate the accumulated information gain at each step. Information gain is mainly used in text-based environments, since vision-based environments have direct

access to scenes without such ambiguity. Therefore, for vision-based environments, we directly use node coverage to measure exploration efficiency.

Belief Representation and Quality Assessment. A core contribution of THEORY OF SPACE is disentangling spatial memory from spatial inference. We structurally decompose the probed cognitive map into two components:

- **(D2) The Cognitive Map (Observed):** *Measures fidelity and coherent integration of observations over time.* We evaluate using two criteria: (1) Correctness, alignment with ground truth, computed as a composite of positional, directional, and facing accuracy; and (2) dynamic reasoning diagnostics, including **Perception** quality, **Self-tracking**, **Stability**, and **Local** \leftrightarrow **Global Consistency**, reflecting internal coherence such as the absence of contradictions within the relational graph and between maps and relations.
- **(D3) The Uncertainty Map (Unobserved):** *Measures how well the agent models plausible hypotheses about unobserved regions.* We assess **Uncertainty Modeling** by providing a candidate set of positions formed by randomly sampled points from both observed and unobserved areas, and measuring the agent’s ability to identify valid locations via F_1 .

This separation lets us diagnose whether failures stem from *misestimating* the observed world or from insufficient *reasoning* about what remains unobserved.

(D4) Belief Revision. *Measures the agent’s ability to revise its spatial belief under latent environment changes.* We evaluate this using the **False Belief** task (§5.2), where objects are covertly manipulated (translated or rotated) following the initial exploration. The agent must re-explore to detect these discrepancies; we measure the accuracy of these identified changes (both object identity and transformation type) using the F_1 score. Furthermore, we introduce **Belief Inertia** to quantify whether belief revision remain biased toward obsolete priors.

(D5) Belief Exploitation Success. *Measures task success when the agent must utilize its spatial belief.* For tasks involving spatial relations (direction, persp.take, action2view), we score direction and distance separately, awarding 0.5 for each correct component. For tasks that output coordinates (view2loc, alloc.map), we compute a coordinate similarity score.

A.4. Cognitive Map Probing

We give detailed definition of metrics and scores defined for cognitive map probing here.

Metrics. We evaluate cognitive map using three complementary metrics. *Positional accuracy* ($pos.acc$) is the Euclidean similarity between predicted and true object coordinates: $(K/N) \cdot e^{-RMSE/L}$, where RMSE is the root

Dynamic Group	Belief on Route	Belief on Survey
Static	Pairwise Relation (<i>direction</i>) report allocentric direction and distance from A to B .	Allocentric Mapping (<i>alloc.map</i>) predict global coordinates (and headings) for all objects.
Forward Dynamics	Perspective Taking (<i>persp.take</i>) output the observation from a specified object’s perspective.	Mental Rotation (<i>ment.rot</i>) predict the sequence of front-facing objects during a 360° self-rotation.
	Action-to-View (<i>act2view</i>) given a sequence of Goto/Rotate, predict the final observation (one object in FOV with ego direction/distance bins).	Location2View (<i>loc2view</i>) given a global pose, predict the observation (one object in FOV with ego bins/distances).
Backward Dynamics	Perspective Decision (<i>perc.dec</i>) infer which object’s perspective the agent is currently adopting.	View2Location (<i>view2loc</i>) localize the agent (and optionally orientation) from a target observation under the map.
	View-to-Action (<i>view2act</i>) recover an action sequence that produces a target observation.	

Table 5. **Task suite comparison:** Route belief emphasizes egocentric, step-by-step path reasoning; Survey belief emphasizes allocentric mapping and novel view inference.

988 mean squared error between predicted and ground-truth ob-
 989 ject positions, L is the RMS ℓ_2 -norm of the positions of
 990 all objects in the scene, and K/N is the coverage (the
 991 ratio of the number of predicted objects K to the number of
 992 ground-truth objects N). *Directional accuracy* (*dir.acc*) is
 993 the accuracy of directional relationship between each pair
 994 of objects. *Facing accuracy* (*facing.acc*) is the fraction of
 995 objects whose predicted facing matches the ground truth.

996 Using *global* and *local* belief representations, we compute
 997 a set of diagnostic scores at each turn t (all per-turn except
 998 **Correctness**, which is computed only at the final turn after
 999 termination). Unless noted, scores are averaged over turns
 1000 and scenes:

1001 • **Correctness (final):** *Measures the accuracy of the*
 1002 *agent’s terminal global spatial belief.* At the last turn,
 1003 we evaluate the predicted global map and report a com-
 1004 posite score given by the (equally weighted) mean of the
 1005 three metrics defined above, with weights $1/3$ each. We
 1006 compute *dir.acc* only for correctness, since the global
 1007 cognitive map prioritizes consistent pairwise spatial rela-
 1008 tions.

1009 • **Perception:** *Measures how accurately the agent inter-*
 1010 *prets newly observed local structure.* We compare the
 1011 predicted local map to the ground-truth local map for the
 1012 current field of view (FOV), counting only objects that
 1013 appear in the FOV for the first time.

1014 • **Self-tracking:** *Measures how well the model estimates*
 1015 *its own pose over time.* We infer the agent’s pose from the
 1016 predicted global map and compare it against the ground-

truth agent state.

1017
 1018 • **Local \leftrightarrow Global consistency:** *Measures whether new* 1018
 1019 *local evidence is incorporated into the global belief co-* 1019
 1020 *herently.* Within the same turn, we compare local and 1020
 1021 global predictions to verify that newly perceived structure 1021
 1022 is integrated without contradictions. 1022

1023 • **Stability:** *Measures whether beliefs about previously* 1023
 1024 *observed objects remain non-degrading over time.* For 1024
 1025 each previously observed object, at every subsequent 1025
 1026 turn we check that its predicted state does not worsen; 1026
 1027 the per-check score is 1 if the prediction is no worse than 1027
 1028 in the previous turn. 1028

A.5. Belief Revision Metrics 1029

We provide detailed definitions of the metrics used to assess 1030
 belief revision here. 1031

Metrics. We evaluate performance along four comple- 1032
 mentary axes: 1033

1034 • **Identification Accuracy (F_1):** *How precisely the agent* 1034
 1035 *pinpoints which objects changed.* We compute the F_1 1035
 1036 score for detecting the subset of objects whose position 1036
 1037 or orientation shifted. 1037

1038 • **Average Steps:** *How efficiently the agent revises its be-* 1038
 1039 *liefs to completion.* We report *Total Steps* needed to 1039
 1040 identify all changes, and *Redundancy Steps*, defined as 1040
 1041 the number of steps taken after the last changed object 1041
 1042 has been observed. Ideally, *Redundancy* $\rightarrow 0$, indicating 1042
 1043 the agent recognizes when updating is complete. 1043



Figure 4. **THEORY OF SPACE exploitation task suite**: it covers **route**-level egocentric reasoning and **survey**-level allocentric mapping. Route tasks evaluate path-based inference and egocentric observations. Survey tasks test global mapping, geometric transformation, and perspective conversion. Together they cover both local navigation reasoning and global spatial abstraction.

- 1044 • **Belief Correctness:** *How accurate the updated beliefs*
1045 *are on the changed subset.* We compute correctness as in
1046 §5.1, but restrict evaluation to changed objects to isolate
1047 the fidelity of re-exploration.
- 1048 • **Belief Inertia:** *Whether updating remains systematically*
1049 *biased toward obsolete priors.* To quantify attraction
1050 back to pre-shift beliefs, we test whether the residual
1051 error of the updated belief aligns with the direction of
1052 the *old* belief. For each shifted object i , let \mathbf{b}_i^{old} denote
1053 the pre-shift belief, \mathbf{b}_i^{new} the post-revision belief, and
1054 \mathbf{g}_i^{new} the post-shift ground truth. Define the *prior-offset*
1055 and *post-revision error* vectors: $\mathbf{v}_i = \mathbf{b}_i^{old} - \mathbf{g}_i^{new}$, $\mathbf{e}_i =$
1056 $\mathbf{b}_i^{new} - \mathbf{g}_i^{new}$. We define positional inertia as

$$1057 s_i^{pos} = \underbrace{\frac{\mathbf{e}_i^\top \mathbf{v}_i}{\|\mathbf{e}_i\| \|\mathbf{v}_i\| + \epsilon}}_{\text{Directional alignment } (\cos \theta_i)} \cdot \underbrace{\exp\left(-\frac{\|\mathbf{b}_i^{new} - \mathbf{b}_i^{old}\|^2}{2\sigma^2}\right)}_{\text{Proximity weight } (w_i)}.$$

1058 Here $\cos \theta_i$ is large when the remaining error after up-

dating still points toward the obsolete location, while
1059 w_i downweights such alignment when the belief has
1060 moved far from \mathbf{b}_i^{old} . We set σ to a dynamic noise scale:
1061 the RMS localization error on the first re-observed
1062 *unchanged* objects during re-exploration; ϵ ensures numerical
1063 stability. Under unbiased updating, $\mathbb{E}[s_i^{pos}] \approx 0$,
1064 whereas $s_i^{pos} > 0$ indicates systematic pull toward the
1065 obsolete prior. For orientation shifts, we measure inertia
1066 via $s_i^{ori} = \mathbb{1}(\phi_i^{new} = \phi_i^{old})$, where ϕ denotes the
1067 predicted orientation. It flags failures to overwrite the
1068 obsolete facing direction.
1069

1070 B. Technical Details

1071 B.1. Benchmark Construction

1072 We expose the ToS world as a *Gym-like* interface [6]: agents
 1073 interact in discrete steps under partial observability at a reso-
 1074 lution of 384×384 to **construct** and **revise** an internal spatial
 1075 belief, which we later **exploit** in evaluation tasks. Scenes
 1076 are procedurally generated multi-room layouts on an $N \times M$
 1077 grid with n named indoor objects (each with integer (x, y)
 1078 and heading in $\{N, E, S, W\}$) and a randomized agent spawn
 1079 pose. We restrict multi-room layouts to a tree topology: the
 1080 room–adjacency graph is connected and acyclic (no loops).

1081 **Text-based World** At each step, OBSERVE returns a sym-
 1082 bolic snapshot of objects in the current room within a 90°
 1083 forward FOV. For every visible object we provide discretized
 1084 egocentric direction (e.g., *front-left*) and distance bins (e.g.,
 1085 *near/mid/far*), plus object identity and facing when deter-
 1086 minable. Egocentric observations are rendered with a 90° -
 1087 degree field of view (FOV), discretized into angular and
 1088 distance bins. Visibility is room-bounded; doorways act as
 1089 transparent portals only when the agent stands in them, en-
 1090 abling dual-room visibility. Optional noise modules perturb
 1091 bins for ablations.

1092 **Vision-based World** We procedurally generate scenes in a
 1093 3D simulator with two controllable parameters: the level
 1094 (number of rooms) and the object count per room. Objects
 1095 are drawn from a library of 293 distinct 3D models, grouped
 1096 into 6 categories and 37 subtypes, primarily everyday house-
 1097 hold items (see Figure 5). To ensure diversity, each object
 1098 type appears at most once in a given scene.

1099 For task setup, we additionally generate instructional (Fig-
 1100 ure 6) and orientation (Figure 7) images that serve as refer-
 1101 ences for the agent in vision-world. We include both images
 1102 in the vision prompt. Object placement follows validity con-
 1103 straints (e.g., collision avoidance, minimum spacing), and
 1104 random seeds control reproducibility across environments.

1105 B.2. Proxy agents

1106 **Information Gain Calculation** We use the AC-3 arc-
 1107 consistency algorithm to maintain, for each object, a do-
 1108 main of feasible grid cells. Initially, every object’s domain
 1109 spans the entire 20×20 map. Each new observation is
 1110 compiled into unary and binary constraints (e.g., egocen-
 1111 tric direction/distance bins, room visibility/occlusion, and
 1112 ALLDIFFERENT to prevent collisions). When a constraint
 1113 is added, AC-3 iteratively prunes any cell in one object’s
 1114 domain that is unsupported by the domains of related objects,
 1115 propagating revisions along incident arcs until a fixed point is
 1116 reached (all arcs are consistent). While AC-3 alone does not
 1117 guarantee global consistency, in our setting all constraints
 1118 are derived from a valid trajectory; therefore the ground-truth

Model Overview

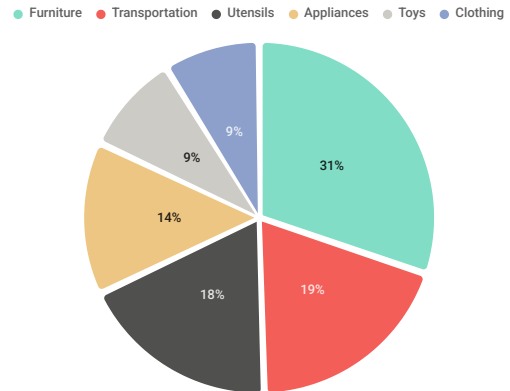


Figure 5. **Demonstration figures for FOV and 3D model distribution. Distribution of all 3D models used in our vision tasks.**

assignment remains supported and is never pruned, ensuring 1119
 that domains stay non-empty throughout propagation. 1120

Scripted Proxy Agents We implement two scripted proxies 1121
 to provide strong, reproducible baselines. 1122

SCOUT. From its spawn pose, the agent performs a 360° 1123
 sweep (four cardinal ROTATE+OBSERVE actions) to capture 1124
 all views at the initial location. It then follows a fixed room- 1125
 visitation order: upon discovering a doorway, it enters the 1126
 adjacent room, executes the same sequential sweep, and 1127
 repeats this “visit–sweep–advance” routine until every room 1128
 has been observed at least once. 1129

STRATEGIST. The first stage mirrors SCOUT: a 1130
 panoramic sweep to register all currently visible objects. 1131
 Thereafter, within the *current room* the agent maintains, for 1132
 each object, a set of feasible positions (“domain”) induced 1133
 by accumulated observations. At each turn it: (i) selects 1134
 the object with the largest remaining domain (highest po- 1135
 sitional uncertainty); (ii) moves to a viewpoint that best 1136
 constrains this object (e.g., near it or along a sightline that 1137
 intersects the most candidate cells); (iii) at that viewpoint, 1138
 orients to test pairwise relations: it computes unresolved 1139
 pairwise directions between the target object and all others 1140
 in the room, identifies the direction bin with the highest out- 1141
 standing count, and OBSERVES in that orientation first. The 1142
 procedure iterates until all objects in the room are resolved 1143
 (domains shrink to singletons), then proceeds to the next 1144
 unvisited room and repeats. 1145

1146 B.3. Prompts

We show the detailed designs of our prompts for exploration 1147
 in Figure 8, evaluation prompts in Figure 9, cognitive map 1148

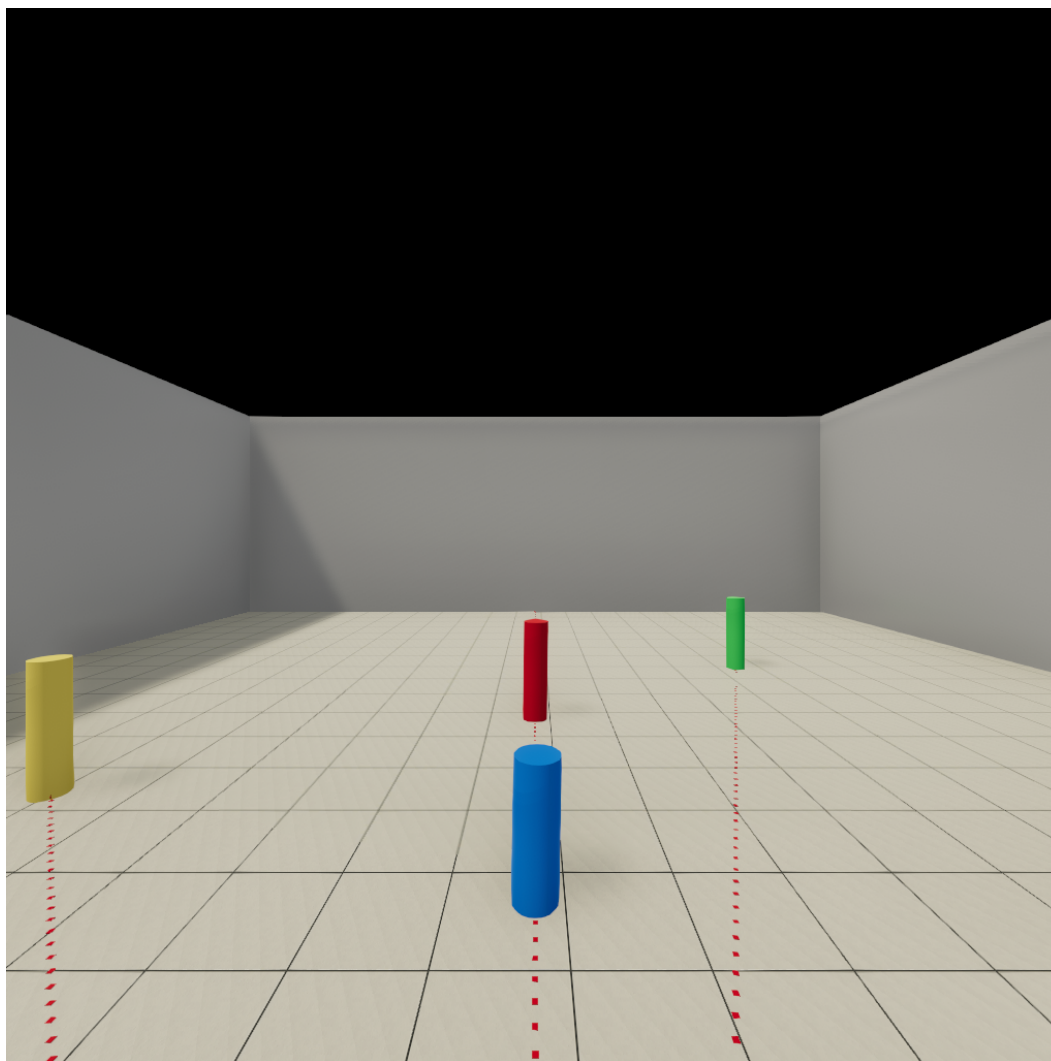


Figure 6. **Example of distance cues in the vision prompt.** The colored cylinders illustrate objects placed at different distances from the agent: yellow at 2 m, blue at 1 m, red at 2 m, and green at 3 m, providing calibration for mapping visual observations to discretized distance bins.

1149 prompts in Figure 10, and top-down view for uncertainty
1150 modeling in Figure 11.

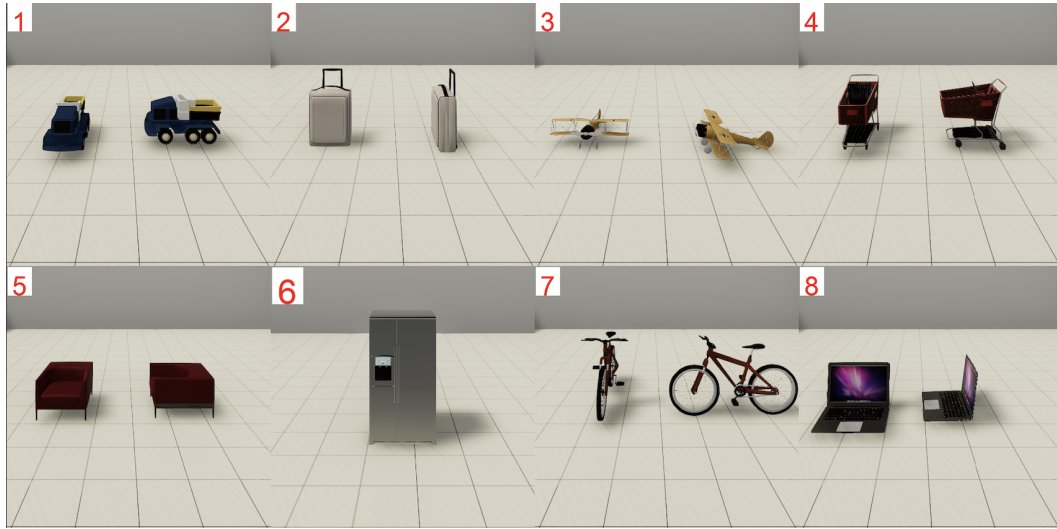


Figure 7. **Object appearance and orientation cues in the vision prompt.** Objects with facing direction are shown from both the front and side views, while objects without inherent orientation are displayed only from the front view. This provides the agent with consistent visual references for recognizing shape and facing.

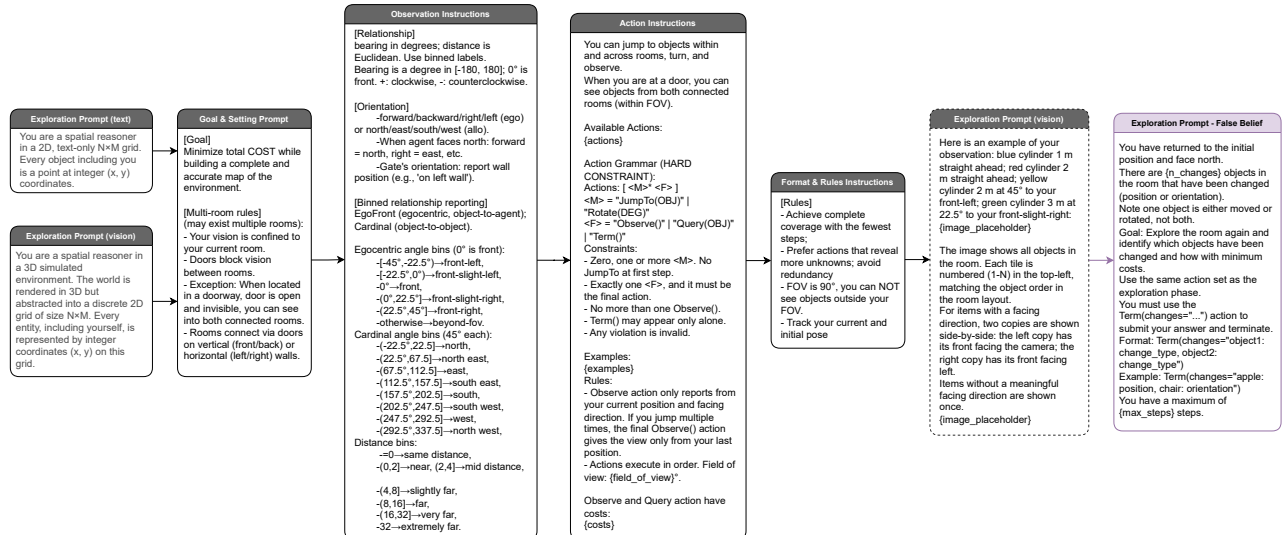


Figure 8. Exploration prompts

Route

Pairwise Direction

You return to your starting position and face north. From a Top-Down map, describe where {obj_name} is relative to {anchor_name}.

Answer format: <cardinal direction>, <distance>

Example: north-west, near

Perspective Taking

Now you jump to {anchor_name}'s direction, facing its direction. Describe where {obj_name} is relative to you.

Answer format: <ego direction>, <distance>

Example: front-left, near

Perspective Determine

Now you jump to an object's position, facing its direction. You observe that {observation}. Which object are you standing at?

Answer format: <object_name>

Example: lamp

Action2View

You return to your starting position and face north. You will execute the following action sequence: {actions}

After executing the actions, what is the ego relation of {target} relative to you?

Answer format: <ego direction>, <distance>

Example: front, near

View2Action

You return to your starting position and face north. Then you have executed an action sequence and changed to a new location and facing direction. You observe the following: {final_obs}

What action sequence led to this final view? The action sequence must be valid and only contain move actions.

Answer format: <sequence of move actions>

Example: JumpTo(lamp), Rotate(90)

Survey

Allocentric Map

Treat your starting position as the origin (0, 0) while facing north. Report allocentric coordinates using (x right/east, y up/north). Objects: {object_list}.

Answer format: (x0, y0); (x1, y1); ... in the same order.

Example: (1, 0); (-2, 3); (0, -1)

Mental Rotation

You return to your starting position and face north. You will perform a full 360-degree rotation by continuously turning {turn_direction} in place. Assume all walls are removed (you can see through walls), so every object is visible.

Focus on this set of objects: {object_pool}. List them in the exact order they appear directly ahead while you rotate. If two objects share a bearing, place the nearer one first.

Answer format: <object_name1>, <object_name2>, ...

Example: mug, sofa, plant

Location2View

{origin_instruction} You move to {loc} and face {direction}.

What is the egocentric relation of {target}?

Answer format: <direction>, <distance>

Example: front, near

View2Location

You move to a new location and face {orientation}. {observations} {origin_instruction} What is your new 2D coordinate (x, y)?

Answer format: (x, y)

Example: (2, -1)

Figure 9. Evaluation prompt design. We show the prompt for each evaluation task.

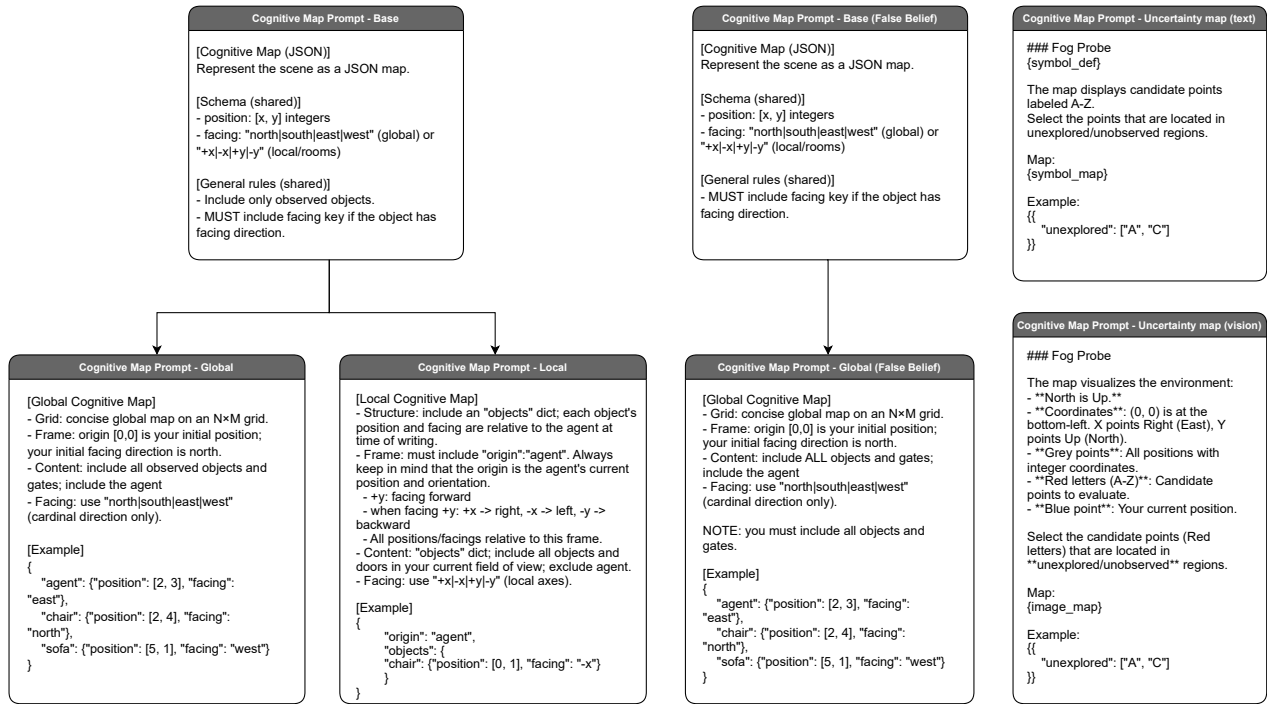


Figure 10. **Belief probing prompt design.** We use these prompts to ask the model to output a cognitive map or select unobserved points.

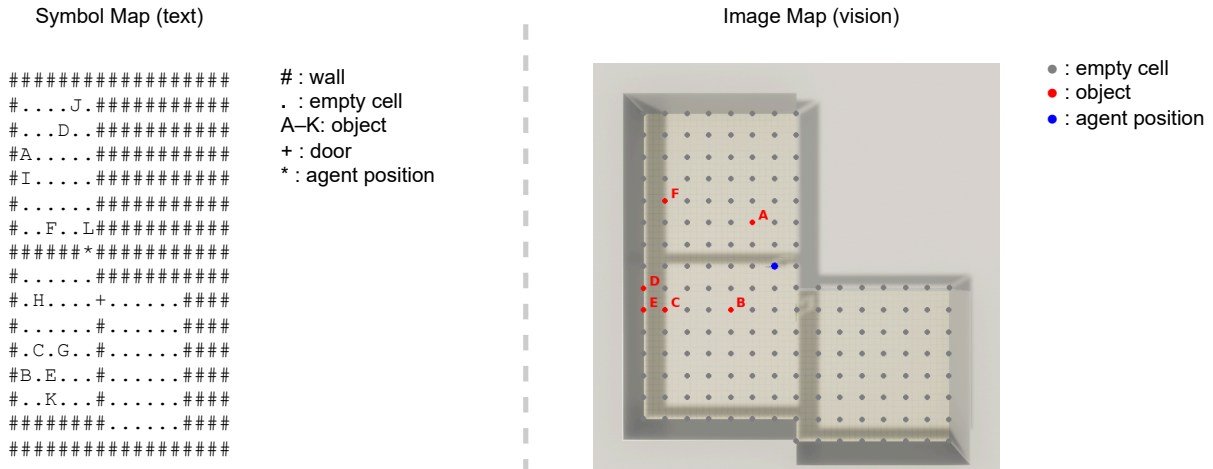


Figure 11. **The symbol map and the image map** provide parallel representations of the same environment for text and vision settings in uncertainty probing prompts.

1151	C. Extra Results		
	C.1. Additional Results for passive setting		1159
	The performance of various models in both text-based and vision-based environments is summarized in Table 6.	1160	1161
	C.2. Additional Results for multi-room settings		1162
	We show detailed results for different room settings including two-room and four-room layouts. In both the two-room and four-room settings, we use the same room size and the same number of objects per room as in the three-room setting. For the four-room setting, we connect the main room with all the others. We evaluate GPT-5.2 and GEMINI-3 PRO, the two best-performing models. Additionally, we tested higher resolution, but found no performance gain. Table 7 and 8 report passive and active performance of the two-room setting. Table 9 and 10 report passive and active performance of the three-room setting. As the number of rooms increases, exploration cost rises accordingly. The results also underscore the importance of efficient exploration: in the four-room setting, which demands more strategic exploration, the gap between active and passive performance becomes substantially larger.	1163	1164
		1165	1166
		1167	1168
		1169	1170
		1171	1172
		1173	1174
		1175	1176
		1177	1178
	C.3. Uncertainty Map Probing		1179
	To probe an agent’s ability to model uncertainty, we provide it with a top-down view of the scene in which all objects are removed, and we overlay a set of candidate points. These points are sampled randomly and include both previously observed and unobserved locations. The agent’s task is to identify which candidate points remain unobserved, thereby revealing its belief over unseen regions.	1180	1181
		1182	1183
		1184	1185
		1186	1187
	Representation. The agent receives an empty top down map that shows only the candidate points and its current position, with no objects present. The agent must select the points that have not yet been observed. In the text based world, the top down map is represented as an $N \times M$ symbolic grid, where different symbols denote the agent, gates, and candidate points. In the vision based world, all objects are removed and the agent instead receives a top down image of the environment, check examples in Figure 11. We use F_1 to evaluate selected points.	1188	1189
		1190	1191
		1192	1193
		1194	1195
		1196	1197
	We report Uncertainty scores in Table 3. GEMINI-3 PRO models uncertainty better than GPT-5.2 in both text- and vision-based settings. These results help explain the information gain and cognitive map trends in Figure 12. GPT-5.2 achieves higher initial information gain (i.e., it ramps up faster), likely because it quickly commits to an explore-the-doors strategy. However, it generalizes poorly to unobserved regions, reflected by the subsequent plateau in Figure 12: additional steps yield little marginal gain. In contrast, although GEMINI-3 PRO improves more slowly at the beginning, its cognitive map accuracy continues to increase with exploration, suggesting it keeps collecting useful evidence and progressively resolving uncertainty.	1198	1199
		1200	1201
		1202	1203
		1204	1205
		1206	1207
		1208	1209
1152	To enable a like-for-like comparison between the text and vision settings, we instantiate identical room layouts across modalities. Concretely, we generate 100 evaluation instances with IDs 0–99; for each ID, we use the ID itself as the random seed to drive task sampling in both environments.		
1153			
1154			
1155			
1156			
1157	This seed tying guarantees deterministic layouts and bit-for-bit reproducibility across modalities.		
1158			

MUSI 2026 Submission #9. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

Methods	direction	persp.take	perc.dec	act2view	view2act	alloc.map	ment.rot	loc2view	view2loc	Avg.
	Static (S)	Dynamic (D)				Static (S)	Dynamic (D)			
	Route					Survey				
Vision-based World										
<i>Proprietary Models</i>										
GPT-5.2	47.3	35.0	63.9	54.5	49.3	64.8	83.3	50.3	65.6	57.1
GEMINI-3 PRO	63.8	36.3	57.5	49.0	58.0	67.2	85.3	70.4	57.0	60.5
CLAUDE-4.5 SONNET	47.3	33.5	37.7	40.8	15.7	54.8	58.3	44.7	54.8	43.1
<i>Open-source Models</i>										
GLM-4.6V	11.5	24.5	4.7	19.0	2.7	22.9	11.7	20.0	33.6	16.7
QWEN3-VL	20.8	28.3	22.7	16.7	4.7	33.2	21.7	27.3	40.8	24.9
Text-based World										
<i>Proprietary Models</i>										
GPT-5.2	84.5	88.2	97.0	89.0	76.0	96.3	98.3	94.8	89.2	90.4
GEMINI-3 PRO	82.7	92.7	97.0	87.5	75.7	86.2	91.3	85.7	80.0	86.5
CLAUDE-4.5 SONNET	73.0	80.7	90.7	77.7	59.0	76.9	74.3	59.2	70.7	73.6
<i>Open-source Models</i>										
GLM-4.6V	22.3	39.8	25.0	25.3	4.7	21.2	9.0	27.0	35.7	23.4
INTERNVL-3.5	36.7	67.8	42.7	41.2	8.7	37.3	19.3	38.7	43.8	37.4
QWEN3-VL	40.8	69.3	56.5	50.0	17.7	42.8	40.3	42.5	54.6	45.6

Table 6. **Exploitation Performance (%) of Belief Construction via Passive Observations.** Models are evaluated as *passive comprehension agents* on Route- and Survey-level reasoning using standardized observation logs from scripted proxy explorers, decoupling exploration from belief construction across text- and vision-based environments. GEMINI-3 PRO leads most tasks in the vision-based world and achieves the best overall average, while GPT-5.2 leads the text-based world and attains the best overall average.

Methods	direction	persp.take	perc.dec	act2view	view2act	alloc.map	ment.rot	loc2view	view2loc	Avg.
	Static (S)	Dynamic (D)				Static (S)	Dynamic (D)			
	Route					Survey				
Vision-based World										
<i>Proprietary Models</i>										
GPT-5.2	39.2	37.3	63.3	53.8	58.3	68.2	92.7	52.3	68.6	59.3
GEMINI-3 PRO	57.8	33.9	53.8	48.5	58.7	64.6	83.3	54.7	69.8	58.3
Text-based World										
<i>Proprietary Models</i>										
GPT-5.2	85.3	92.0	99.0	90.0	83.0	97.2	99.7	89.5	95.2	92.3
GEMINI-3 PRO	88.2	86.7	91.7	87.3	79.3	90.1	92.7	81.5	82.9	86.7

Table 7. Exploitation Performance (%) via Passive Observations under **two rooms** settings.

Methods	Avg.cost	direction	persp.take	perc.dec	act2view	view2act	alloc.map	ment.rot	loc2view	view2loc	Avg.
		Static (S)	Dynamic (D)				Static (S)	Dynamic (D)			
		Route					Survey				
Vision-based World											
<i>Proprietary Models</i>											
GPT-5.2	10.8	41.3	36.2	48.2	49.0	54.7	56.9	72.0	45.2	59.7	51.5
GEMINI-3 PRO	6.6	51.7	36.3	63.0	47.2	56.0	63.4	85.0	50.3	67.5	57.8
Text-based World											
<i>Proprietary Models</i>											
GPT-5.2	6.2	68.7	67.3	90.0	76.8	64.0	83.4	92.7	73.7	83.7	77.8
GEMINI-3 PRO	6.2	76.0	68.3	89.0	77.2	72.7	83.1	96.0	77.5	86.2	80.6

Table 8. Exploitation Performance (%) via Active Exploration under **two rooms** settings.

Methods	direction	persp.take	perc.dec	act2view	view2act	alloc.map	ment.rot	loc2view	view2loc	Avg.
	Static (S)	Dynamic (D)				Static (S)	Dynamic (D)			
Vision-based World										
Proprietary Models										
GPT-5.2	47.0	37.7	59.7	38.3	40.3	60.1	73.7	50.5	65.9	52.6
GEMINI-3 PRO	63.5	35.5	58.7	42.8	43.0	64.4	81.7	48.8	67.4	56.2
Text-based World										
Proprietary Models										
GPT-5.2	83.8	88.2	94.3	86.8	62.7	94.8	93.7	82.0	92.5	86.5
GEMINI-3 PRO	81.2	91.3	96.7	82.2	68.3	76.8	81.3	74.2	79.0	81.2

Table 9. Exploitation Performance (%) via Passive Observations under **four rooms** settings.

Methods	Avg.cost	direction	persp.take	perc.dec.	act2view	view2act	alloc.map	ment.rot	loc2view	view2loc	Avg.
		Static (S)	Dynamic (D)				Static (S)	Dynamic (D)			
Vision-based World											
Proprietary Models											
GPT-5.2	23.2	41.2	33.2	49.0	30.8	30.7	32.5	49.7	40.5	55.4	40.3
GEMINI-3 PRO	19.7	59.8	34.2	60.3	34.7	46.0	56.8	62.7	44.0	64.8	51.5
Text-based World											
Proprietary Models											
GPT-5.2	16.4	65.3	69.0	74.3	62.8	44.3	66.6	76.3	57.5	77.8	66.0
GEMINI-3 PRO	19.7	76.3	77.2	91.7	73.3	64.3	77.0	83.7	74.0	81.9	77.7

Table 10. Exploitation Performance (%) via Active Exploration under **four rooms** settings.

Accumulated Info Gain & Cognitive Map Correctness

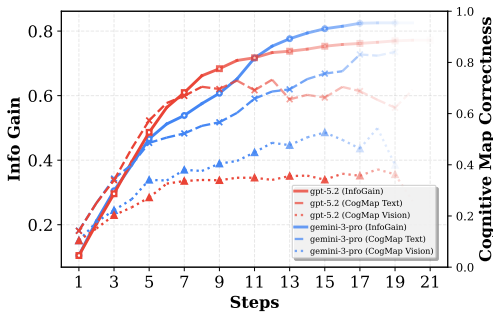


Figure 12. Accumulated Information Gain and Cognitive Map Correctness over steps.

1210 C.4. Cognitive map Validation & Correlation

1211 **Cognitive Map Validation & Correlation.** To validate the
1212 utility of the probed cognitive map and investigate whether
1213 it faithfully reflects the agent’s reasoning process, we first
1214 conducted two ablation studies:

- 1215 • **Sufficiency Test (Oracle Map):** We conditioned the
1216 model on the ground-truth cognitive map before gener-
1217 ating answers for evaluation. Performance rose to near-

perfect levels ($\approx 95\%$ for both models in both worlds).
This confirms that our cognitive map representation cap-
tures *all* necessary information for the tasks; performance
bottlenecks stem from the agent’s inability to accurately
construct the map, not the representation format itself.

- **Alignment Test (Explicit Reasoning):** We prompted the
model to explicitly generate the cognitive map before an-
swering the evaluation questions. This resulted in a slight
performance degradation compared to direct answering.

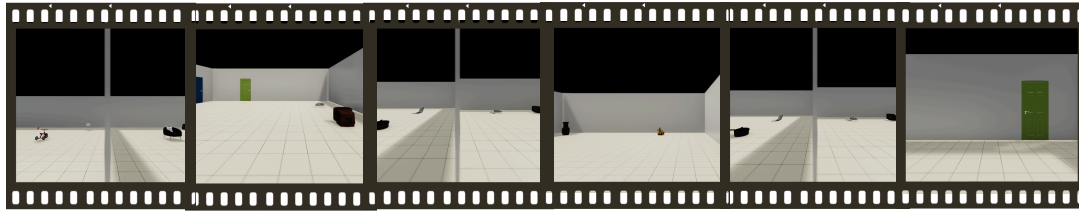
These results reveal an **externalization gap**: the model’s
latent internal spatial belief is richer or more accurate than
the discretized JSON output it produces. **While it is a lossy**

Methods	Text (%)	Vision (%)
GPT-5.2	41.8	57.0
GEMINI-3 PRO	46.6	64.5

Table 11. **Pearson correlation** (r) between belief correctness and downstream performance. All correlations are significant ($p < .001$).

compression of the agent’s true internal state, the explicit map remains a strong diagnostic signal. We support this claim by computing the Pearson correlation between the agent’s cognitive map **Correctness** and downstream task

1234	performance. To ensure a robust correlation, we calculate	D. Additional Visualization Examples	1249
1235	the average performance across five independent cognitive	We include concrete examples of task formats and answer	1250
1236	map runs for each sample. As shown in Table 11, belief	styles with open-ended, format-constrained outputs in Fig-	1251
1237	correctness is consistently and positively correlated with	ure 13.	1252
1238	downstream success in both modalities, with all correlations	D.1. Cognitive map output by models	1253
1239	significant ($p < .001$). The association is stronger in vi-	We visualize the turn-by-turn cognitive maps (in Figures 14	1254
1240	sion ($r=0.570/0.645$) than in text ($r=0.418/0.466$). The	and 15 of GPT-5.2, comparing them against ground-truth	1255
1241	stronger vision correlation suggests that perception-driven	maps. The performance is noticeably stronger in text-based	1256
1242	mapping errors and unstable belief updates more directly	environments than in vision-based ones.	1257
1243	translate into task failures. Thus, we establish map probing	D.2. Exploration pattern examples by models	1258
1244	as a <i>validated diagnostic proxy</i> for failure analysis. While	We include representative trajectories from each model to	1259
1245	acknowledging that correlation does not imply causality, we	illustrate the active exploration patterns identified in our	1260
1246	treat the explicit map as a robust, albeit conservative, signal	analysis, shown in Figure 16, 17, 18, 19, and 20 . These	1261
1247	for diagnosing reasoning breakdowns rather than definitive	examples highlight how different models manifest recurring	1262
1248	evidence.	exploration behaviors: for instance, GPT-5.2 often adopts	1263
		a “finding-gate” strategy, rotating until a doorway is de-	1264
		tected before moving toward it, while other models more	1265
		frequently repeat redundant checks. All figures mark the	1266
		agent’s position and orientation explicitly, with actions anno-	1267
		tated beneath each frame and a shared legend provided for	1268
		each trajectory.	1269
		D.3. Analysis Platform	1270
		We also include some demonstrations in Figure 21, 23, 22,	1271
		24, and 25 of our designed platform for better analysis	1272



Pairwise Direction

Q: You return to your starting position and face north.
From a Top-Down map, describe where shelf is relative to truck.

A: south east, mid distance

Allocentric Map

Q: Treat your starting position as the origin (0, 0) while facing north. Report allocentric coordinates using (x right/east, y up/north). Objects: shelf, truck, lamp.

A: [[12, -1], [10, 1], [0, 4]]

Perspective Taking

Q: Now you jump to backpack's direction, facing its direction.
Describe where chair is relative to you.

A: front-left, mid distance

Mental Rotation

Q: You return to your starting position and face north. You will perform a full 360-degree rotation by continuously turning counterclockwise in place. Assume all walls are removed (you can see through walls), so every object is visible. Focus on this set of objects: bike, pan, television. List them in the exact order they appear directly ahead while you rotate. If two objects share a bearing, place the nearer one first.

A: ['television', 'pan', 'bike']

Perspective Determine

Q: Now you jump to an object's position, facing its direction. You observe that truck is front-left, mid distance, facing backward; shelf is front, mid distance. Which object are you standing at?

A: laptop

Action2View

Q: You return to your starting position and face north.
You will execute the following action sequence:
1. Jump to the object at front-right, mid distance.
2. Rotate(-90)
3. Jump to the object at front-right, mid distance.
4. Rotate(-180)

After executing the actions, what is the ego relation of bike relative to you?

A: front-right, mid distance

Location2View

Q: Treat the green door as the new 'origin' (0, 0). You move to (2, -5) and face north. What is the egocentric relation of pan?

A: front, mid distance

View2Action

Q: You return to your starting position and face north. Then you have executed an action sequence and changed to a new location and facing direction. You observe the following: pan is at front-right, slightly far, facing backward; truck is at front-right, mid distance, facing forward; laptop is at front, mid distance, facing backward. What action sequence led to this final view?

A: [['rotate', 90], ['jumpto', 'green door'], ['jumpto', 'shelf'], ['rotate', 180]]

View2Location

Q: You move to a new location and face north. You observe: pan is at front, mid distance, facing right; truck is at front-right, mid distance, facing left; green door is at front-slight-left, slightly far, on left wall. Treat the green door as the new 'origin' (0, 0).

A: [2, -5]

Figure 13. **Examples of task formats and answer styles used.** Each block illustrates a spatial reasoning task type in our suite (Route-level and Survey-level), including the corresponding input context and an example open-ended answer that must follow a strict output format. In the vision setting, textual scene descriptions in the questions are replaced by rendered observation images.

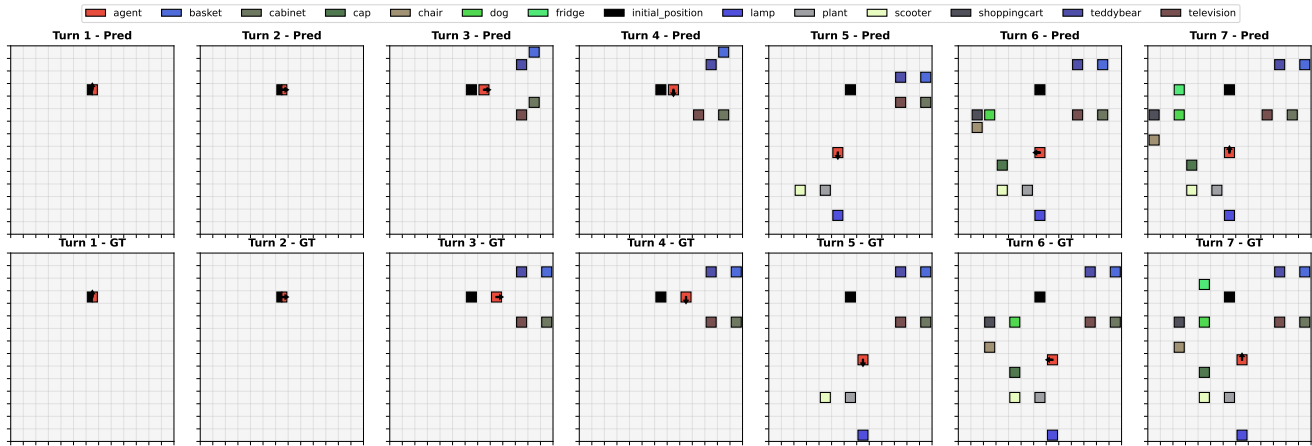


Figure 14. GPT-5.2's turn-by-turn cognitive map in text world during exploration.

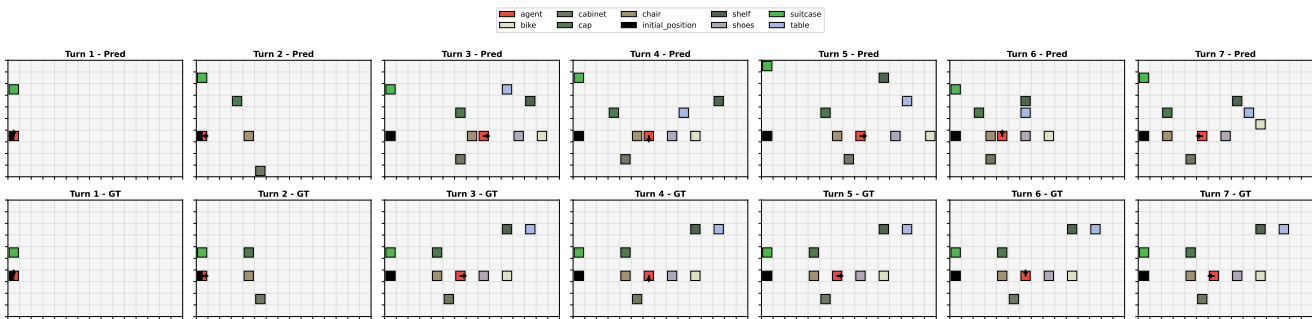


Figure 15. GPT-5.2's turn-by-turn cognitive map in vision world during exploration.

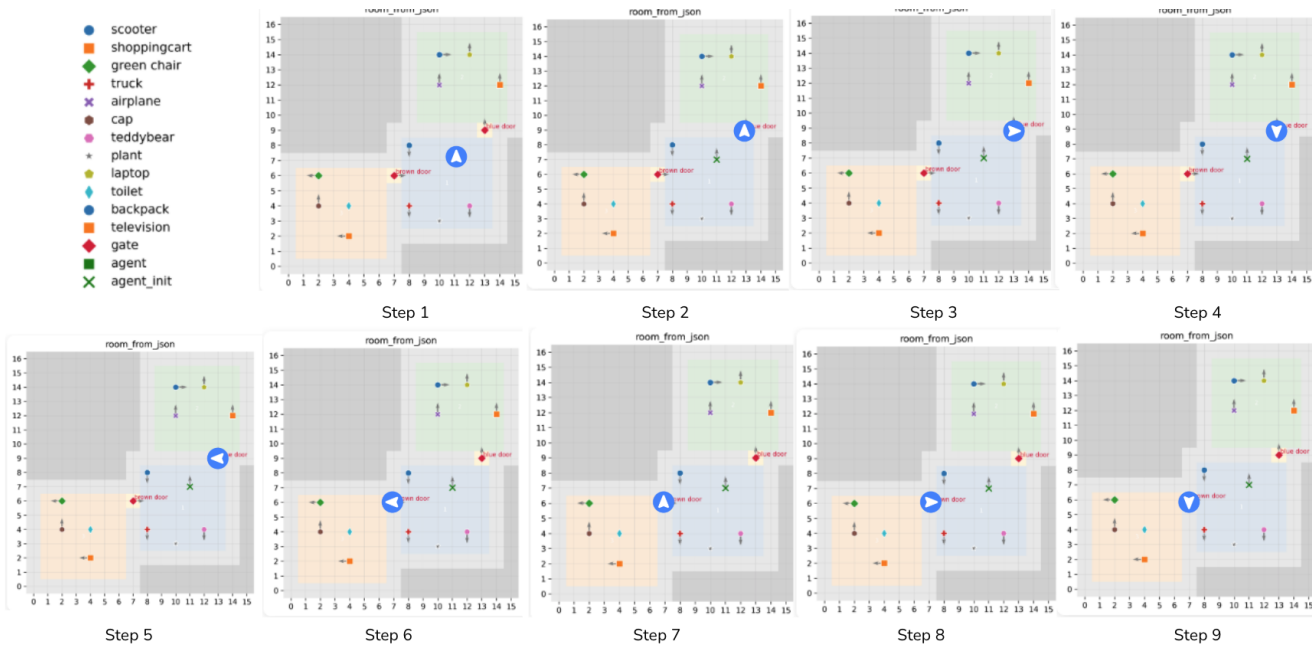


Figure 16. Example trajectory illustrating GPT-5.2's door-finding strategy and systematic sweeping pattern: Upon detecting a door, the agent navigates toward it and executes a strategic rotation to maximize environmental coverage. The process terminates once all target objects have been successfully identified.

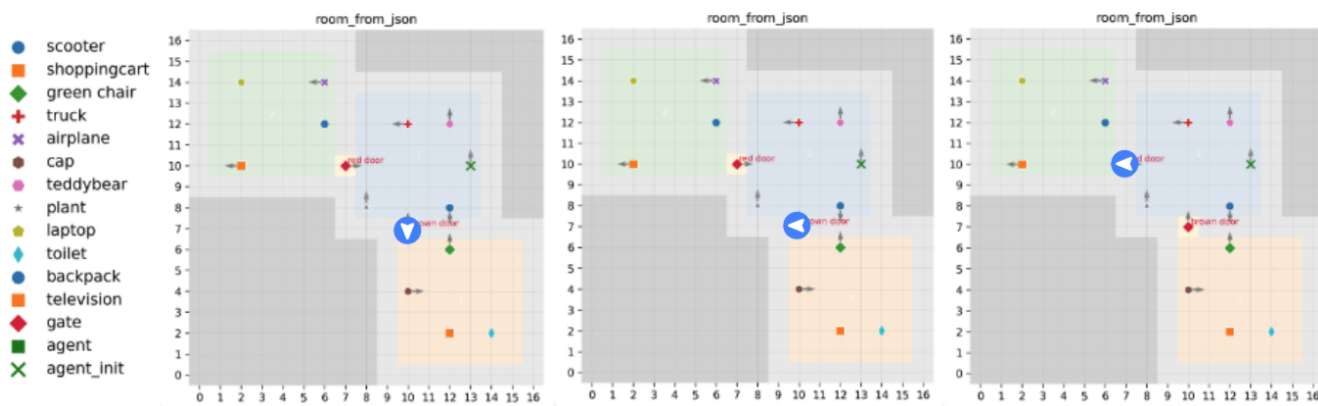


Figure 17. Example trajectory illustrating GPT-5.2’s omission pattern: Observing the door too early may lead the agent to skip the rest of the exploration, causing incomplete environmental discovery.

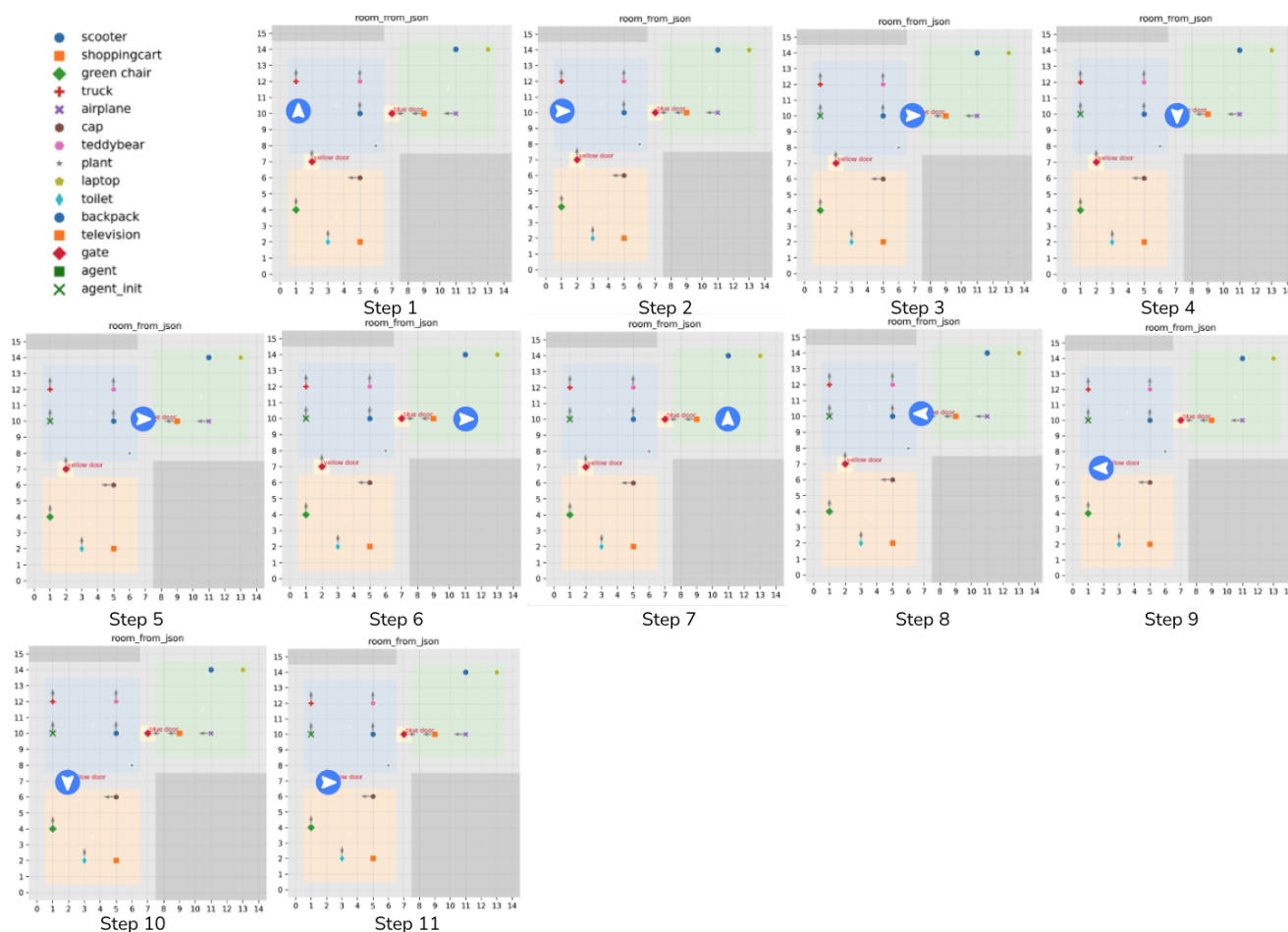


Figure 18. Example trajectory illustrating GEMINI-3 PRO’s door-finding strategy and systematic sweeping pattern in vision world: Upon detecting a door, the agent navigates toward it and executes a strategic rotation to maximize environmental coverage. The process terminates once all target objects have been successfully identified.

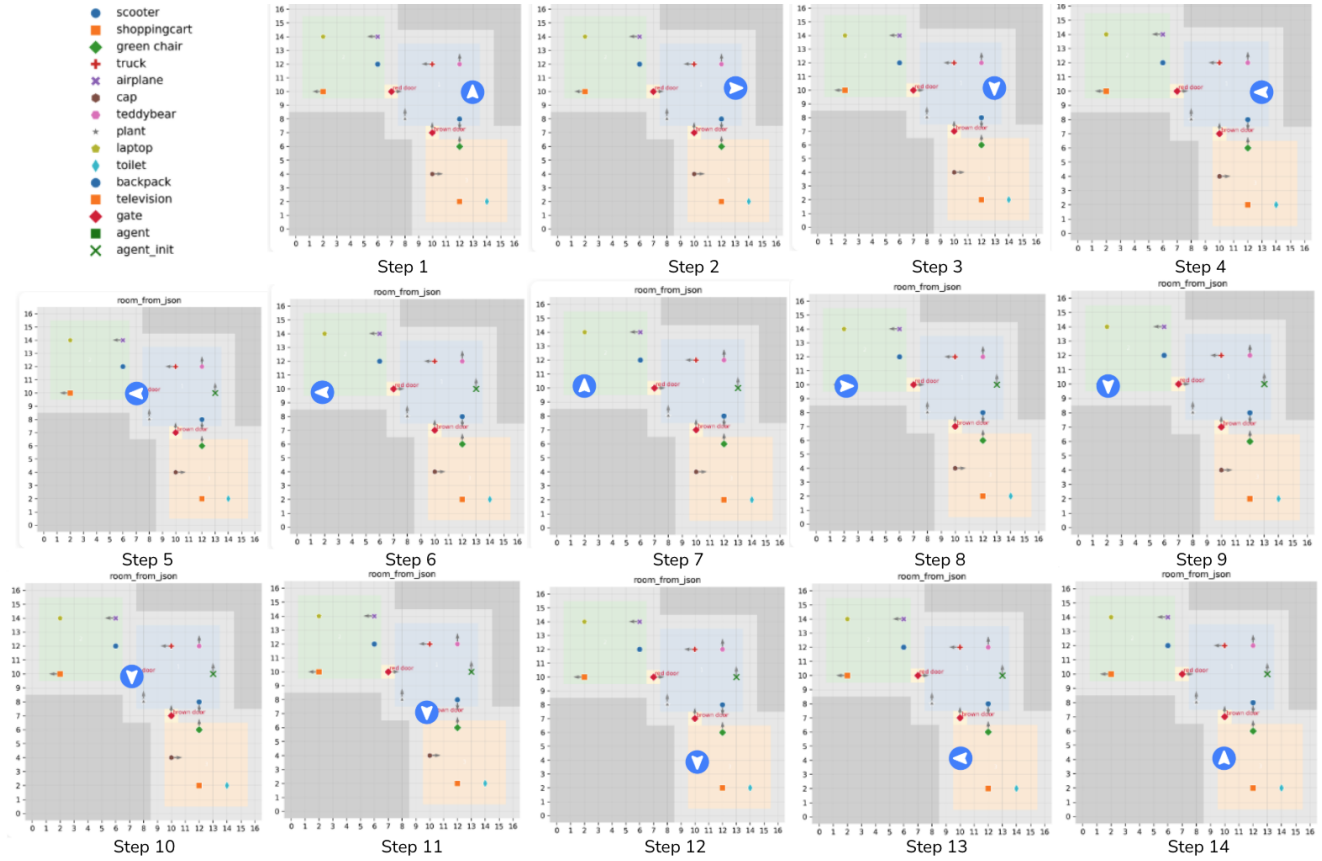


Figure 19. Example trajectory illustrating GEMINI-3 PRO’s object sweeping pattern mostly found in text world: Orbit the starting object using it as the pivot point. Randomly select an observed door to jump to a new object, then resume pivoting around the new target in a continuous loop.

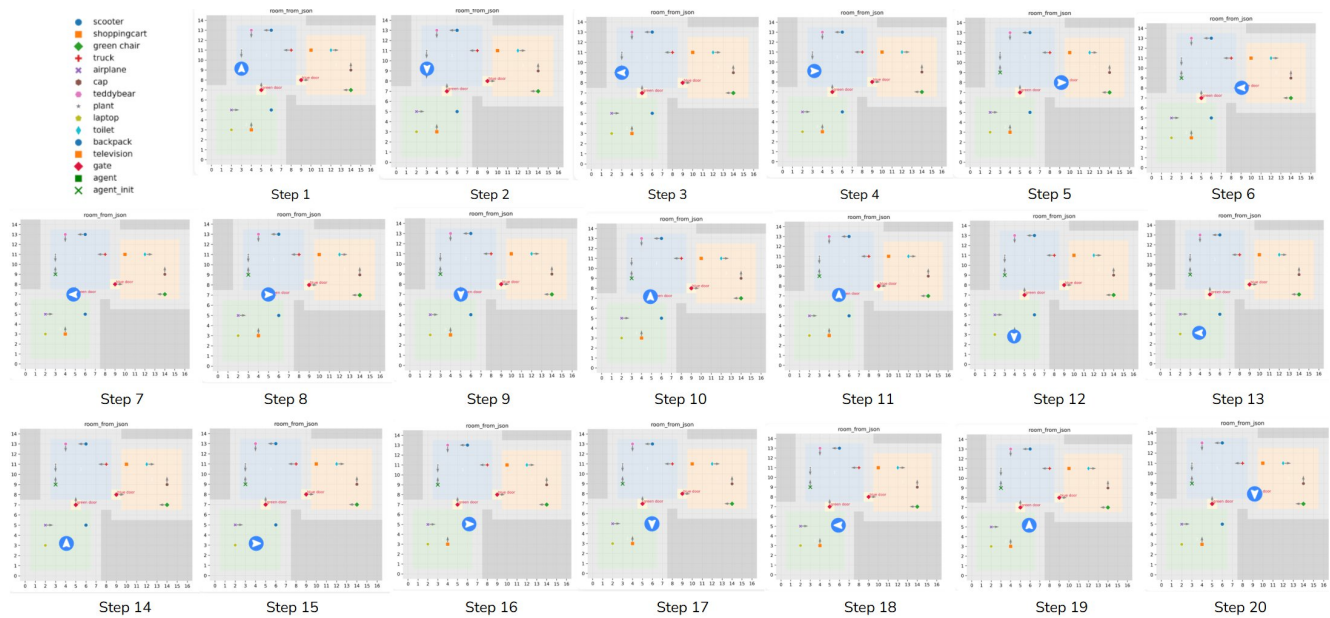


Figure 20. Example trajectory illustrating CLAUDE-4.5 SONNET’s exploration pattern: There is no clear exploration pattern.

Performance Charts

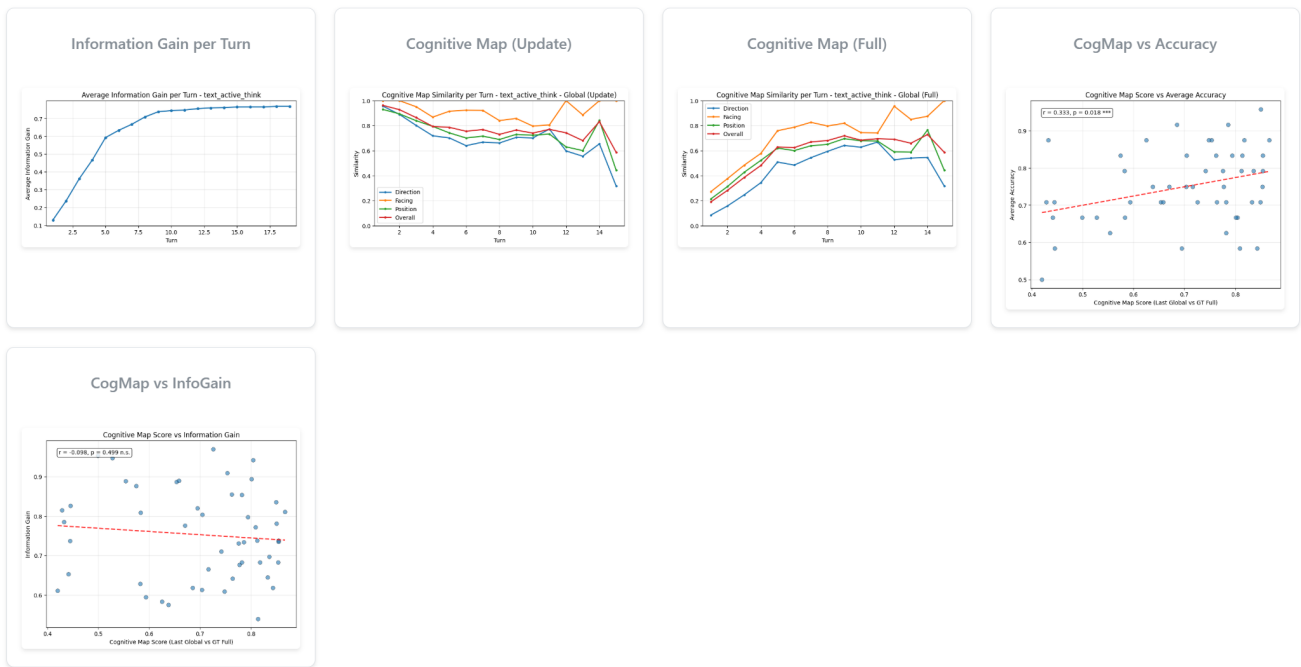


Figure 21. Platform designed by us for analysis (chart)

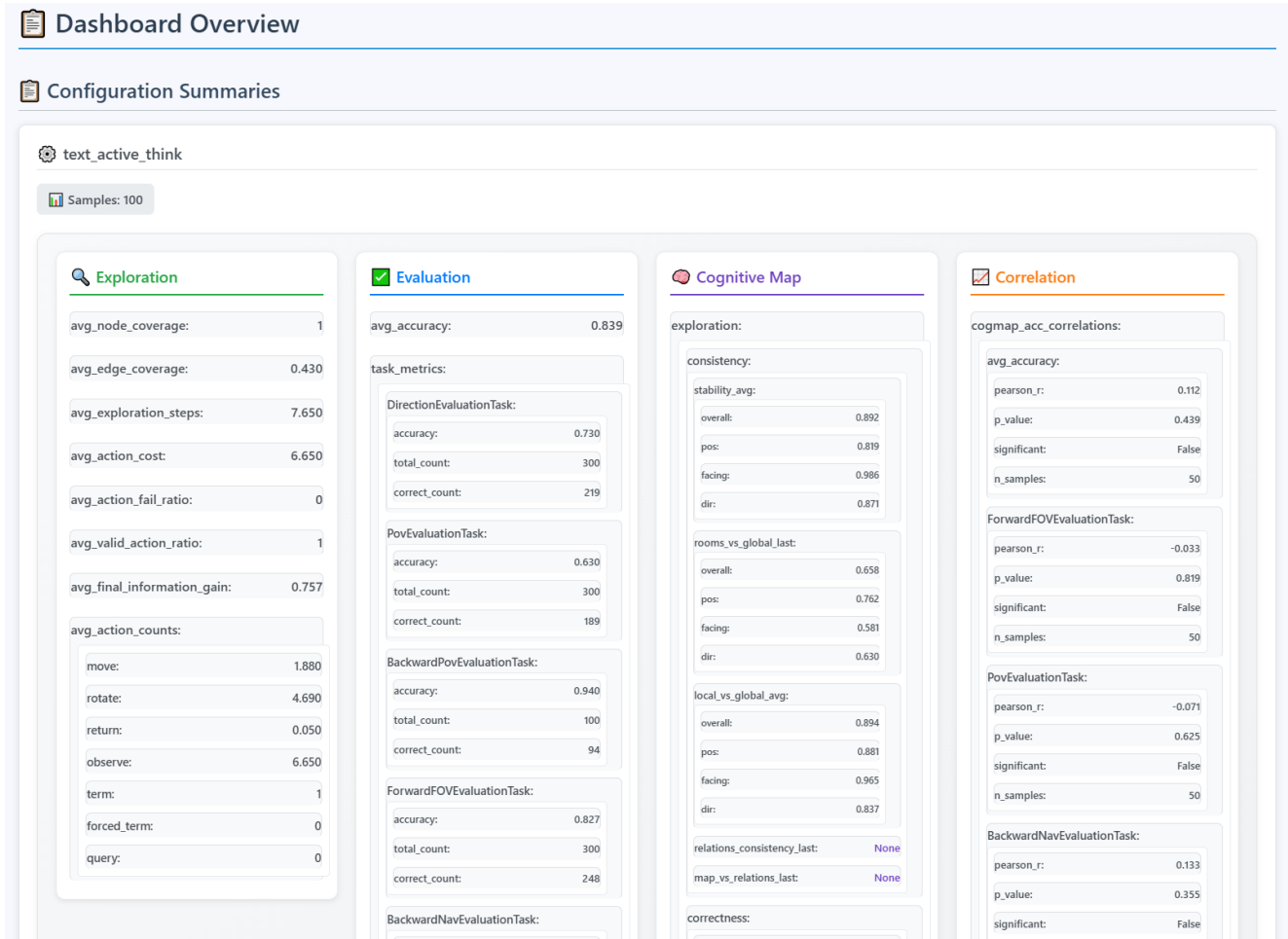


Figure 22. Visualization Platform for analysis: Metrics for active exploration in text world

vision_active_think

Samples: 100

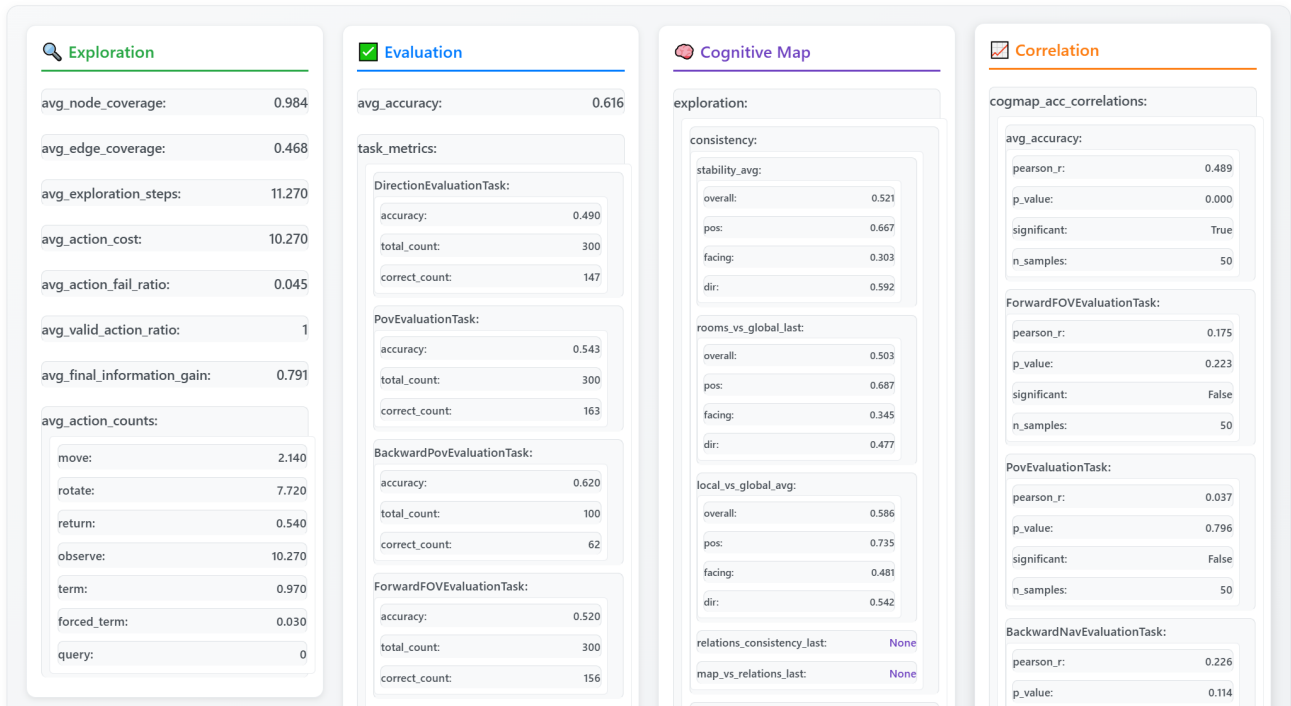


Figure 23. Visualization Platform for analysis: Metrics for active exploration in vision world

Turn 2

Environment Observation (click to toggle)

You observe:

- bike: front-right, mid distance, facing left
 - lamp: front, mid distance
 - blue door: front-right, slightly far, on front wall.
- You have a maximum of 19 exploration steps left.

Strictly follow this format:

THINK:

[Your thoughts on next step actions]

FINAL ANSWER:

Actions: [...]...

Assistant Thinking (click to toggle)

Jump to the blue door to stand in the doorway, which should open visibility into both connected rooms and reveal the most new objects/doors in one observe....

Assistant Action

Actions: [JumpTo(blue door), Observe()]

Global Cognitive Map Response (click to toggle)

THINK: Infer a consistent integer-coordinate layout from binned bearings/distances; keep agent at the blue door position and place objects relative to initial origin and subsequent doorway observation.

FINAL ANSWER:

```
{
  "agent": { "position": [3, 5], "facing": "north" },
  "lamp": { "position": [0, ...
```

Cognitive Map JSONs

Predicted

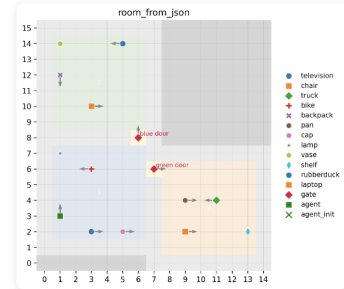
```
{
  "agent": {
    "position": [
      3,
      5
    ],
    "facing": "north"
  },
  "lamp": {
    "position": [
      0,
      3
    ]
  },
  "bike": {
    "position": [
      2,
      3
    ],
    "facing": "west"
  },
}
```

Ground Truth (Observed)

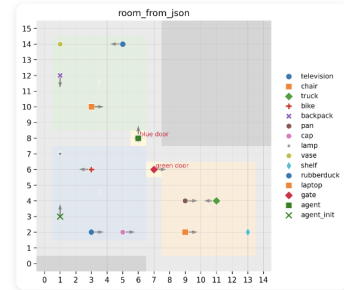
```
{
  "bike": {
    "position": [
      2,
      3
    ],
    "facing": "west"
  },
  "lamp": {
    "position": [
      0,
      4
    ],
    "facing": "east"
  },
  "vase": {
    "position": [
      0,
      11
    ],
    "facing": "east"
  },
}
```

Ground Truth (Full)

```
{
  "television": {
    "position": [
      2,
      -1
    ],
    "facing": "east"
  },
  "chair": {
    "position": [
      2,
      7
    ],
    "facing": "east"
  },
  "truck": {
    "position": [
      10,
      1
    ],
    "facing": "west"
  },
}
```



State before Turn 2



State at Turn 2

Figure 24. Visualization Platform for analysis: one turn of active exploration in text-world, including agent's action and cognitive map.

Turn 4

Environment Observation (click to toggle)
 You rotated counterclockwise 90°. You observe: <image>.
 You have a maximum of 17 exploration steps left.
 Strictly follow this format:
 THINK:
 [Your thoughts on next step actions]
 FINAL ANSWER:
 Actions: [...]...

Assistant Thinking (click to toggle)
 Still no objects/doors visible; rotate to face east to re-scan the remaining quadrant of room 1 (and potentially spot the green door or other objects)....

Assistant Action
 Actions: [Rotate(-90), Observe()]

Global Cognitive Map Response (click to toggle)
 THINK: Need a minimal global cognitive map including only entities observed so far (agent, 2 doors, bike, lamp, television, cap).
 Use initial agent position as origin; agent currently facing east after three -90° rotations. Approximate integer grid positions consistent with observations (blue door o...

Cognitive Map JSONs

Predicted

```
{
  "agent": {
    "position": [
      0,
      0
    ],
    "facing": "east"
  },
  "blue door": {
    "position": [
      3,
      6
    ]
  },
  "green door": {
    "position": [
      6,
      3
    ]
  },
  "lamp": {
    "position": [
      -1,
      5
    ]
  },
  "bike": {
```

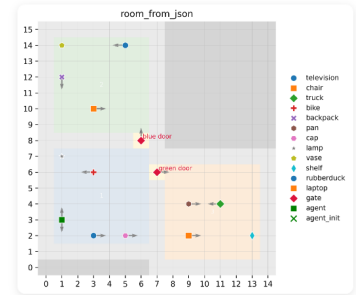
Ground Truth (Observed)

```
{
  "television": {
    "position": [
      2,
      -1
    ],
    "facing": "east"
  },
  "bike": {
    "position": [
      2,
      3
    ],
    "facing": "west"
  },
  "cap": {
    "position": [
      4,
      -1
    ],
    "facing": "east"
  },
  "lamp": {
    "position": [
      0,
      4
    ],
```

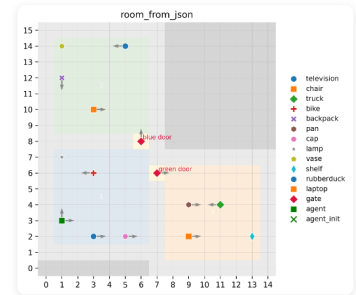
Ground Truth (Full)

```
{
  "television": {
    "position": [
      2,
      -1
    ],
    "facing": "east"
  },
  "chair": {
    "position": [
      2,
      7
    ],
    "facing": "east"
  },
  "truck": {
    "position": [
      10,
      1
    ],
    "facing": "west"
  },
  "bike": {
    "position": [
      2,
      3
    ],
```

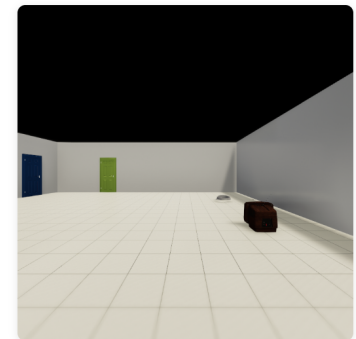
Local Cognitive Map Response (click to toggle)
 THINK: Build a local (agent-centric) map from the current view (agent facing +y in local frame), including only currently visible doors/objects; add facing only for the television.



State before Turn 4



State at Turn 4



Observation 1

Figure 25. Visualization Platform for analysis: one turn of active exploration in vision-world