DOES SPATIAL COGNITION EMERGE IN FRONTIER MODELS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Not yet. We present SPACE, a benchmark that systematically evaluates spatial cognition in frontier models. Our benchmark builds on decades of research in cognitive science. It evaluates large-scale mapping abilities that are brought to bear when an organism traverses physical environments, smaller-scale reasoning about object shapes and layouts, and cognitive infrastructure such as spatial attention and memory. For many tasks, we instantiate parallel presentations via text and images, allowing us to benchmark both large language models and large multimodal models. Results suggest that contemporary frontier models fall short of the spatial intelligence of animals, performing near chance level on a number of classic tests of animal cognition.

1 INTRODUCTION

Frontier models have achieved impressive performance in mathematics, coding, general knowledge, and commonsense reasoning (Hendrycks et al., 2021a;b; Chen et al., 2021; Sakaguchi et al., 2021; Yue et al., 2024). This remarkable progress has inspired characterizations of frontier models as possessing the intelligence of a smart high schooler and predictions of the imminent arrival of super-intelligence (Aschenbrenner, 2024). These characterizations are often underpinned by the premise that competence (or even mastery) in some aspects of cognition is symptomatic of broad cognitive competence. This is not self-evident. To quote Brooks's first law of artificial intelligence, "When an AI system performs a task, human observers immediately estimate its general competence in areas that seem related. Usually that estimate is wildly overinflated." (Brooks, 2024).

Our work focuses on spatial cognition, a foundational form of intelligence that is present in a broad spectrum of animals including humans (Marshall & Fink, 2001; Waller & Nadel, 2013; Mallot, 2024). Spatial cognition refers to the ability of animals to perceive and interact with their surroundings, build mental representations of objects and environments, and draw upon these representations to support navigation and manipulation. Decades of research in animal cognition have characterized the spatial cognition of rats, bats, dogs, chimpanzees, wolves, and humans (Tolman, 1948; Menzel, 1973; Peters, 1974; Gillner & Mallot, 1998; Marshall & Fink, 2001; Tommasi et al., 2012; Geva-Sagiv et al., 2015). Human infants already possess rudimentary spatial cognition, which subsequently improves along developmental schedules that have been characterized (Blades & Spencer, 1994; Newcombe, 2000; Vasilyeva & Lourenco, 2012). Spatial cognition is known to underpin more advanced cognitive abilities (Kozhevnikov et al., 2007; Newcombe, 2010; Young et al., 2018).

The emergence of spatial cognition has been linked to embodiment (Smith & Gasser, 2005; Jansen & Heil, 2010; Frick & Möhring, 2016), without which the development of spatial cognition may be impaired (Foreman et al., 1990; Anderson et al., 2013). However, frontier models are typically trained in a disembodied manner on corpora of text, images, and video. Does spatial cognition emerge in disembodied frontier models? To study this question systematically, we develop SPACE, a benchmark that builds on decades of research in cognitive science. Our benchmark comprises two broad classes of tasks, covering large-scale and small-scale spatial cognition (Hegarty et al., 2006; Meneghetti et al., 2022; Newcombe, 2024). The tasks are schematically illustrated in Figure 1.

Large-scale spatial cognition has to do with a model's ability to understand its surroundings. In large-scale spatial cognition tasks, the model is familiarized with an environment and is then asked to estimate distances and directions to landmarks, sketch a map of the environment, retrace a known route, or identify a shortcut to the goal. Small-scale spatial cognition has to do with a model's ability

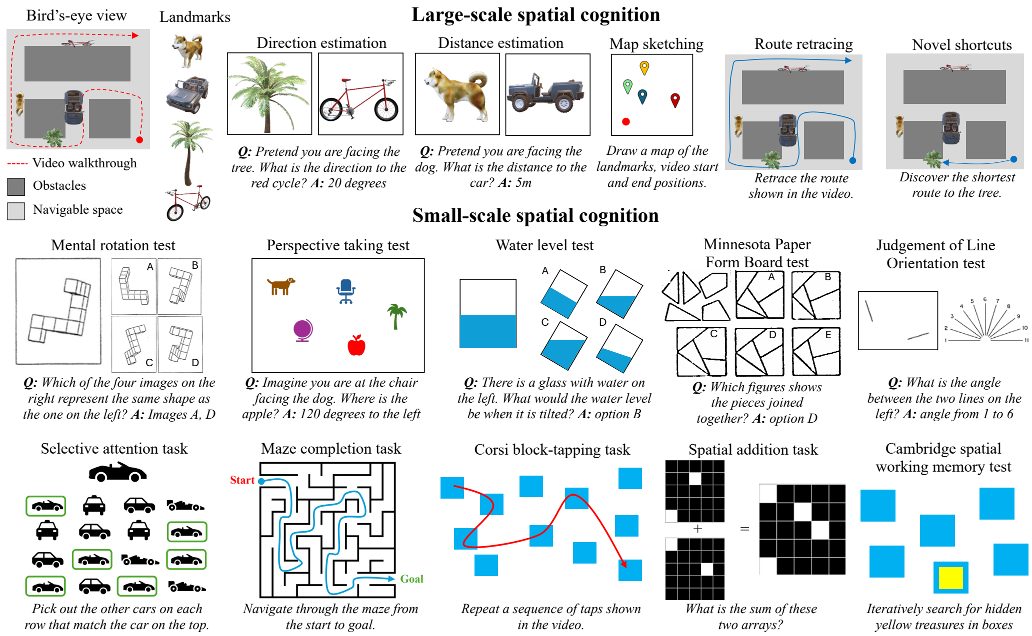

Figure 1: **SPACE: Spatial Perception And Cognition Evaluation.** We design a suite of spatial cognition tasks based on the cognitive science literature. These are broadly classified into large-scale and small-scale spatial cognition. Large-scale tasks require understanding space at the level of environments, while small-scale tasks require understanding space at the level of objects or object arrangements. We develop multimodal as well as purely textual presentations, which support evaluation of both large language models (LLMs) and vision-language models (VLMs).

to perceive, imagine, and mentally transform objects in two or three dimensions. Together, large-scale and small-scale tasks evaluate core cognitive abilities such as spatial perception, visualization, selective attention, and visuospatial memory.

We design text-based and image-based presentations to evaluate both language-only and vision-language models (LLMs and VLMs, respectively). Our results indicate that contemporary frontier models have not yet reached competency – let alone mastery – in spatial cognition. On key large-scale spatial cognition tasks, frontier multimodal models perform near chance level, even when presented with an allocentric (map) view of the environment. The strongest models exhibit much better performance on some small-scale tasks, especially with purely textual presentations via character arrays, but perform near chance on image-based tasks such as mental rotation (Vandenberg & Kuse, 1978), perspective taking (Kozhevnikov & Hegarty, 2001), maze completion (Lacroix et al., 2021), or the classic Minnesota Paper Form Board Test (Likert & Quasha, 1941; 1969).

2 Related work

Spatial cognition. Spatial cognition is a branch of cognitive science that seeks to understand how humans and animals perceive, interpret, represent, and interact with objects and environments (Marshall & Fink, 2001; Landau, 2002; Waller & Nadel, 2013; Mallot, 2024; Newcombe, 2024). This involves the perception of object sizes, shapes, and scales, as well as the relationships between objects and landmarks in the environment (including location, distance, direction, and orientation). Spatial cognition is broadly divided into two categories: large-scale and small-scale (Hegarty et al., 2006; Jansen, 2009; Meneghetti et al., 2022; Newcombe, 2024). *Large-scale spatial cognition* refers to the ability to build spatial representations of environments and use them effectively for navigation and spatial reasoning. Large-scale spatial cognition tasks typically involve egocentric spatial transformations, where the viewer's perspective changes with respect to the environment while the spatial relationships between parts of the environment remain constant (Wang et al., 2014). *Small-scale spatial cognition* refers to the ability to perceive, imagine, and mentally transform objects or shapes in 2D or 3D. This is typically evaluated using paper and pencil tasks that require allocentric spatial transformations of objects and shapes (Wang et al., 2014). While large-scale spatial cognition has been demonstrated in a wide range of animals (Tolman, 1948; Menzel, 1973; Peters, 1974; O'Keefe & Nadel, 1978; Gillner & Mallot, 1998; Richardson et al., 1999; Geva-Sagiv et al., 2015; Toledo et al., 2020), the study of small-scale spatial cognition is specific to humans.

Spatial reasoning in large language models. PlanBench (Valmeekam et al., 2024) and CogEval (Momennejad et al., 2023) evaluate LLMs on text-based planning tasks such as navigation, delivery logistics and block stacking to evaluate cognitive mapping and planning. Yamada et al. (2024) evaluates spatial reasoning in LLMs by performing map traversals of different types of graphs and evaluate the model's self-localization ability. EWOK (Ivanova et al., 2024) studies spatial plausibility reasoning in LLMs (i.e., given some context text, does a target text sound plausible?). Unlike these benchmarks, SPACE evaluates a broader umbrella of cognitive abilities and implements multimodal presentations of classic animal cognition experiments.

Benchmarks for large multimodal models. The recent successes of multimodal models (OpenAI, 2024; Li et al., 2024a; Reid et al., 2024) have been facilitated by large-scale training on text and multimodal corpora (Rana, 2010; Together Computer, 2023; Chen et al., 2023; Laurençon et al., 2023; Gadre et al., 2023), followed by tuning on human preferences (Liu et al., 2023a; Awadalla et al., 2024; Ouyang et al., 2022; Rafailov et al., 2023). The remarkable advances in the capabilities of these models inspired a variety of benchmarks that evaluate their performance. Early multimodal benchmarks consisted of single-task datasets such as visual question answering (Antol et al., 2015; Goyal et al., 2019; Marino et al., 2019) and image captioning (Chen et al., 2015). However, due to the limited scope of early datasets and concerns regarding potential test-data leakage, newer benchmarks use diverse collections of tasks (Fu et al., 2023; Yu et al., 2024; Liu et al., 2023b; Yue et al., 2024; Lu et al., 2024; Ying et al., 2024). While these datasets primarily focus on image understanding, newer datasets that emphasize spatiotemporal reasoning have been proposed for video (Li et al., 2024b; Fu et al., 2024a; Majumdar et al., 2024).

Recent studies highlight a number of shortcomings of frontier multimodal models (Moskvichev et al., 2023; Tong et al., 2024; Chen et al., 2024a; Fu et al., 2024b). One such shortcoming is that models may not perceive the image in detail, often missing fine-grained details or ignoring the image entirely (Chen et al., 2024b; Guan et al., 2024; Tong et al., 2024). HallusionBench proposes a new dataset of image pairs, where tiny edits are made from one image to another that change the answer to the question (Guan et al., 2024). MMVP identifies issues with CLIP-based pretraining of visual encoders, which make current models blind to certain visual patterns, and proposes a benchmark of CLIP-blind image pairs where the same question has opposite answers (Tong et al., 2024). MMStar shows that many questions in multimodal benchmarks can be answered correctly without the image and proposes a new split of existing benchmarks that addresses this issue (Chen et al., 2024b).

Another shortcoming of existing models is their lack of spatial perception and reasoning (Chen et al., 2024a; Cheng et al., 2024). SpatialVLM proposes a VQA dataset that requires answering questions about relative spatial arrangements and metric relationships (Chen et al., 2024a). SpatialRGPT further includes region-level understanding (Cheng et al., 2024). 'Perception test' aims to overcome shortcomings of standard video datasets by creating a diagnostic dataset where participants record videos while following complex scripts depicting interesting events (Patraucean et al., 2023). It evaluates fundamental perceptual skills (memory, abstraction, intuitive physics, and semantics) and various types of reasoning.

Another line of work considers skill acquisition (the ability to learn a skill and apply it to new scenarios). Prior work has studied this using visual analogical reasoning (Chollet, 2019; Moskvichev et al., 2023; Yiu et al., 2024). The ARC dataset contains samples consisting of a few examples of abstract grids and their transformations and one or more test inputs (Chollet, 2019). The objective is to understand the transformation performed using the examples and apply it to test inputs. The transformations have been further organized into specific concepts with varying degrees of difficulty in the ConceptARC dataset (Moskvichev et al., 2023). Inspired by ARC and developmental psychology, the KiVA dataset studies visual analogies in the context of visually realistic 3D shapes with concepts like transformations in color, size, rotations, reflections, and counting (Yiu et al., 2024).

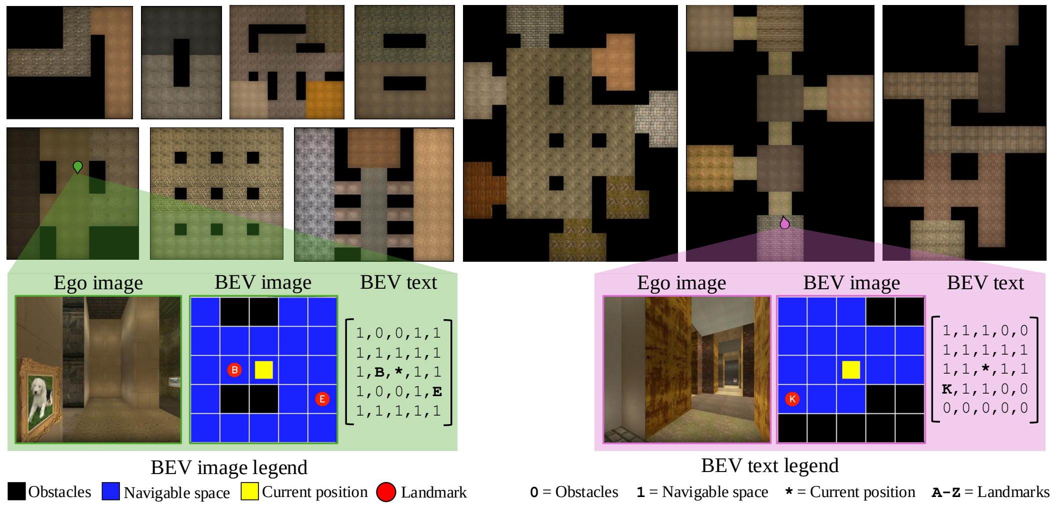

Figure 2: **Large-scale spatial cognition.** We design ten environment layouts based on experimental protocols in cognitive science. The top row shows bird's-eye view (BEV) renderings of these environments. To evaluate large-scale spatial cognition in frontier models, we implement three observation spaces: egocentric image, BEV image, and BEV text (see bottom row). **Ego image** shows a first-person view within the environment. **BEV image** shows an allocentric bird's-eye view of the $2.5\text{m} \times 2.5\text{m}$ region centered on the current position. **BEV text** shows a depiction of the allocentric bird's-eye view using text characters. We use the ego image and BEV image presentations to evaluate multimodal models. Large language models are evaluated using the BEV text presentation.

3 SPACE: A BENCHMARK FOR SPATIAL PERCEPTION AND COGNITION EVALUATION

We develop a benchmark for evaluating the spatial cognition of frontier models. The benchmark comprises large-scale and small-scale tasks and is designed for compatibility with both text-only and multimodal models.

3.1 LARGE-SCALE SPATIAL COGNITION

In large-scale spatial cognition tasks, we evaluate the ability of models to build spatial representations of their surrounding environment, and whether they can use these representations to reason about and navigate in the environment. There are two stages to these tasks. First, we familiarize the model with an environment by showing a video walkthrough.[1] The model must build a mental representation of the environment that captures the locations of start, goal and landmark locations, and their spatial relationships. After the model is familiarized with the environment, we evaluate the model's spatial representation using five tasks derived from the cognitive science literature (Meneghetti et al., 2022). See Figure 1(top) and Figure 2 for an overview.

1. **Direction estimation.** The goal is to determine the directions to other landmarks from a given landmark. The participant is asked to pretend that they are facing a landmark A, and then asked to estimate the direction (in degrees) to another landmark B. These are also known as pointing trials in the cognitive science literature (Allen et al., 1996; Hegarty et al., 2006; Pazzaglia & Taylor, 2007; Weisberg et al., 2014; Meneghetti et al., 2016). We formulate this as a multiple-choice QA task with four options for the direction (only one correct option).
2. **Distance estimation.** The goal is to determine the straight-line distances from one landmark to all other landmarks (Allen et al., 1996; Hegarty et al., 2006). The participant is asked to pretend that they are facing a landmark A, and then asked to estimate the Euclidean distance to all the other landmarks. We pose this as a multiple-choice QA with four options for the list of distances to each landmark. Since current models are not good at estimating metric measurements (Chen

[1]For text-only models, the 'video walkthrough' is a sequence of bird's-eye view (BEV) observations presented as arrays of letters, see Figure 2 for examples.

et al., 2024a; Cheng et al., 2024), we generate incorrect options such that the ratios of distances between landmarks are not preserved, making it easier to identify the correct option.

3. **Map sketching.** The goal is to draw a map of the environment that contains the start, goal and landmark positions (Allen et al., 1996; Hegarty et al., 2006; Pazzaglia & Taylor, 2007; Weisberg et al., 2014; Meneghetti et al., 2016; 2021). We again formulate this as multiple-choice QA with four options for the map sketches. The correct option preserves the true spatial relationships between the different map elements, while the incorrect options skew the spatial relationships randomly.

4. **Route retracing.** The goal is to retrace the route shown in the video from the start to the goal (Allen et al., 1996; Pazzaglia & Taylor, 2007; Meneghetti et al., 2016; 2021). This task evaluates the model's ability to remember landmarks seen in the route and the actions required along the route to reach the goal. We formulate this as an interactive task where the model receives the current observation, decides which action to take, and receives updated observations based on the actions taken. We measure performance using the SPL metric (success weighted by path length), which penalizes the model for taking unnecessary detours (Anderson et al., 2018). (The route shown in the demonstration, which the model is asked to retrace, is always the shortest path from the start to the goal.)

5. **Novel shortcut discovery.** The goal is to discover a novel shortcut (i.e., a route never observed before) from the start to the goal after observing a video walkthrough that takes detours to reach the goal (Tolman, 1948; Allen et al., 1996; Pazzaglia & Taylor, 2007; Meneghetti et al., 2016; 2021). The ability to take novel shortcuts in familiar environments is a key indicator of cognitive mapping ability (Tolman, 1948). When designing environments and walkthrough paths, we ensured that a novel shortcut exists that the model can exploit. Similar to route retracing, we treat this as an interactive navigation task and measure performance using the SPL metric.

3.1.1 Implementation

3D environment generation. We create ten environment layouts based on prior work in cognitive science and artificial intelligence (Tolman, 1948; Gillner & Mallot, 1998; Richardson et al., 1999; Banino et al., 2018; Bouchekioua et al., 2021). Figure 2 shows bird's-eye view (BEV) images of each layout. We populate each environment with visual landmarks in the form of paintings hanging on the walls, where the painting frames are 3D meshes and the paintings are images from ImageNet (Deng et al., 2009). To create a 3D environment for a given layout, we first randomly sample textures for walls, floors, and ceilings from a database of textures to create the base 3D mesh. Next, we randomly assign ImageNet images and 3D frame meshes to predefined landmark locations in the environment. We create the 3D environment using the Trimesh library and export it in glTF format (Dawson-Haggerty et al., 2019). We simulate the environment using the Habitat simulator (Savva et al., 2019). We create 3 environments per layout, for a total of 30 environments in our benchmark.

Observation spaces. We create multiple observation spaces to support evaluating both text-only and vision+text models. These are egocentric images, bird's-eye view (BEV) images, and bird's-eye view (BEV) text presentations.

- **Ego image.** The environment is captured using a forward-facing perspective camera placed at the model's location in the environment. This is similar to the setup of an animal navigating through an immersive environment.
- **BEV image.** This is a bird's-eye view image of a $2.5\text{m}\times 2.5\text{m}$ area in the environment surrounding the model's location. This is akin to a human using a map to navigate. The current location is always at the center of the BEV image. We use a Pacman-like coloring scheme highlighting the obstacles, navigable space, current postion, and landmarks.
- **BEV text.** This is a presentation of the BEV image in the form of an array of text characters. Specifically, we encode the image into a text array. We carefully select the encoding to ensure compatibility with text tokenizers of popular models and ensure that each element of the array is encoded by the tokenizers of all evaluated models as a distinct token.

See Figure 2 (bottom) for examples of these presentations. The first two observation spaces are used for models that support visual inputs, while the last observation space is used for text-only models. See the appendix for additional illustrations of these tasks and dataset statistics.

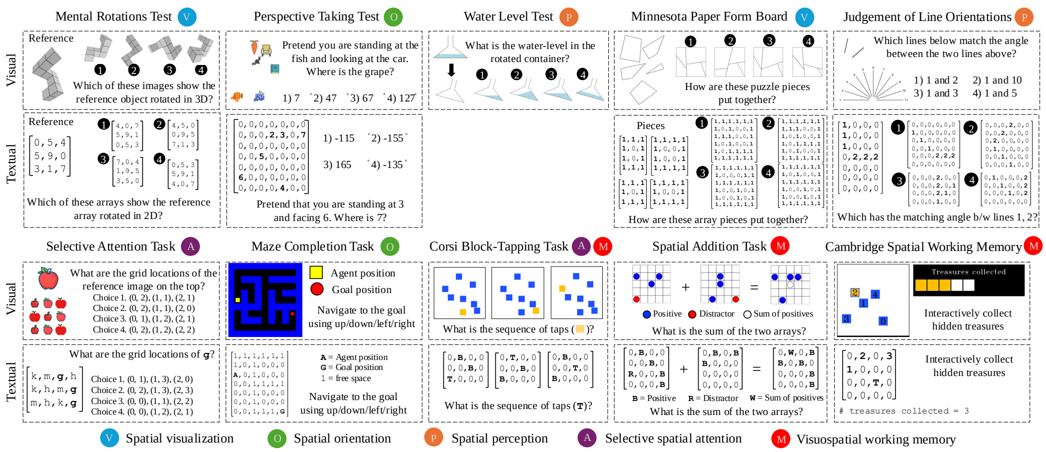

Figure 3: **Small-scale spatial cognition.** We show an example from each small-scale spatial cognition task. All tests with the exception of the water level come with both multimodal and purely textural presentations (for evaluating VLMs and LLMs, respectively). These tasks evaluate cognitive abilities such as spatial visualization, orientation, perception, selective spatial attention and visuospatial working memory. Bolding of characters in this figure is for illustration purposes only.

3.2 Small-scale spatial cognition

In small-scale spatial cognition tasks, we evaluate the models' ability to perceive, imagine, and mentally transform objects or shapes in two and three dimensions. We build on the body of work on visuospatial abilities, which are evaluated in humans via paper-and-pencil tasks (Allen et al., 1996; Weisberg et al., 2014; Meneghetti et al., 2022). These abilities may be used to explain individual differences between participants in large-scale spatial cognition (Meneghetti et al., 2022). We define ten small-scale tasks to evaluate abilities such as spatial perception, spatial visualization, spatial orientation, mental rotation, selective attention, and visuospatial working memory. See Figure 1(bottom) and Figure 3 for an overview.

6. **Mental rotation test (MRT).** This was introduced by Vandenberg & Kuse (1978) as a test of spatial visualization, i.e., the ability to mentally manipulate 2D or 3D stimuli. The original MRT contained 20 items, where each item consisted of a criterion figure, two correct alternatives, and two distractors (Vandenberg & Kuse, 1978). The criterion figure is a perspective rendering of a 3D criterion shape from Shepard & Metzler (1971). The correct alternatives are rotated versions of the criterion shape, where the rotation is applied in the 2D image space on the criterion figure, or along the vertical axis in 3D for the criterion shape. The distractors are rotated mirror-images of the criterion shape or renderings of other criterion shapes. The goal was to identify the two correct alternatives from the four choices. We implement a version of MRT with one correct choice and three distractors, and incorporate rotations along multiple axes (Peters et al., 1995). Since the 3D implementation of MRT can only be evaluated with images, our text-only version of MRT uses a 2D implementation, akin to the card rotations test (French et al., 1963).

7. **Perspective taking test (PTT).** This was introduced by Kozhevnikov & Hegarty (2001) as a test of spatial orientation, i.e., the ability to imagine being in a different position in space and seeing the surroundings from the new perspective. An arrangement of objects is shown on a piece of paper. A test participant is asked to take the perspective of standing next to an object (say, object A) facing another (say, object B), and is required to point to a third object (say, object C). This task has been used extensively in subsequent literature (Hegarty & Waller, 2004; Weisberg et al., 2014; Meneghetti et al., 2022). To implement it, we randomly sample N icons of objects like cars, carrots, chairs, and grapes and place them at random locations in an image (with no overlap between objects). We then randomly sample three of the N objects as A, B, and C. We treat this as multiple-choice QA with four options (only one of them correct).

8. **Water level test (WLT).** This was introduced by Piaget et al. (1957) as a test of visuospatial perception. Originally, the test was designed to evaluate children's knowledge about the horizontal nature of the surface of water in a sealed bottle regardless of its orientation. Children were presented with bottles partially filled with colored water and asked to imagine the position of

the water if it were tilted. Children had to gesture, draw, or use cardboard cutouts to answer the question (Piaget et al., 1957; Foltz, 1978; Wittig & Allen, 1984). Performance on the water-level test was found to be related to performance on spatial ability tests (Foltz, 1978; Wittig & Allen, 1984). We implement this test in the form of multiple-choice QA (Wittig & Allen, 1984).

9. **Minnesota Paper Form Board test (MPFB).** This is a test of spatial visualization, i.e., the ability to perform multi-step manipulations of complex spatial information (Meneghetti et al., 2022). Specifically, a participant is provided with pieces of a figure and is asked to identify how the pieces fit together (Likert & Quasha, 1941; 1969). We programmatically segment a square into five pieces, and rotate the pieces randomly to generate the final segments. We generate alternate segmentations of a square as negative choices for a multiple-choice QA presentation.

10. **Judgement of Line Orientation test (JLO).** This was introduced by Benton (1994) as a measure of visuospatial perception. The original implementation contained 30 samples presented in a flip-book style, where two lines are shown at the top of each page. The goal is to determine the angles between the two lines by comparing them to an array of reference lines (i.e., pick two reference lines that have same angle between them as the lines at the top). There have been multiple variations of JLO with subsets of the 30 questions for faster evaluation (Spencer et al., 2013). We recreate the JLO test suite by randomly sampling pairs of lines on a 2D plane with an angle between 0 to 180 degrees (in multiples of 18 degrees) and formulate it as multiple-choice QA.

11. **Selective attention task (SAtt).** This is designed to evaluate selective spatial attention, i.e., the ability to selectively attend to a particular region of space while ignoring others (Serences & Kastner, 2014; Pahor et al., 2022). In particular, we use the widely used cancellation task, where the goal is to search for and mark out target stimuli embedded amidst distractors (Della Sala et al., 1992; Brickenkamp & Zillmer, 1998; Dalmaijer et al., 2015; Lacroix et al., 2021; Pahor et al., 2022; Kalina & Walgrave, 2004). The stimuli may be characters (Brickenkamp & Zillmer, 1998; Dalmaijer et al., 2015; Pahor et al., 2022; Della Sala et al., 1992; Kalina & Walgrave, 2004), pictures (Lacroix et al., 2021; Pahor et al., 2022), or icons (Lacroix et al., 2021). We design the task as multiple-choice QA with pictures as the stimuli for visual evaluation and characters as stimuli for textual evaluation. The target stimuli and distractors are arranged on a grid. The answer must be selected from one out of four options. The correct option lists the (row, column) pairs that localize the target stimuli in the grid.

12. **Maze completion task (MCT).** This task was designed to evaluate spatial orientation, planning, and executive functioning (Lacroix et al., 2021). It was used as a neuropsychological test to assess executive function disorders in children (Marquet-Doléac et al., 2010). We programmatically create mazes using Mazelib (Stilley, 2014) and visually render them using a Pacman-like color scheme (similar to BEV images in Figure 2). We treat this as an interactive task, where the model is provided with the maze rendering and is prompted to sequentially select an up/down/left/right action to reach the goal. Upon reaching the goal, the model must select a stop action to successfully complete the task. If the model does not reach the goal within 250 actions, the task is considered a failure. We measure the success rate, i.e., the percentage of mazes where the model reaches the goal within the allotted time.

13. **Corsi block-tapping task (CBTT).** This is designed to assess visuospatial working memory and attention in healthy participants and patients with known or suspected brain damage (Corsi, 1972; Claessen et al., 2015). An examiner demonstrates a sequence of block-tapping movements on a board containing fixed blocks placed in pseudo-random positions. Participants are required to reproduce the same sequence (forward condition) or the inverted sequence (backward condition) of block-tapping movements to succeed. We evaluate frontier models on the forward condition since prior work has not found significant differences between task performance in the forward and backward conditions (Claessen et al., 2015). Specifically, we create a digital Corsi board with N blue-colored blocks that are randomly placed on the board with no overlap (N varies from 5 to 8). We randomly sample a sequence of K taps, where $K \in [4, N]$, where each block is tapped at most once. The taps are digitally rendered on the blocks by changing their color from blue to yellow when tapped, yielding an sequence of K images. After the K images are presented, we provide a rendering of the board with integer IDs assigned to each block and ask the model to reproduce the sequence of taps using these IDs. We treat this as multiple-choice QA and provide four choices of tap sequences, only one of which is correct.

14. **Spatial addition task (SAdd).** This was introduced in the fourth edition of the Wechsler Memory Scale, a suite of neuropsychological tests to evaluate memory function in individuals aged 16 to

90 (Wechsler, 2009). SAdd evaluates visuospatial storage and manipulation in working memory. A test participant is shown a grid with blue and red dots for five seconds. The participant is asked to remember the location of the blue dots and ignore the red dots. The participant is then shown another such grid. The objective is to add the two grids together by following certain rules. If a grid location has a blue dot in exactly one of grids, the result should be blue. If a grid location has blue dots on both grids, the result should be white. We programmatically generate grid pairs with sizes sampled from $\{3, 5, 7, 9\}$ and pseudo-randomly populate them with blue and red dots. We formulate the task as multiple-choice QA, presenting four grids as possible answers, exactly one of which is correct.

15. **Cambridge spatial working memory test (CSWM).** This was designed to evaluate spatial working memory in human subjects (Sahakian et al., 1988). Multiple colored boxes are shown on a screen. A yellow 'treasure' is initially hidden in one of the boxes. The participant must select boxes one at a time to open them and search for the treasure. Once the treasure is found, another treasure is placed in one of the remaining boxes. The intention is for the participant to locate all the yellow treasures via a process of elimination. We programmatically generate instances of this task by randomly sampling $N \in \{3, 4, 5, 6, 7\}$ blue boxes, assigning them to random locations (without overlap), and placing the treasures in each box in random order. The model must proceed interactively. At each step, the game screen is presented to the model with random integers assigned to each box.[2] The model selects a box to check for the hidden treasure, and is presented with an updated game screen. If the treasure was found in the previously selected box, the box becomes yellow. The treasures found so far are also displayed alongside the game screen to indicate progress. The objective is to find all the treasures before a time limit T (determined based on N). The model passes the test if it finds all the treasures.

As with large-scale spatial cognition, we also implement purely textual presentations of these tasks to support evaluation of large language models (LLMs). Figure 3 illustrates both the multimodal and the purely textual presentations. The key idea in instantiating the textual presentations is to encode all spatial information via 2D character arrays. We did not identify a natural such encoding for the Water Level Test (WLT) and did not include a text-only presentation for it for this reason. See the appendix for additional illustrations of these tasks.

4 Experiments

Baselines. We evaluate a number of LLMs and VLMs. Using text-only presentations, we evaluate GPT-4v and GPT-4o (OpenAI, 2023; 2024), Claude 3.5 Sonnet (Anthropic, 2024), the Llama3 family (Dubey et al., 2024), Mistral models such as Mixtral 8x7B, Mixtral 8x22B, and Mistral 123B (Jiang et al., 2024; Mistral AI team, 2024a), and two Yi 1.5 models (Young et al., 2024). Using multimodal presentations, we evaluate GPT-4v and GPT-4o (OpenAI, 2023; 2024), Claude 3.5 Sonnet (Anthropic, 2024), LlaVA-NeXT-Interleave (Li et al., 2024a), Pixtral (Mistral AI team, 2024b), and Phi-3.5-vision (Abdin et al., 2024). We also list the results of a chance baseline that selects an answer at random. For multiple-choice QA tasks, chance is at 25%. For interactive tasks, the chance baseline samples an action at random in each step. We further include human performance for reference for the multiple-choice QA tasks (see the appendix for details).

Implementation details. We use the vLLM inference engine for evaluating the open-source models (Kwon et al., 2023). Since LlaVa-Next-Interleave is not supported by vLLM, it is evaluated via HuggingFace (Wolf et al., 2019). For multiple-choice QA, we randomize the placement of the correct answer among the four choices. By performing multiple trials, we can compute means and standard deviations for each model on each task. (See the appendix for details.) For each task, we implement a prompt that provides a detailed description of the task and the expected response format. The prompts are reproduced in the appendix.

Large-scale spatial cognition results. The results are shown in Table 1, grouped by presentation modality (ego image, BEV image, BEV text). For image-based presentations, we evaluate GPT-4v and GPT-4o because they support video understanding (via a succession of images). For BEV text, we evaluate both open and closed LLMs. We also list the performance of the chance baseline for calibration, as well as human performance (see the appendix for details). In the text-only

[2] The boxes' integer IDs are randomized in each step, forcing the model to remember their locations.

Observation space: Ego image

Method	Direction estimation	Distance estimation	Map sketching	Route retracing	Novel shortcuts	Average
Human	82.8	83.2	96.6	-	-	-
GPT-4o	32.0 ±4.1	36.5 ±5.0	33.3 ±4.1	6.6 ±3.6	6.4 ±1.0	23.0
GPT-4v	29.7 ±0.3	31.9 ±2.7	20.0 ±11.8	1.6 ±1.2	3.9 ±0.9	17.4
Chance	25.0	25.0	25.0	0.0	0.0	15.0

Observation space: BEV image

Method	Direction estimation	Distance estimation	Map sketching	Route retracing	Novel shortcuts	Average
Human	82.9	82.5	100.0	-	-	-
GPT-4o	29.5 ±5.5	31.9 ±1.0	33.3 ±3.3	23.6 ±3.1	25.9 ±2.0	28.8
GPT-4v	26.3 ±3.0	29.3 ±4.1	45.0 ±5.0	13.7 ±5.2	15.3 ±3.0	25.9
Chance	25.0	25.0	25.0	0.0	0.0	15.0

Observation space: BEV text

Method	Direction estimation	Distance estimation	Map sketching	Route retracing	Novel shortcuts	Average
Human	66.7	76.5	66.7	-	-	-
GPT-4o	28.7 ±4.1	33.3 ±1.7	46.7 ±4.1	27.5 ±3.2	26.6 ±0.1	32.6
Mistral 123B	30.5 ±5.1	28.9 ±5.7	38.3 ±5.5	20.3 ±2.8	19.9 ±3.0	27.6
Llama 3.1 405B	28.3 ±4.0	24.8 ±1.9	35.8 ±4.3	22.6 ±3.6	25.6 ±2.5	27.4
GPT-4v	30.7 ±4.1	26.5 ±2.7	40.8 ±6.0	20.6 ±5.8	15.4 ±2.0	26.8
Llama 3 70B	27.0 ±2.2	30.4 ±1.9	35.0 ±8.3	13.2 ±9.2	5.3 ±4.1	22.2
Llama 3.1 70B	24.5 ±1.4	18.9 ±4.2	32.5 ±6.4	13.2 ±1.2	19.8 ±2.6	21.8
Yi 1.5 34B	26.2 ±4.7	35.7 ±1.4	35.0 ±10.7	3.2 ±0.2	1.1 ±1.6	20.2
Mixtral 8x22B	21.3 ±1.9	19.4 ±1.4	39.2 ±12.6	1.5 ±1.4	3.9 ±1.7	17.0
Yi 1.5 9B	10.8 ±1.0	20.0 ±3.7	35.0 ±5.0	5.0 ±2.2	1.3 ±1.5	14.4
Llama 3 8B	22.5 ±2.9	24.6 ±2.1	23.3 ±7.1	0.0 ±0.0	1.1 ±1.6	14.3
Llama 3.1 8B	14.0 ±1.6	16.9 ±3.4	33.3 ±9.7	0.9 ±1.2	2.5 ±0.7	13.5
Mixtral 8x7B	15.8 ±2.0	16.1 ±1.4	30.0 ±8.2	1.1 ±1.6	1.1 ±1.6	12.8
Chance	25.0	25.0	25.0	0.0	0.0	15.0

Table 1: **Large-scale spatial cognition results.** The three tables show results for different observation spaces. Results below 50% of human performance are gray. Methods are sorted based on their overall performance.

modality, GPT-4o attains the highest average performance. Mistral 123B is the highest-performing open model. All evaluated models struggle with large-scale spatial cognition, falling significantly below human performance on direction estimation, distance estimation, and map sketching, and less than 30% SPL on route retracing and novel shortcuts, even with allocentric presentation. With egocentric multimodal presentation (the closest counterpart to classic experimental protocols in animal cognition), the models are near chance level on all tasks.

Human performance ranges from 80% to 100% accuracy on image-based presentations of the multiple-choice QA tasks. Since perceiving large sequences of text arrays is non-trivial for humans, the performance drops to 65%–80% for the text presentations.

Small-scale spatial cognition results. The results are shown in Table 2. With multimodal presentations, we benchmark GPT-4o, GPT-4v, Claude 3.5 Sonnet, and a number of open multimodal models. With purely textual presentations, we benchmark both open and closed models. We also list the performance of the chance baseline for calibration, as well as human performance (see the appendix for details).

Performance of some model classes (e.g., GPT-4o, GPT-4v, Claude) on purely textual presentations is considerably higher than on multimodal presentations. The best-performing models, Claude and GPT-4o, achieve 43.8% and 40.1% average accuracies in the multimodal regime and 64.5% and 65.2% average accuracies with purely textual presentations. (Chance is $< 25\%$.) We attribute this in part to the simplified nature of the text-only implementations of the tasks (e.g., the text-only presentation of mental rotation uses only 2D shapes and constrained 2D rotations) and in part to the relative developmental maturity of large language models (LLMs) versus multimodal models.

On tasks that evaluate visuospatial working memory (specifically SAtt, CBTT, SAdd, and CSWM), the strongest LLMs perform well. On selective attention (SAtt), GPT-4o, Claude, Mistral 123B, and GPT-4v all achieve over 95% accuracy, matching or outperforming the human performance on this task. On the other hand, models perform poorly on maze completion (MCT), in both presentation modalities. (Note that the models operate with full visibility, as illustrated in Figure 3.) With multimodal presentation, all evaluated models are near chance on perspective taking (PTT) and the

Multimodal											
Method	MRT	PTT	WLT	MPFB	JLO	SAtt	MCT	CBTT	SAdd	CSWM	Average
Human	78.5	80.0	94.0	84.0	82.0	95.0	-	100.0	98.0	-	-
Claude 3.5 Sonnet	29.9 ±3.8	21.8 ±2.9	37.0 ±4.6	35.5 ±7.0	40.5 ±3.8	**90.5** ±3.5	2.2 ±1.8	56.5 ±3.8	48.0 ±6.2	76.7 ±2.5	43.8
GPT-4o	33.3 ±1.9	26.5 ±3.6	59.0 ±10.8	27.0 ±2.2	26.5 ±5.9	70.2 ±1.8	10.4 ±1.0	68.0 ±2.0	40.5 ±7.1	40.0 ±0.0	40.1
GPT-4v	32.3 ±0.3	28.0 ±2.0	35.0 ±7.7	22.5 ±4.1	26.5 ±6.8	59.8 ±4.4	0.7 ±1.0	44.5 ±3.0	32.0 ±4.5	26.7 ±3.4	30.8
Pixtral 12B	28.3 ±3.1	23.2 ±4.9	43.0 ±7.0	30.5 ±7.9	24.5 ±7.3	36.0 ±3.9	OOM	39.5 ±3.0	28.5 ±6.1	OOM	25.4*
Phi-3.5-vision	24.1 ±1.0	27.0 ±3.2	22.5 ±7.9	26.0 ±0.0	21.0 ±4.1	44.0 ±4.6	OOM	33.0 ±4.6	22.0 ±6.8	OOM	22.0*
Llava interleave 7B	25.1 ±3.2	25.8 ±5.8	25.0 ±8.5	25.0 ±3.3	24.0 ±5.7	32.0 ±4.9	OOM	25.5 ±5.7	27.0 ±4.1	OOM	20.9*
Chance	25.0	25.0	25.0	25.0	25.0	25.0	0.0	25.0	25.0	33.8 ±5.4	23.4

Text-only										
Method	MRT	PTT	MPFB	JLO	SAtt	MCT	CBTT	SAdd	CSWM	Average
Human	90.0	75.0	92.0	98.0	96.0	-	100.0	98.0	-	-
GPT-4o	41.9 ±6.2	55.5 ±3.9	50.5 ±9.6	66.5 ±4.8	**98.8** ±0.4	21.5 ±3.8	82.5 ±1.7	**93.5** ±3.6	76.7 ±2.5	65.2
Claude 3.5 Sonnet	37.5 ±1.8	50.0 ±7.5	45.0 ±6.7	70.5 ±4.3	**97.0** ±1.0	10.0 ±1.1	**97.5** ±0.9	**91.5** ±4.3	82.0 ±3.3	64.5
Mistral 123B	39.4 ±6.2	44.8 ±4.0	48.5 ±5.2	57.0 ±5.4	**97.5** ±0.5	14.8 ±2.8	88.5 ±0.9	**92.5** ±0.9	62.0 ±2.8	60.5
GPT-4v	41.2 ±7.2	67.5 ±6.5	34.0 ±6.0	62.0 ±4.0	**95.8** ±1.3	3.7 ±1.0	87.5 ±3.6	79.0 ±2.2	45.3 ±2.5	57.3
Llama 3 70B	28.1 ±9.2	29.2 ±2.4	38.5 ±3.8	42.5 ±0.9	71.8 ±3.8	1.5 ±1.0	52.5 ±5.7	62.5 ±5.4	34.0 ±5.9	40.0
Mixtral 8x22B	26.9 ±3.2	24.5 ±5.2	31.0 ±5.9	36.0 ±5.1	73.5 ±3.6	1.5 ±2.1	55.0 ±3.3	68.0 ±6.8	17.3 ±2.5	37.0
Yi 1.5 34B	20.6 ±6.0	28.0 ±2.1	34.5 ±4.6	33.5 ±3.6	58.2 ±4.3	0.7 ±1.0	35.5 ±8.4	41.5 ±0.9	24.0 ±0.0	30.7
Yi 1.5 9B	21.2 ±1.2	23.8 ±2.7	30.0 ±3.2	24.5 ±5.5	48.2 ±4.0	0.7 ±1.0	36.5 ±4.6	51.5 ±8.9	24.7 ±8.4	29.0
Llama 3 8B	14.4 ±1.1	25.8 ±5.1	26.0 ±4.2	27.0 ±1.7	46.0 ±3.7	0.0 ±0.0	27.5 ±7.1	30.0 ±7.3	26.0 ±6.5	24.7
Mixtral 8x7B	19.4 ±4.5	10.5 ±0.9	29.5 ±5.7	27.5 ±7.5	39.0 ±5.4	0.0 ±0.0	22.5 ±3.8	43.5 ±3.3	22.7 ±4.1	23.8
Chance	25.0	25.0	25.0	25.0	25.0	0.0	25.0	25.0	33.0 ±5.3	23.1

Table 2: **Small-scale spatial cognition results.** The two tables show results for multimodal and text-only presentations, respectively. Results below 50% of human performance are gray, results above 90% of human performance are **bold**. Methods are sorted based on their average performance. (Some multimodal models ran out of memory on MCT and CSWM tasks; their accuracy is taken to be 0 for calculating the average.)

Minnesota paper form board test (MPFB). On mental rotation (MRT), the best models are near chance with multimodal presentation, which uses 3D shapes, and only marginally better with purely textual presentation, which uses 2D arrays and constrained rotations.

Humans perform well, achieving over 80% accuracy on the majority of the multiple-choice QA tasks with both text-only and multimodal presentations. Humans perform better on the textual presentations of tasks like MRT, MPFB and JLO than their vision counterparts due to the simplified nature of the text-only implementations.

5 Discussion

We presented SPACE, a benchmark for spatial cognition in frontier models. Our evaluation of contemporary models brings up intriguing questions and opportunities for further investigation. First, our results underscore that frontier models exhibit a fundamentally different form of intelligence from what has been observed (and studied) in humans and animals. No biological intelligence we have encountered has exhibited such advanced skill in some aspects of higher cognition (Trinh et al., 2024) while failing so profoundly in basic spatial cognition. This is particularly intriguing because in biological intelligence, spatial cognition is considered a prerequisite for higher cognition, and breakdowns in spatial cognition are diagnostic of higher-level disorders (Cappa, 2008; Possin, 2010; Verghese et al., 2017; Cammisuli et al., 2024). From a scientific standpoint, the constellation of traits exhibited by frontier models is fascinating and may inspire a new cognitive science (Simon, 2019). As a precautionary stance, we can refrain from drawing analogies based on experience with biological cognition. (E.g., "a model won the Mathematics Olympiad therefore it possesses a comparable cognitive repertoire to a human Olympiad winner and could be expected to have comparable skill in other domains".)

Could deficiencies in spatial cognition be causally linked to some of the puzzling breakdowns exhibited by contemporary frontier models in higher-level tasks? What is the roadmap for bringing spatial cognition in frontier models up to the level of animal cognition (and perhaps beyond)? Is this a prerequisite for attaining some of the more far-reaching aspirations of contemporary artificial intelligence research? Does embodiment play a role, as it has in prior forms of intelligence (Smith & Gasser, 2005; Savva et al., 2019)? Or will artificial cognition continue to develop along a fundamentally different ontogenetic path? We expect further advances to increase the robustness and generality of frontier models, and to continue to broaden our understanding of the nature of intelligence.

REFERENCES

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv:2404.14219*, 2024.

Gary L Allen, Kathleen C Kirasic, Shannon H Dobson, Richard G Long, and Sharon Beck. Predicting environmental learning from spatial abilities: An indirect route. *Intelligence*, 22(3), 1996.

David I Anderson, Joseph J Campos, David C Witherington, Audun Dahl, Monica Rivera, Minxuan He, Ichiro Uchiyama, and Marianne Barbu-Roth. The role of locomotion in psychological development. *Frontiers in Psychology*, 4, 2013.

Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. On evaluation of embodied navigation agents. *arXiv:1807.06757*, 2018.

Anthropic. Introducing claude 3.5 sonnet, 2024. `https://www.anthropic.com/news/claude-3-5-sonnet`.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.

Leopold Aschenbrenner. Situational awareness: The decade ahead, 2024. `https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf`.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, Caiming Xiong, Ran Xu, Yejin Choi, and Ludwig Schmidt. MINT-1T: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *arXiv:2406.11271*, 2024.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705), 2018.

Arthur Lester Benton. *Contributions to neuropsychological assessment: A clinical manual*. Oxford University Press, USA, 1994.

Mark Blades and Christopher Spencer. The development of children's ability to use spatial representations. *Advances in child development and behavior*, 25, 1994.

Youcef Bouchekioua, Aaron P Blaisdell, Yutaka Kosaki, Iku Tsutsui-Kimura, Paul Craddock, Masaru Mimura, and Shigeru Watanabe. Spatial inference without a cognitive map: the role of higher-order path integration. *Biological Reviews*, 96(1), 2021.

R Brickenkamp and E Zillmer. Test d2: concentration-endurance test. *Gottingen Ger. CJ Hogrefe*, 1998.

Rodney Brooks. Rodney brooks' three laws of artificial intelligence, 2024. `https://rodneybrooks.com/rodney-brooks-three-laws-of-artificial-intelligence/`.

Davide Maria Cammisuli, Gloria Marchesi, Virginia Bellocchio, Edoardo Nicolò Aiello, Barbara Poletti, Federico Verde, Vincenzo Silani, Nicola Ticozzi, Stefano Zago, Teresa Difonzo, et al. Behavioral disorders of spatial cognition in patients with mild cognitive impairment due to alzheimer's disease (the bdsc-mci project): Ecological validity of the corsi learning suvra-span test. *Journal of Personalized Medicine*, 14(5), 2024.

SF Cappa. *Cognitive neurology: a clinical textbook*. Oxford University Press, 2008.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas J. Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024b.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv:2107.03374*, 2021.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv:2406.01584*, 2024.

François Chollet. On the measure of intelligence. *arXiv:1911.01547*, 2019.

Michiel HG Claessen, Ineke JM Van Der Ham, and Martine JE Van Zandvoort. Computerization of the standard corsi block-tapping task affects its underlying cognitive concepts: a pilot study. *Applied Neuropsychology: Adult*, 22(3), 2015.

Philip Michael Corsi. *Human memory and the medial temporal region of the brain.* Phd thesis, McGill University, 1972. https://escholarship.mcgill.ca/concern/theses/05741s554.

Edwin S Dalmaijer, Stefan Van der Stigchel, Tanja CW Nijboer, Tim HW Cornelissen, and Masud Husain. Cancellationtools: All-in-one software for administration and analysis of cancellation tasks. *Behavior Research Methods*, 47, 2015.

Dawson-Haggerty et al. trimesh, 2019. https://trimesh.org/.

Sergio Della Sala, Marcella Laiacona, Hans Spinnler, and Chiara Ubezio. A cancellation test: its reliability in assessing attentional deficits in alzheimer's disease. *Psychological medicine*, 22(4), 1992.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024.

Pul Ashby Foltz. Adult performance on piaget's water level task and its relation of spatial orientation and visualization. Master's thesis, University of Richmond, 1978. https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=1424&context=masters-theses.

Nigel Foreman, Denny Foreman, Alison Cummings, and Sandra Owens. Locomotion, active choice, and spatial memory in children. *The Journal of general psychology*, 117(2), 1990.

John W French, Ruth B Ekstrom, and Leighton A Price. Manual for kit of reference tests for cognitive factors (revised 1963). *(No Title)*, 1963.

Andrea Frick and Wenke Möhring. A matter of balance: Motor control is related to children's spatial and proportional reasoning skills. *Frontiers in Psychology*, 6, 2016.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024a.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. *arXiv:2404.12390*, 2024b.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023.

Maya Geva-Sagiv, Liora Las, Yossi Yovel, and Nachum Ulanovsky. Spatial cognition in bats and rats: from sensory acquisition to multiscale maps and navigation. *Nature Reviews Neuroscience*, 16(2), 2015.

Sabine Gillner and Hanspeter A Mallot. Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of cognitive neuroscience*, 10(4), 1998.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4), 2019.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Mary Hegarty and David Waller. A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32(2), 2004.

Mary Hegarty, Daniel R Montello, Anthony E Richardson, Toru Ishikawa, and Kristin Lovelace. Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, 34(2), 2006.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021b.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*, 2024.

Petra Jansen. The dissociation of small-and large-scale spatial abilities in school-age children. *Perceptual and Motor Skills*, 109(2), 2009.

Petra Jansen and Martin Heil. The relation between motor development and mental rotation ability in 5-to 6-year-old children. *International Journal of Developmental Science*, 4(1), 2010.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv:2401.04088*, 2024.

Ashley N Kalina and Suzie A Walgrave. Normative evaluation of a letter cancellation instrument for the assessment of sustained attention: A construct validation study. *The Journal of Undergraduate Research*, 2(1), 2004.

Maria Kozhevnikov and Mary Hegarty. A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29, 2001.

Maria Kozhevnikov, Michael A Motes, and Mary Hegarty. Spatial visualization in physics problem solving. *Cognitive science*, 31(4), 2007.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, 2023.

Emilie Lacroix, Stéphanie Cornet, Naima Deggouj, and Martin Gareth Edwards. The visuo-spatial abilities diagnosis (vsad) test: Evaluating the potential cognitive difficulties of children with vestibular impairment through a new tablet-based computerized test battery. *Behavior Research Methods*, 53, 2021.

Barbara Landau. Spatial cognition. In *Encyclopedia of the Human Brain*. Elsevier, 2002.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In *NeurIPS*, 2023.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv:2407.07895*, 2024a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

Rensis Likert and WH Quasha. *Minnesota Paper Form Board Test*. Psychological Corporation, 1941.

Rensis Likert and William H Quasha. *Revised Minnesota paper form board test*. Psychological Corporation, 1969.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*, 2024.

Hanspeter A Mallot. *From Geometry to Behavior: An Introduction to Spatial Cognition*. MIT Press, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019.

Jérôme Marquet-Doléac, Régis Soppelsa, and Jean-Michel Albaret. *Laby 5-12: Test des labyrinthes*. Hogrefe, 2010.

John C Marshall and Gereon R Fink. Spatial cognition: Where we were and where we are. *Neuroimage*, 14(1), 2001.

Chiara Meneghetti, Clara Zancada-Menéndez, Patricia Sampedro-Piquero, Laudino Lopez, Massimiliano Martinelli, Lucia Ronconi, and Barbara Rossi. Mental representations derived from navigation: The role of visuo-spatial abilities and working memory. *Learning and Individual Differences*, 49, 2016.

Chiara Meneghetti, Laura Miola, Enrico Toffalini, Massimiliano Pastore, and Francesca Pazzaglia. Learning from navigation, and tasks assessing its accuracy: The role of visuospatial abilities and wayfinding inclinations. *Journal of Environmental Psychology*, 75, 2021.

Chiara Meneghetti, Laura Miola, Tommaso Feraco, and Veronica Muffato. Individual differences in navigation: An introductory overview. In *Prime Archives in Psychology*. Vide Leaf, 2nd edition, 2022.

Emil W Menzel. Chimpanzee spatial memory organization. *Science*, 182(4115), 1973.

Mistral AI team. Large enough, 2024a. `https://mistral.ai/news/mistral-large-2407/`.

Mistral AI team. Announcing pixtral 12b, 2024b. `https://mistral.ai/news/pixtral-12b/`.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps in large language models with cogeval: No emergent planning. *Advances in neural information processing systems*, 37, 2023.

Arsenii Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. The conceptarc benchmark: Evaluating understanding and generalization in the ARC domain. *Trans. Mach. Learn. Res.*, 2023, 2023.

Nora S. Newcombe. *Making space: The development of spatial representation and reasoning*. MIT Press, 2000.

Nora S. Newcombe. Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, 34(2), 2010.

Nora S. Newcombe. Spatial Cognition. In *Open Encyclopedia of Cognitive Science*. MIT Press, 2024.

John O'Keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford University Press, 1978.

OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.

OpenAI. Hello gpt-4o, 2024. `https://openai.com/index/hello-gpt-4o/`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Anja Pahor, Randy E Mester, Audrey A Carrillo, Eunice Ghil, Jason F Reimer, Susanne M Jaeggi, and Aaron R Seitz. Ucancellation: A new mobile measure of selective attention and concentration. *Behavior Research Methods*, 54(5), 2022.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2023.

Francesca Pazzaglia and Holly A Taylor. Perspective, instruction, and cognitive style in spatial representation of a virtual environment. *Spatial Cognition and Computation*, 7(4), 2007.

Michael Peters, Bruno Laeng, Kerry Latham, Marla Jackson, Raghad Zaiyouna, and Chris Richardson. A redrawn vandenberg and kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition*, 28(1), 1995.

Roger Paul Peters. *Wolf-sign: Scents And Space In A Wide-ranging Predator.* University of Michigan, 1974.

Jean Piaget, Baerbel Inhelder, F. J. Langdon, and J. L. Lunzer. The child's conception of space. *British Journal of Educational Studies*, 5(2), 1957.

Katherine L Possin. Visual spatial cognition in neurodegenerative disease. *Neurocase*, 16(6), 2010.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Ahad Rana. Common crawl – building an open web-scale crawl using hadoop, 2010. https://www.slideshare.net/hadoopusergroup/common-crawlpresentation.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024.

Anthony E Richardson, Daniel R Montello, and Mary Hegarty. Spatial knowledge acquisition from maps and from navigation in real and virtual environments. *Memory & cognition*, 27(4), 1999.

Barbara J Sahakian, Robin G Morris, John L Evenden, Andrew Heald, Raymond Levy, Michael Philpot, and Trevor W Robbins. A comparative study of visuospatial memory and learning in alzheimer-type dementia and parkinson's disease. *Brain*, 111(3), 1988.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9), 2021.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

John T. Serences and Sabine Kastner. A multi-level account of selective attention. *The Oxford Handbook of Attention*, 2014.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972), 1971.

Herbert A. Simon. *The Sciences of the Artificial*. MIT Press, 3rd edition, 2019.

Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial Life*, 11(1-2), 2005.

Robert J Spencer, Carrington R Wendell, Paul P Giggey, Stephen L Seliger, Leslie I Katzel, and Shari R Waldstein. Judgment of line orientation: an examination of eight short forms. *Journal of clinical and experimental neuropsychology*, 35(2), 2013.

John Stilley. mazelib: A python api for creating and solving mazes, 2014. https://github.com/john-science/mazelib.

Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023. https://github.com/togethercomputer/RedPajama-Data.

Sivan Toledo, David Shohami, Ingo Schiffner, Emmanuel Lourie, Yotam Orchan, Yoav Bartan, and Ran Nathan. Cognitive map–based navigation in wild bats revealed by a new high-throughput tracking system. *Science*, 369(6500), 2020.

Edward C Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4), 1948.

Luca Tommasi, Cinzia Chiandetti, Tommaso Pecchia, Valeria Anna Sovrano, and Giorgio Vallortigara. From natural geometry to spatial cognition. *Neuroscience & Biobehavioral Reviews*, 36(2), 2012.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, 2024.

Trieu Trinh, Yuhuai Wu, Quoc Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 2024.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36, 2024.

Steven G Vandenberg and Allan R Kuse. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, 47(2), 1978.

Marina Vasilyeva and Stella F Lourenco. Development of spatial cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3), 2012.

Joe Verghese, Richard Lipton, and Emmeline Ayers. Spatial navigation and risk of cognitive impairment: A prospective cohort study. *Alzheimer's & Dementia*, 13(9), 2017.

David Ed Waller and Lynn Ed Nadel. *Handbook of Spatial Cognition*. American Psychological Association, 2013.

Lu Wang, Allan S Cohen, and Martha Carr. Spatial ability at two scales of representation: A meta-analysis. *Learning and Individual Differences*, 36, 2014.

David Wechsler. *WMS-IV: Wechsler memory scale*. Pearson, 2009.

Steven M Weisberg, Victor R Schinazi, Nora S Newcombe, Thomas F Shipley, and Russell A Epstein. Variations in cognitive maps: understanding individual differences in navigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3), 2014.

Michele Andrisin Wittig and Mary J Allen. Measurement of adult performance on piaget's water horizontality task. *Intelligence*, 8(4), 1984.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771*, 2019.

Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=xkiflfKCw3`.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *ICML*, 2024.

Eunice Yiu, Maan Qraitem, Charlie Wong, Anisa Noor Majhi, Yutong Bai, Shiry Ginosar, Alison Gopnik, and Kate Saenko. Kiva: Kid-inspired visual analogies for testing large multimodal models. *arXiv:2407.17773*, 2024.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01.ai. *arXiv:2403.04652*, 2024.

Christopher J Young, Susan C Levine, and Kelly S Mix. The connection between spatial and mathematical ability across development. *Frontiers in psychology*, 9, 2018.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

A APPENDIX

A.1 ECOLOGICAL COMPATIBILITY OF MULTIMODAL INPUTS WITH FRONTIER MODELS

Our results in Section 4 suggest that state-of-the-art frontier models fail in basic spatial cognition tasks presented in SPACE. These failures could be attributed to the lack of spatial cognition in these models. Alternatively, these failures could be due to models not comprehending the inputs presented to them (i.e., the inputs are not *ecologically compatible* with the models). To rule out this alternate possibility, we design additional tests unrelated to spatial cognition on the same vision / text inputs used in our benchmark. If models succeed on these tests, it would suggest that the inputs are ecologically compatible with them since they can understand and perform tasks using these inputs. In each test, we pose a series of multiple-choice questions evaluating a model's fine-grained understanding of the visual inputs (and textual in some cases). Next, we describe the tests we designed.

Test 1: Given BEV image / text inputs (see Figure 2), answer the following questions:

Q1. What is the size of the grid (H x W)?
Q2. What is your current (x, y) location?
Q3. What are the (x, y) locations of all navigable cells? Include cells containing landmarks and your current position.
Q4. What are the (x, y) locations of all obstacle cells?
Q5. What are the landmarks visible in the image / array?
Q6. What are the locations of the landmarks visible in the image / array?

Test 2: Given an ego image (see Figure 2), answer the following questions:

Q1. What is the name of the landmark visible in the image?
Q2. Is the landmark <name> in the left half of the image?
Q3. Is the landmark <name> in the right half of the image?
Q4. Is the landmark <name> in the central section of the image?

Test 3: Given two consecutive ego images from a walkthrough (see Figure 4), answer the following question:

Q1. What is the action taken to go from image 1 to image 2 (move forward, turn left, turn right, wait/do nothing)?

Test 4: Given a perspective taking test image / text array (see Figures 10 and 11), answer the following questions:

Q1. How many objects / non-zero locations are present in the image / array?
Q2. What objects / non-zero locations are present in the image / array?
Q3. Is <object / location> to the left of <object / location> in the image / array?
Q4. Is <object / location> to the above <object / location> in the image / array?

Test 5: Given water level test images, answer the following questions:

Q1. Is there water in the water container?
Q2. From image 1 to image 2, is the water container rotated to the left, right or not rotated at all?
Q3. From image 1 to image 2, what is the absolute rotation angle of the water container (in degrees)?

Test 6: Given a selective attention task grid of icons / characters, answer the following questions:

Q1. How many total objects / characters are present in the image / grid (including repetitions)?
Q2. What is the size of the grid of objects / characters (width x height)?
Q3. How many unique objects / characters are present in the grid (ignore repetitions)?

Results discussion: We evaluate GPT-4o and GPT-4v on these tests. The results are shown in Tables 3 and 4. Both models largely understand BEV image and text inputs (test 1). However, they

Multimodal evaluation

Model	Test 1							Test 2					Test 3
	Q1	Q2	Q3	Q4	Q5	Q6	Avg.	Q1	Q2	Q3	Q4	Avg.	Q1
GPT-4o	30.4	78.2	89.4	87.8	100.0	93.6	79.9	100.0	84.0	95.5	83.5	92.6	59.3
GPT-4v	55.8	86.2	89.8	91.2	99.8	79.2	83.6	98.0	45.0	36.5	56.5	66.8	48.0

Text-only evaluation

Model	Test 1							Test 2					Test 3
	Q1	Q2	Q3	Q4	Q5	Q6	Avg.	Q1	Q2	Q3	Q4	Avg.	Q1
GPT-4o	100.0	100.0	77.5	85.6	100.0	82.8	90.9	-	-	-	-	-	-
GPT-4v	100.0	100.0	96.8	91.0	100.0	77.6	94.2	-	-	-	-	-	-

Table 3: Measuring ecological compatibility of multimodal inputs with frontier models (part 1)

Multimodal evaluation

Model	Test 4					Test 5				Test 6			
	Q1	Q2	Q3	Q4	Avg.	Q1	Q2	Q3	Avg.	Q1	Q2	Q3	Avg.
GPT-4o	99.6	99.6	89.8	87.7	92.4	100.0	73.9	38.7	64.0	83.0	90.5	58.2	77.2
GPT-4v	78.3	87.4	78.4	76.0	79.0	100.0	56.3	32.4	55.4	74.5	88.0	35.8	66.0

Text-only evaluation

Model	Test 4					Test 5				Test 6			
	Q1	Q2	Q3	Q4	Avg.	Q1	Q2	Q3	Avg.	Q1	Q2	Q3	Avg.
GPT-4o	97.6	99.6	99.5	94.2	97.4	-	-	-	-	100.0	100.0	99.5	99.8
GPT-4v	96.8	98.1	92.2	76.8	96.8	-	-	-	-	99.5	100.0	94.8	98.1

Table 4: Measuring ecological compatibility of multimodal inputs with frontier models (part 2)

fall short in calculating the grid size for BEV images (Q1). GPT-4o understands egocentric images, i.e., recognizes and localizes landmarks in egocentric images (test 2). GPT-4v recognizes landmarks well (Q1), but performs poorly in localization (Q2, Q3 and Q4). Both GPT-4o and GPT-4v perform poorly on action estimation (test 3) and estimation of water container rotations (Q2 and Q3 in test 5). GPT-4o excels in understanding the perspective taking inputs with multimodal and text-only presentations (test 4). GPT-4v also performs well on test 4, but is worse with multimodal inputs when compared to text-only inputs. Finally, both GPT-4o and GPT-4v perform adequately with counting objects (Q1 in test 6) and grid sizes (Q2 in test 6) on selective attention task inputs with multimodal inputs. However, they struggle to calculate the number of unique objects / characters (Q3 in test 6). Both GPT-4o and GPT-4v excel at the text-only presentation of test 6.

Our results indicate that state-of-the-art models can understand multimodal and text-only inputs provided in our benchmark. They perform well in most of the tests, but have very specific shortcomings (e.g., localizing landmarks in ego images for GPT-4v, understanding rotations of water containers and counting unique characters / objects in a grid). Importantly, the average results on each test is much higher than the SPACE task counterparts. For example, even though GPT-4o and GPT-4v understand BEV text inputs nearly perfectly (test 1), they perform poorly in the BEV text versions of the large-scale spatial cognition tasks (see Table 1). Similarly, even though GPT-4o understands the perspective taking inputs nearly perfectly for both text-only and multimodal presentations, it performs poorly on the perspective taking task in SPACE (see Table 2). Therefore, the failure of frontier models on SPACE is most likely due to their lack of spatial cognition, and not because they cannot understand the inputs presented to them.

A.2 SPACE EXAMPLES

We illustrate examples for each task from our proposed SPACE benchmark.

Large-scale spatial cognition

- **Egocentric image observations:** Figures 4, 5

- **BEV image observations**[3]**:** Figures 6, 7

Small-scale spatial cognition

- **MRT:** Figures 8 and 9
- **PTT:** Figures 10 and 11
- **WLT:** Figure 12
- **MPFB:** Figures 13 and 14
- **JLO:** Figures 15 and 16
- **MCT:** Figures 23 and 24
- **CBTT:** Figures 19 and 20
- **SAdd:** Figures 21 and 22
- **CSWM:** Figures 25 and 26

A.3 Implementation details

We provide additional implementation details about our experimental setup in this section.

Human performance: We obtain human performance on SPACE tasks by evaluating 29 participants (aged 20 - 50). We evenly divide the questions from our benchmark across all participants. For each participant, we provide HTML files containing a subset of questions from each SPACE task and the corresponding choices. The HTML files contain formatted versions of the prompts used to evaluate frontier models. We do not provide any additional instructions or background information about how to solve the tasks. For efficiency, we group all questions corresponding to a single environment in the large-scale spatial cognition tasks. Each participant is assigned to view a video walkthrough from one environment and asked to answer a series of questions about that same environment. This is in line with classical protocols in human cognition (Allen et al., 1996; Hegarty et al., 2006; Pazzaglia & Taylor, 2007; Weisberg et al., 2014; Meneghetti et al., 2016; 2021). We further provide a CSV file where the participant is instructed to enter the answers. We instruct the participants to perform all tasks mentally without any aids like pen and paper. Each participant is estimated to have taken 60 to 90 minutes to answer all the questions. The participants send us their responses and we evaluate them collectively. We denote the collective performance of all participants as the human performance in Tables 1 and 2.

Note that we establish the human baseline only for the multiple-choice QA tasks since it was straightforward to share the test materials with the participants online and obtain their answers. The interactive tasks would require us to meet participants in person to perform evaluations and we were not equipped to do this.

Randomized trails for evaluation: For multiple-choice QA, we randomize the placement of the correct answer among the four choices such that it appears in each of the four positions once, yielding four trials per question. For each trial, we evaluate the performance over all questions to obtain the average accuracy. For interactive tasks like route retracing, novel shortcuts, MCT and CSWM, we evaluate each model in three independent trials and obtain the corresponding metrics. By performing multiple trials, we can compute means and standard deviations for each model on each task across the trials.

Image pre-processing: For majority of our experiments, we use square images. We provide the images to models as is without pre-processing. For most models (especially closed-source ones), the processing of the image beyond the input stage is outside our control. We rely on the model creators to correctly process the images. The exact image resolution and aspect ratios are task-dependent and listed in Table 5. For egocentric video inputs in the large-scale spatial cognition tasks, the number of frames varies from 61 to 240. Since GPT-4o and GPT-4v APIs did not permit 240+ frames as inputs, we subsample the video frames by a factor of 2 before providing them to the model. For BEV video inputs, the number of frames varies from 13 to 72. We provide them as is to the model.

[3]BEV text observations are obtained by simply converting the BEV image to text as illustrated in Figure 2.

Task	Image resolutions (W × H)
Large-scale spatial cognition (Ego images)	512×512
Large-scale spatial cognition (BEV images)	512×512
Mental rotation test (MRT)	Varies from 595×541 to 1133×1432 since we crop the redundant white space around images.
Perspective taking test (PTT)	640×480
Water level test (WLT)	Varies from 239×488 to 787×631 since we crop the redundant white space around images.
Minnesota paper form board (MPFB)	480×480 for choice images (i.e., puzzle pieces put together). Varies from 831×578 to 1211×740 for the image containing the puzzle pieces.
Judgement of line orientation (JLO)	512×512 for the input image, 1656×910 for the legend
Selective attention task (SAtt)	512×512 for 3×3 grids, 768×768 for 4×4 grids, and 1024×1024 for 5×5 and 6×6 grids
Maze completion task (MCT)	1100×1100 for 11×11 mazes, 2300×2300 for 23×23 mazes, and 3100×3100 for 31×31 mazes
Corsi block-tapping test (CBTT)	1024×1024
Spatial addition task (SAdd)	300×300 for 3×3 grids, 500×500 for 5×5 grids, 700×700 for 7×7 grids, and 900×900 for 9×9 grids
Cambridge spatial working memory (CSWM)	1024×1024

Table 5: Image resolutions and aspect ratios for images and videos in SPACE.

Prompting frontier models for SPACE tasks: We evaluate frontier models on each of the SPACE tasks using zero-shot prompting. For each task, we design a prompt that provides a detailed description of the task and the expected response format. Below, we provide the prompt templates for each of the SPACE tasks. While the prompts have been formatted for visual display in LaTeX, the content remains the same. We have replaced images and arrays (in some cases) with placeholders for brevity.

Large-scale spatial cognition

- **Direction estimation:** Prompts 1, 2 and 3
- **Distance estimation:** Prompts 4, 5 and 6
- **Map sketching:** Prompts 4, 5 and 6
- **Route retracing:** Prompts 10, 11, 12 and 13
- **Novel shortcuts:** Prompts 14, 15, 16 and 17

Small-scale spatial cognition

- **Mental rotation test:** Prompts 18 and 19
- **Perspective taking test:** Prompts 20 and 21
- **Water level test:** Prompt 22
- **Minnesota Paper Form Board test:** Prompts 23 and 24
- **Judgement of Line Orientation test:** Prompts 25 and 26
- **Selective attention task:** Prompts 27 and 28
- **Maze completion task:** Prompts 29 and 30
- **Corsi block-tapping task:** Prompts 32 and 31
- **Spatial addition task:** Prompts 33 and 34
- **Cambridge spatial working memory test:** Prompts 35 and 36

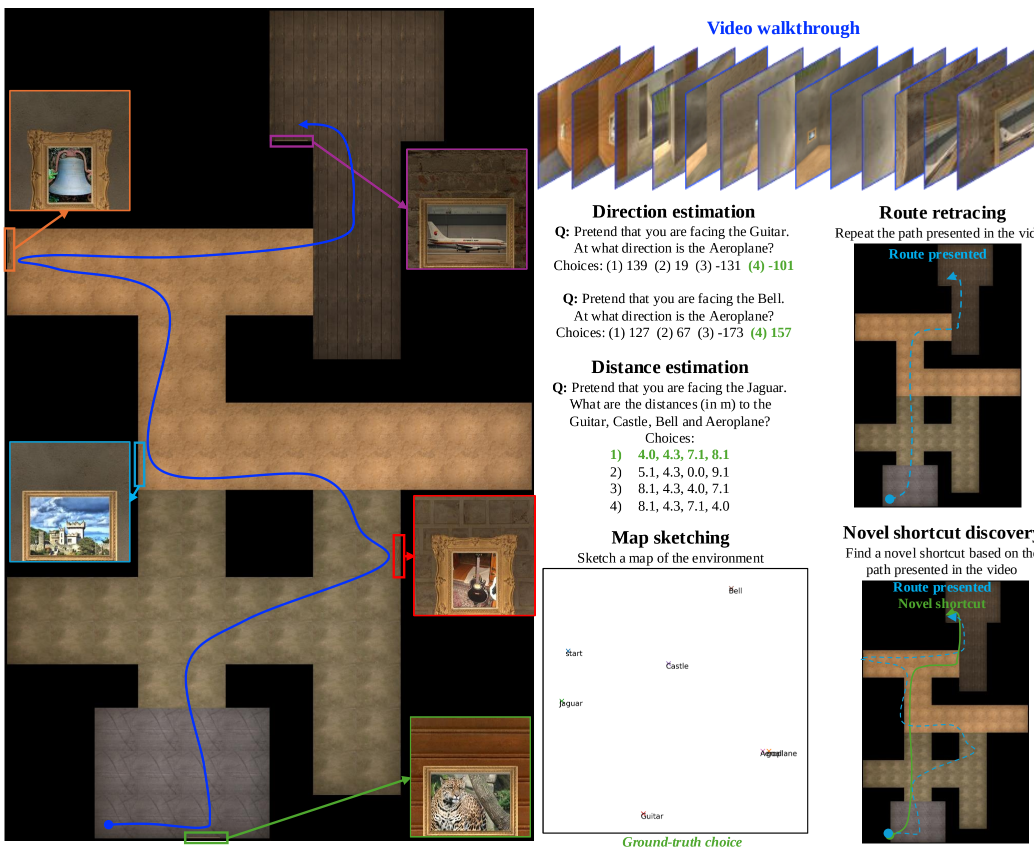

Figure 4: **Large-scale spatial cognition with ego image observations**

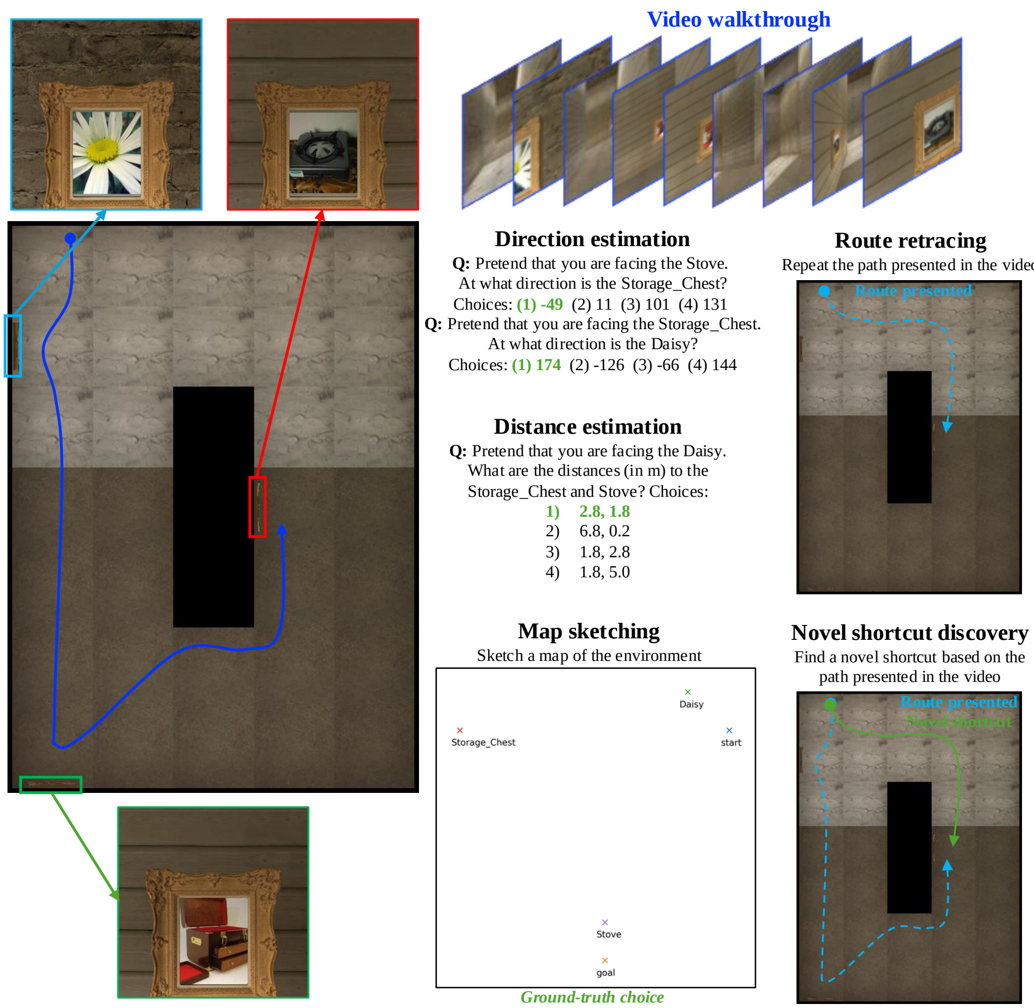

Figure 5: **Large-scale spatial cognition with ego image observations**

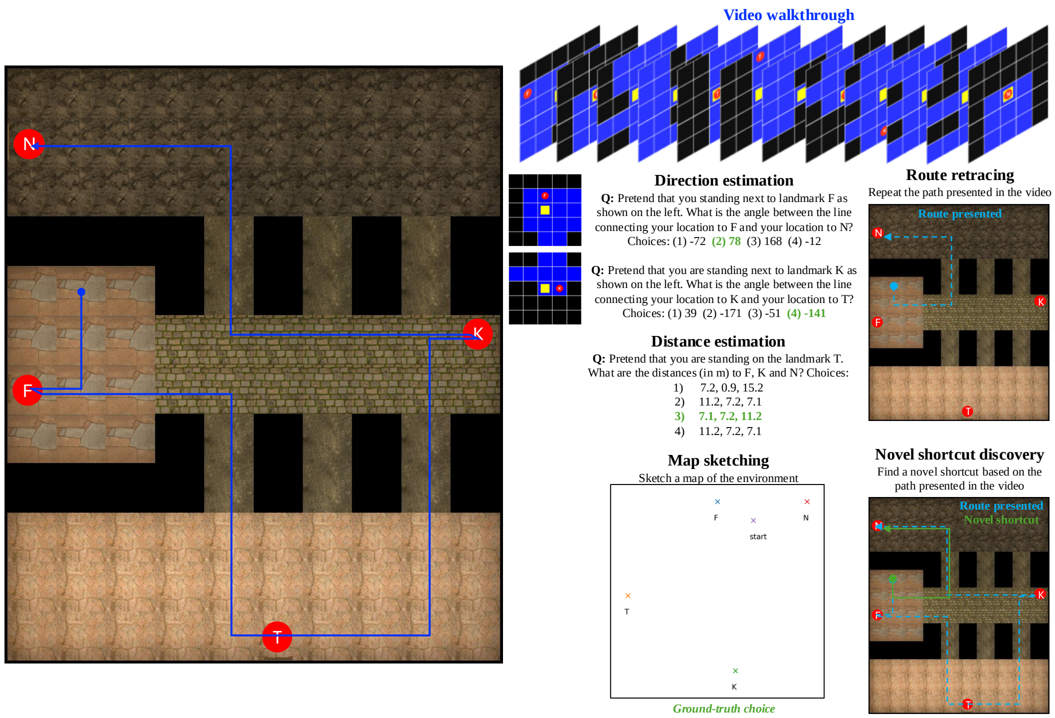

Figure 6: **Large-scale spatial cognition with BEV image observations.** Please note that the top-down visualization on the left needs to be rotated by 90° clockwise to get the BEV images.

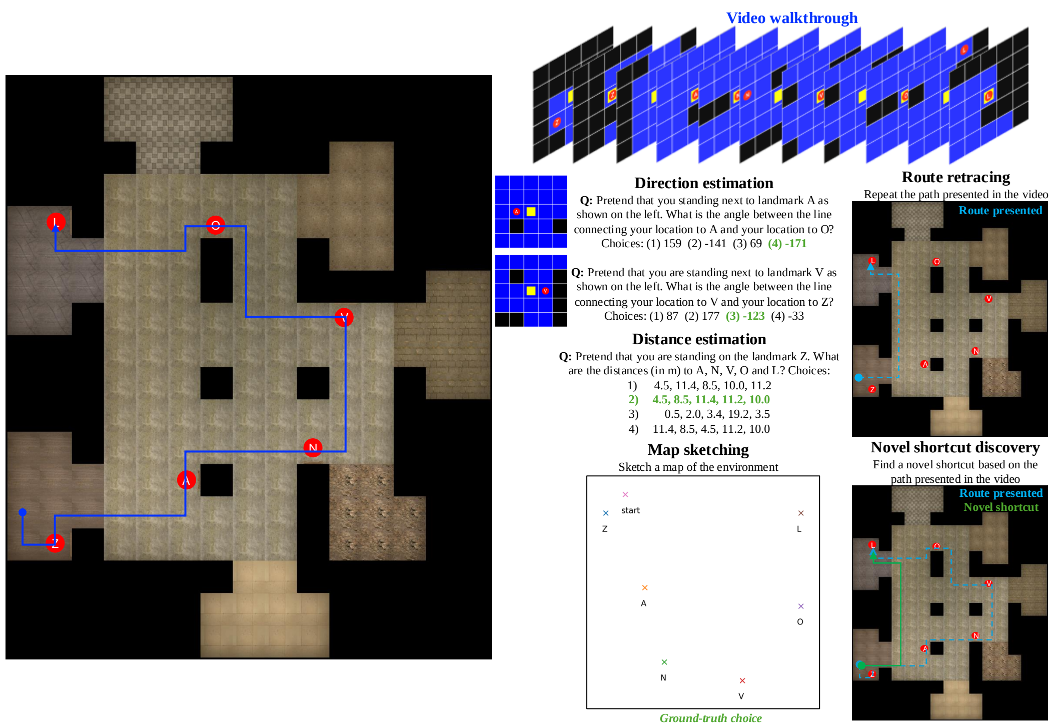

Figure 7: **Large-scale spatial cognition with BEV image observations.** Please note that the top-down visualization on the left needs to be rotated by 90° clockwise to get the BEV images.

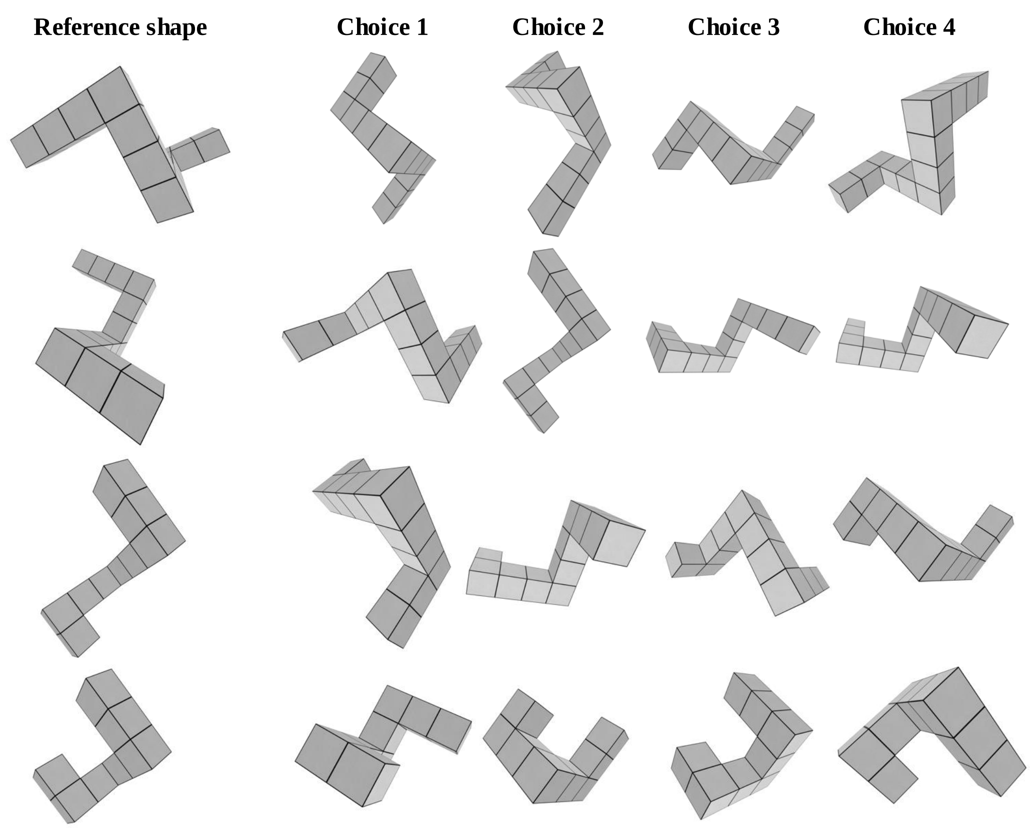

Figure 8: **Mental rotation task (MRT) with visual inputs:** Given a reference shape, find the choice that shows the same object rotated in 3D.

Reference array	Choice 1	Choice 2	Choice 3	Choice 4
0,0,8,1,8,0,0 0,0,2,0,1,0,0 0,0,7,0,7,0,4 9,0,9,5,0,0,0 0,0,0,7,2,0,1 0,0,5,1,0,2,0 0,2,3,0,0,0,7	0,0,8,1,8,0,0 0,0,1,0,2,0,0 4,0,7,0,7,0,0 0,0,0,5,9,0,9 1,0,2,7,0,0,0 0,2,0,1,5,0,0 7,0,0,0,3,2,0	7,0,1,0,4,0,0 0,2,0,0,0,0,0 0,0,2,0,7,1,8 0,1,7,5,0,0,1 3,5,0,9,7,2,8 2,0,0,0,0,0,0 0,0,0,9,0,0,0	0,0,0,9,0,0,0 2,0,0,0,0,0,0 3,5,0,9,7,2,8 0,1,7,5,0,0,1 0,0,2,0,7,1,8 0,2,0,0,0,0,0 7,0,1,0,4,0,0	0,0,0,9,0,0,0 0,0,0,0,0,0,2 8,2,7,9,0,5,3 1,0,0,5,7,1,0 8,1,7,0,2,0,0 0,0,0,0,0,2,0 0,0,4,0,1,0,7
view,none,back used,year,none view,been,none	view,used,view been,year,none none,none,back	back,none,view none,year,used none,been,view	view,been,none used,year,none view,none,back	view,used,view none,year,been back,none,none
0,0,0,0,9,1,9,0,0 0,0,0,8,2,0,3,0,0 3,0,0,0,1,0,0,0,3 9,0,9,1,0,0,0,0,0 2,0,2,0,3,0,0,0,0 4,0,1,0,0,0,0,9,0 7,0,0,0,0,0,2,2,0 0,4,0,0,0,9,4,0,0 0,5,0,3,0,0,0,0,6	0,5,0,3,0,0,0,0,6 0,4,0,0,0,9,4,0,0 7,0,0,0,0,0,2,2,0 4,0,1,0,0,0,0,9,0 2,0,2,0,3,0,0,0,0 9,0,9,1,0,0,0,0,0 3,0,0,0,1,0,0,0,3 0,0,0,8,2,0,3,0,0 0,0,0,0,9,1,9,0,0	0,0,3,9,2,4,7,0,0 0,0,0,0,0,0,0,4,5 0,0,0,9,2,1,0,0,0 0,8,0,1,0,0,0,0,3 9,2,1,0,3,0,0,0,0 1,0,0,0,0,0,0,9,0 9,3,0,0,0,0,2,4,0 0,0,0,0,0,9,2,0,0 0,0,3,0,0,0,0,0,6	0,0,7,4,2,9,3,0,0 5,4,0,0,0,0,0,0,0 0,0,0,1,2,9,0,0,0 3,0,0,0,0,1,0,8,0 0,0,0,0,3,0,1,2,9 0,9,0,0,0,0,0,0,1 0,4,2,0,0,0,0,3,9 0,0,2,9,0,0,0,0,0 6,0,0,0,0,0,3,0,0	0,0,9,1,9,0,0,0,0 0,0,3,0,2,8,0,0,0 3,0,0,0,1,0,0,0,3 0,0,0,0,0,1,9,0,9 0,0,0,0,3,0,2,0,2 0,9,0,0,0,0,1,0,4 0,2,2,0,0,0,0,0,7 0,0,4,9,0,0,0,4,0 6,0,0,0,0,3,0,5,0
0,0,0,3,9,0,8 0,0,3,0,8,0,5 0,4,1,0,0,0,2 0,0,0,0,3,4,0 0,0,0,9,0,1,0 0,0,0,4,0,1,5 0,0,1,8,0,6,0	0,0,0,0,0,0,0 0,0,4,0,0,0,0 0,3,1,0,0,0,1 3,0,0,0,9,4,8 9,8,0,3,0,0,0 0,0,0,4,1,1,6 8,5,2,0,0,5,0	8,0,9,3,0,0,0 5,0,8,0,3,0,0 2,0,0,0,1,4,0 0,4,3,0,0,0,0 0,1,0,9,0,0,0 5,1,0,4,0,0,0 0,6,0,8,1,0,0	0,5,0,0,2,5,8 6,1,1,4,0,0,0 0,0,0,3,0,8,9 8,4,9,0,0,0,3 1,0,0,0,1,3,0 0,0,0,0,4,0,0 0,0,0,0,0,0,0	0,0,0,0,0,0,0 0,0,0,0,4,0,0 1,0,0,0,1,3,0 8,4,9,0,0,0,3 0,0,0,3,0,8,9 6,1,1,4,0,0,0 0,5,0,0,2,5,8

Figure 9: **Mental rotation task (MRT) with text inputs:** Given a reference array, find the choice that shows the same array rotated in 2D.

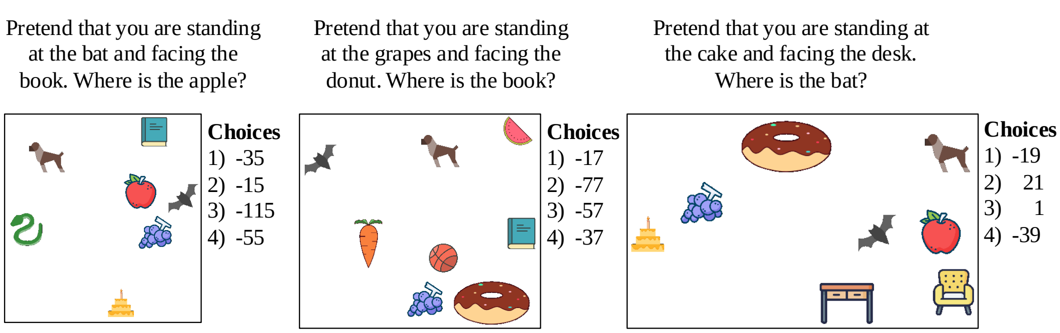

Figure 10: **Perspective taking task (PTT) with visual inputs**.

Pretend that you are standing at 6 and facing 9. Where is 1?

```
0,0,0,8,0,0,0,0
0,0,4,7,0,0,0,0
0,0,0,0,0,0,0,0
0,0,6,2,3,0,0,0
9,0,0,0,0,0,0,0
0,0,0,0,0,1,0,0
0,0,0,0,0,0,0,5
0,0,0,0,0,0,0,0
```

Choices
1) -119
2) -59
3) -99
4) -159

Pretend that you are standing at 7 and facing 2. Where is 1?

```
0,3,0,0
1,6,0,0
0,7,2,0
0,4,8,0
```

Choices
1) -115
2) 165
3) -135
4) -75

Pretend that you are standing at 4 and facing 9. Where is 6?

```
0,0,0,0,0,0,0
0,0,0,0,0,0,0
0,0,0,1,0,6,0
7,0,0,0,0,0,0
2,0,5,0,0,4,8
0,3,0,9,0,0,0
0,0,0,0,0,0,0
```

Choices
1) 136
2) 76
3) 116
4) 156

Pretend that you are standing at 9 and facing 4. Where is 2?

```
0,9,0,5
0,3,0,0
0,0,8,0
4,7,2,0
```

Choices
1) -96
2) -76
3) -36
4) 4

Figure 11: **Perspective taking task (PTT) with text inputs**. *The array colors are only for illustration purposes.*

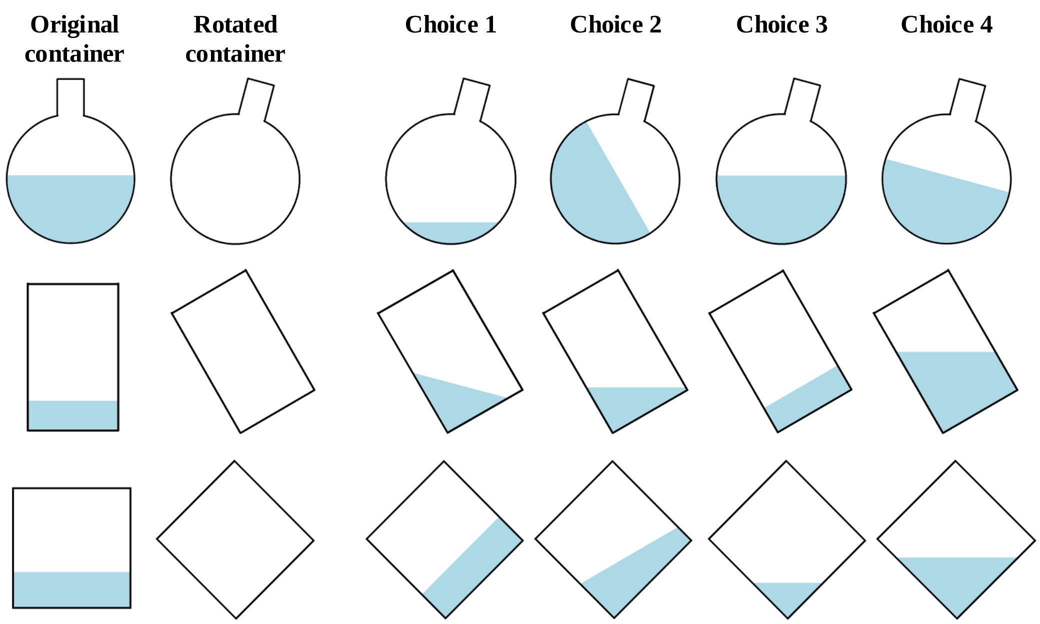

Figure 12: **Water level test (WLT) with vision inputs:** Given a water container filled with water, predict the water level in the rotated container.

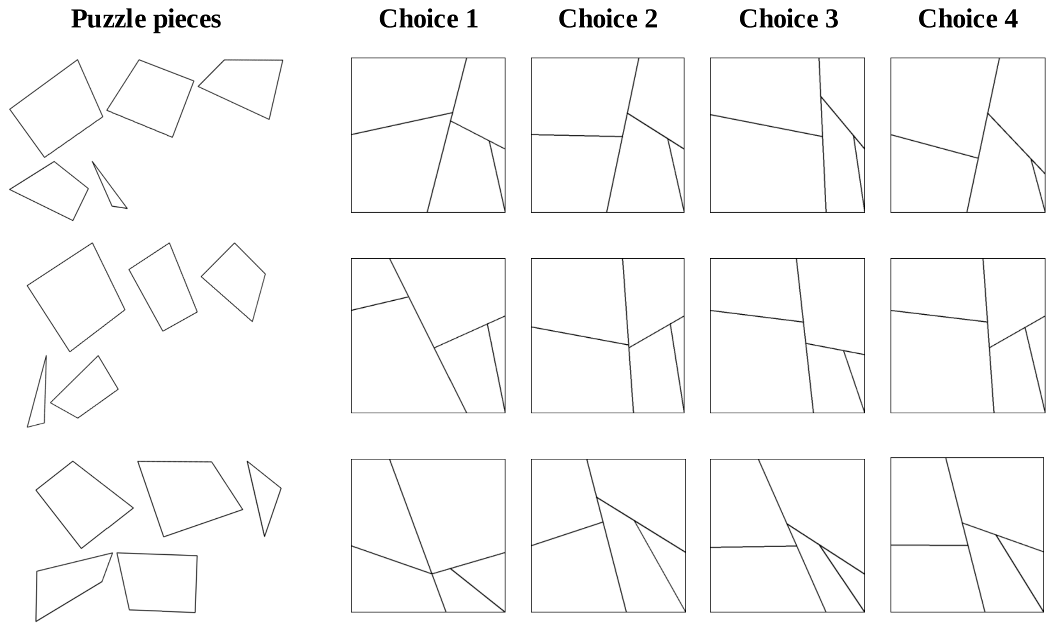

Figure 13: **Minnesota Paper Form Board test (MPFB) with visual inputs:** Which one of the four choices shows what it would be like when the puzzle pieces are put together?

Array pieces | **Choice 1** | **Choice 2** | **Choice 3** | **Choice 4**

```
1,1,1,1  1,1,1,1,1,1,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,1,1,1  1,1,1,1,1,1,1

1,1,1,1,1,1,1  1,1,1,1
1,0,0,0,0,0,1  1,0,0,1
1,0,0,0,0,0,1  1,0,0,1
1,1,1,1,1,1,1  1,1,1,1
```

```
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
1,0,0,0,0,0,1,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,0,0,0,0,1,0,1
1,0,0,0,0,0,1,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,0,0,0,0,1,0,1
1,1,1,1,1,1,1,1,1,1   1,0,0,0,0,0,1,0,0,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
1,0,0,1,0,0,0,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1
1,0,0,1,0,0,0,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1
1,0,0,1,0,0,0,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1
1,0,0,1,0,0,0,0,0,1   1,1,1,1,1,1,1,1,1,1   1,0,0,1,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1
1,0,0,1,0,0,0,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
```

```
1,1,1,1    1,1,1,1,1,1
1,0,0,1    1,0,0,0,0,1
1,0,0,1    1,0,0,0,0,1
1,0,0,1    1,0,0,0,0,1
1,0,0,1    1,0,0,0,0,1
1,1,1,1    1,0,0,0,0,1
           1,1,1,1,1,1
1,1,1,1,1
1,0,0,0,1    1,1,1,1,1
1,0,0,0,1    1,0,0,0,1
1,0,0,0,1    1,0,0,0,1
1,0,0,0,1    1,0,0,0,1
1,1,1,1,1    1,1,1,1,1
```

```
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
1,0,1,0,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,1,1,1,1,1,1,1,1,1   1,0,1,0,0,0,0,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,0,1,0,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,0,1,0,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,0,1,0,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1   1,1,1,1,1,1,1,1,1,1   1,0,1,0,0,0,0,0,0,1
1,0,1,0,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,0,1,0,0,0,0,0,0,1   1,1,1,1,1,1,1,1,1,1   1,0,0,0,1,0,0,0,0,1   1,1,1,1,1,1,1,1,1,1
1,0,1,0,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,0,1,0,0,0,0,1   1,0,0,0,0,0,0,1,0,1
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
```

```
1,1,1,1  1,1,1,1,1,1,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,1,1,1,1,1,1
1,0,0,1
1,0,0,1  1,1,1,1,1,1,1
1,1,1,1  1,0,0,0,0,0,1
         1,0,0,0,0,0,1
1,1,1,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,0,0,1  1,0,0,0,0,0,1
1,1,1,1  1,1,1,1,1,1,1
```

```
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
1,0,0,0,0,0,1,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,0,0,0,1,0,0,1
1,0,0,0,0,0,1,0,0,1   1,0,1,0,0,0,0,0,0,1   1,0,0,0,0,0,1,0,0,1   1,0,0,0,0,0,1,0,0,1
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,0,0,0,0,0,1,0,0,1
1,0,0,0,0,1,0,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,1,0,0,0,0,0,1   1,0,0,0,0,0,1,0,0,1
1,0,0,0,0,1,0,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,1,0,0,0,0,0,1   1,1,1,1,1,1,1,1,1,1
1,0,0,0,0,1,0,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,0,0,0,0,1,0,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,0,0,0,0,1,0,0,0,1   1,0,0,0,0,0,0,1,0,1   1,0,0,1,0,0,0,0,0,1   1,0,1,0,0,0,0,0,0,1
1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1   1,1,1,1,1,1,1,1,1,1
```

Figure 14: **Minnesota Paper Form Board test (MPFB) with text inputs:** Which one of the four choices shows what it would be like when the array pieces are put together? *The array colors are purely for illustration purposes.*

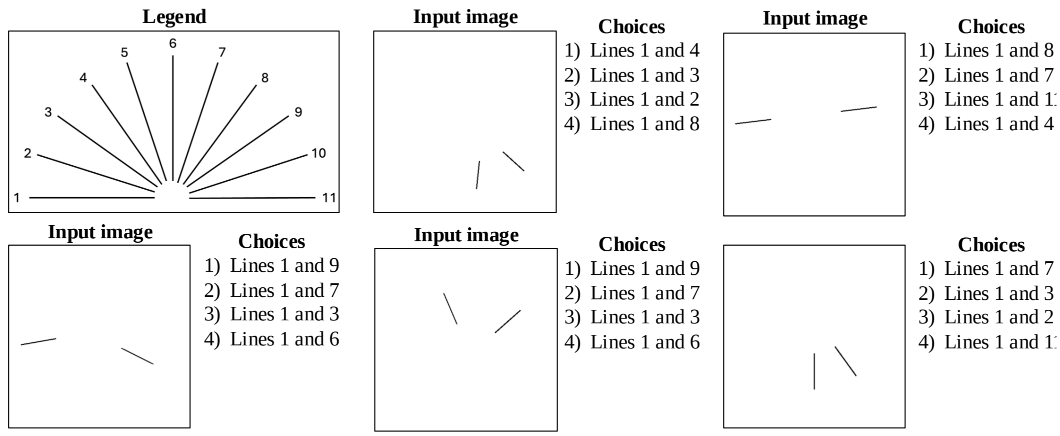

Figure 15: **Judgement of Line Orientation test (JLO) with visual inputs:** Which pair of lines from the legend have a matching angle to the lines in the input image?

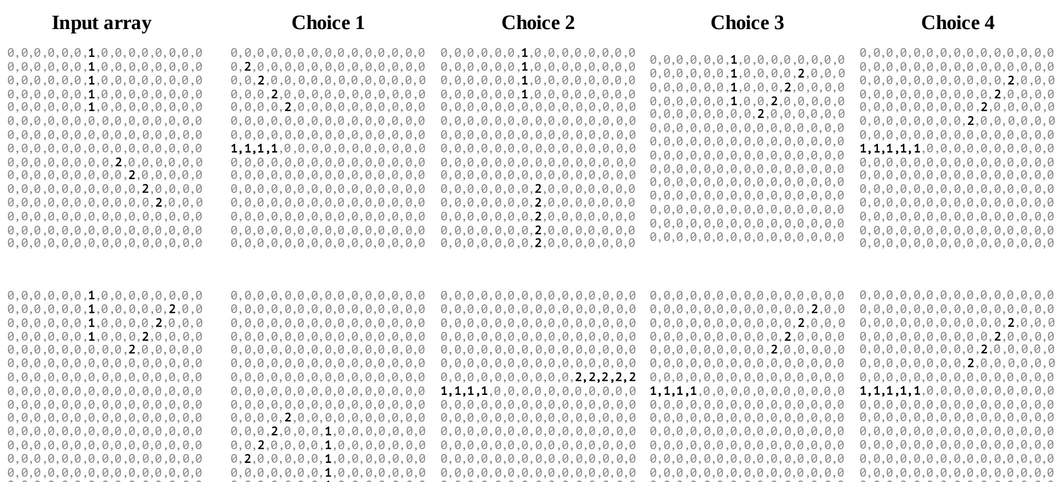

Figure 16: **Judgement of Line Orientation test (JLO) with text inputs:** Which choice has an angle between lines 1 and 2 that matches the angle from the input array? *The array colors are purely for illustration purposes.*

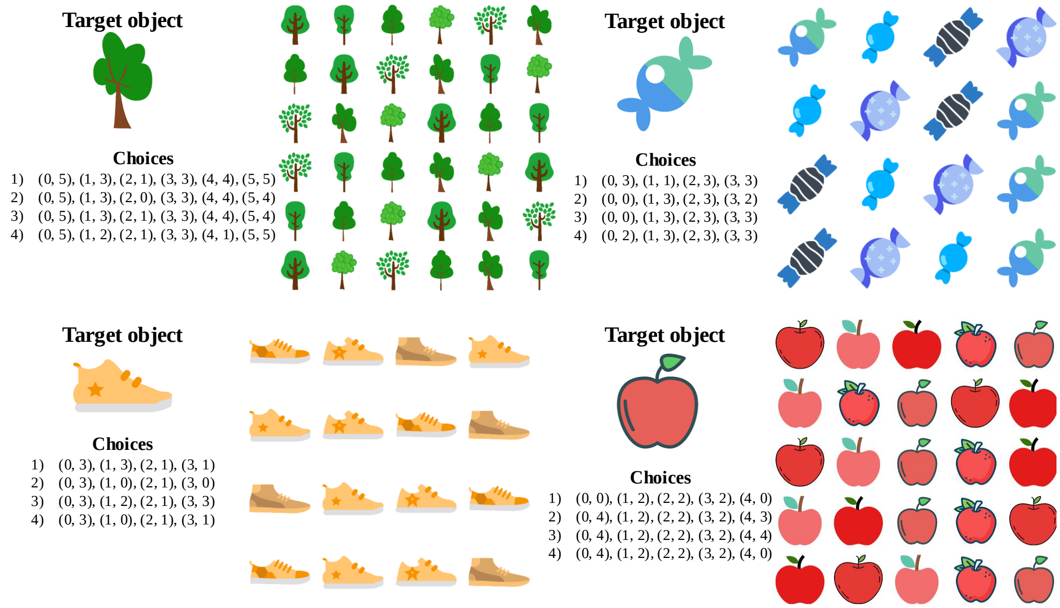

Figure 17: **Spatial attention task (SAtt) with visual inputs:** Which grid locations contain the target object? Locations are represented as (row, column). Top-left of the grid is (0, 0).

Target character: b

```
l,k,h,s,p,z,x,b,n
b,p,x,s,z,n,h,l,k
n,p,x,s,k,h,l,z,b
k,b,s,x,p,l,z,h,n
x,l,s,z,p,b,k,n,h
p,k,h,n,l,b,x,s,z
n,h,l,z,k,x,p,s,b
x,b,n,l,p,h,k,z,s
h,x,n,k,p,l,b,z,s
```

Choices

1) (0, 7), (1, 0), (2, 8), (3, 1), (4, 5), (5, 5), (6, 8), (7, 1), (8, 2)
2) (0, 7), (1, 0), (2, 8), (3, 1), (4, 5), (5, 5), (6, 8), (7, 1), (8, 6)
3) (0, 3), (1, 1), (2, 0), (3, 1), (4, 5), (5, 6), (6, 3), (7, 5), (8, 0)
4) (0, 7), (1, 0), (2, 5), (3, 1), (4, 5), (5, 5), (6, 8), (7, 1), (8, 6)

Target character: 5

```
7,0,6,5,4
0,7,6,5,4
6,7,5,4,0
7,4,5,6,0
0,6,7,4,5
```

Choices

1) (0, 3), (1, 3), (2, 2), (3, 2), (4, 2)
2) (0, 2), (1, 2), (2, 2), (3, 0), (4, 1)
3) (0, 3), (1, 3), (2, 1), (3, 2), (4, 4)
4) (0, 3), (1, 3), (2, 2), (3, 2), (4, 4)

Target character: k

```
l,c,k,d,z
l,d,z,c,k
d,c,l,z,k
c,z,k,d,l
d,l,k,c,z
```

Choices

1) (0, 2), (1, 4), (2, 4), (3, 2), (4, 2)
2) (0, 1), (1, 1), (2, 4), (3, 2), (4, 1)
3) (0, 2), (1, 4), (2, 4), (3, 2), (4, 4)
4) (0, 2), (1, 4), (2, 4), (3, 1), (4, 2)

Target character: 9

```
9,3,5,2,4,7,6
5,4,2,9,7,6,3
5,7,6,9,2,4,3
3,4,7,2,5,6,9
3,6,7,4,2,9,5
2,6,9,4,7,5,3
3,5,2,4,6,9,7
```

Choices

1) (0, 3), (1, 3), (2, 3), (3, 6), (4, 5), (5, 3), (6, 2)
2) (0, 0), (1, 3), (2, 3), (3, 6), (4, 5), (5, 2), (6, 5)
3) (0, 0), (1, 3), (2, 3), (3, 6), (4, 5), (5, 2), (6, 3)
4) (0, 0), (1, 1), (2, 3), (3, 6), (4, 5), (5, 2), (6, 5)

Figure 18: **Spatial attention task (SAtt) with text inputs:** Which grid locations contain the target character? Locations are represented as (row, column). Top-left of the grid is (0, 0).

Figure 19: **Corsi block-tapping task (CBTT) with visual inputs:** What is the sequence of block taps observed in the image?

Sequence of taps ("T") **Reference numbers**

```
0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0
0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0
0,0,B,0,0,B,0   0,0,B,0,0,B,0   0,0,T,0,0,B,0   0,0,B,0,0,B,0   0,0,B,0,0,B,0   0,0,6,0,0,1,0
0,0,0,0,0,B,0   0,0,0,0,0,B,0   0,0,0,0,0,B,0   0,0,0,0,0,T,0   0,0,0,0,0,B,0   0,0,0,0,0,5,0
0,0,0,0,0,0,B   0,0,0,0,0,0,T   0,0,0,0,0,0,B   0,0,0,0,0,0,B   0,0,0,0,0,0,B   0,0,0,0,0,0,2
0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0
B,T,0,0,0,0,0   B,B,0,0,0,0,0   B,B,0,0,0,0,0   B,B,0,0,0,0,0   T,B,0,0,0,0,0   3,4,0,0,0,0,0
```

Choices
(1) 2, 3, 6, 5, 4
(2) 4, 2, 6, 3, 5
(3) 1, 3, 4, 5, 2
(4) 4, 2, 6, 5, 3

```
0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0
0,B,0,0,0,0,0   0,B,0,0,0,0,0   0,B,0,0,0,0,0   0,B,0,0,0,0,0   0,B,0,0,0,0,0   0,3,0,0,0,0,0
0,0,0,0,0,B,0   0,0,0,0,0,B,0   0,0,0,0,0,T,0   0,0,0,0,0,B,0   0,0,0,0,0,B,0   0,0,0,0,0,4,0
0,B,0,0,0,0,B   0,B,0,0,0,0,B   0,B,0,0,0,0,B   0,B,0,0,0,0,T   0,T,0,0,0,0,B   0,2,0,0,0,0,1
0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0
T,0,0,B,0,0,0   B,0,0,T,0,0,0   B,0,0,B,0,0,0   B,0,0,B,0,0,0   B,0,0,B,0,0,0   5,0,0,6,0,0,0
0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0   0,0,0,0,0,0,0
```

Choices
(1) 2, 1, 5, 4, 6
(2) 5, 6, 4, 1, 2
(3) 5, 6, 4, 2, 1
(4) 2, 4, 1, 3, 5

```
0,B,0,0,0   0,T,0,0,0   0,B,0,0,0   0,B,0,0,0   0,1,0,0,0
B,0,0,0,0   B,0,0,0,0   B,0,0,0,0   B,0,0,0,0   5,0,0,0,0
0,0,0,B,B   0,0,0,B,B   0,0,0,T,B   0,0,0,B,T   0,0,0,7,6
0,T,B,0,0   0,B,B,0,0   0,B,B,0,0   0,B,B,0,0   0,2,4,0,0
0,0,0,0,B   0,0,0,0,B   0,0,0,0,B   0,0,0,0,B   0,0,0,0,3
```

Choices
(1) 2, 1, 7, 6
(2) 2, 7, 6, 1
(3) 2, 1, 6, 7
(4) 1, 3, 7, 6

Figure 20: **Corsi block-tapping task (CBTT) with text inputs:** What is the sequence of block taps observed in the arrays? *The array colors are purely for illustration purposes.*

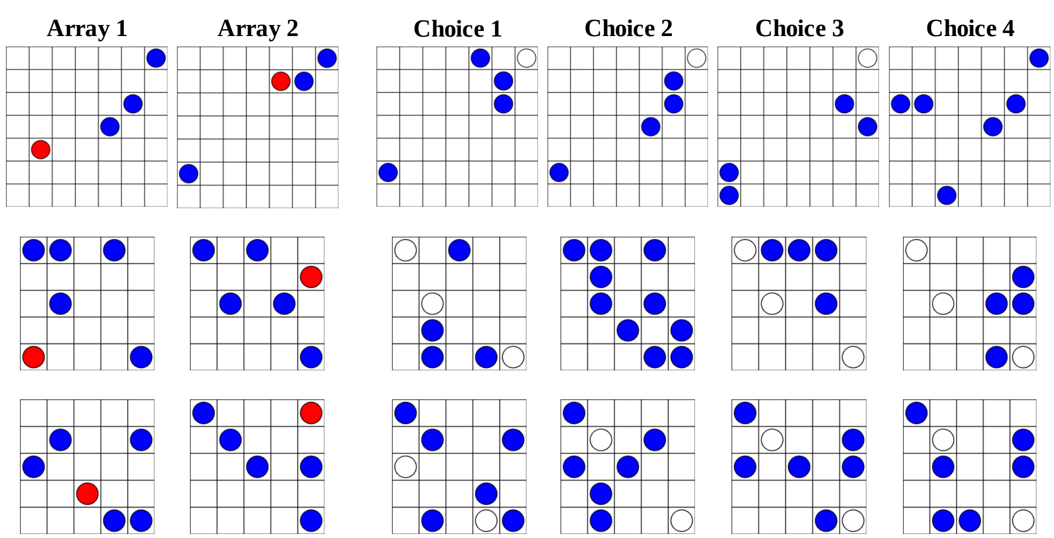

Figure 21: **Spatial addition task (SAdd) with visual inputs:** What is the sum of the two arrays? Ignore red circles. Blue circles represent 1, white circles represent 2 and empty spaces represent 0.

Array 1	Array 2	Choice 1	Choice 2	Choice 3	Choice 4
E,B,E,R,E	E,B,E,E,E	E,W,E,E,E	E,B,E,E,E	E,W,E,E,E	E,W,E,E,E
E,E,E,E,E	E,B,E,E,E	E,B,E,E,E	E,E,E,E,E	E,E,E,E,B	E,B,E,E,E
E,E,E,E,E	E,E,E,E,E	E,E,E,E,E	E,E,E,E,E	E,E,E,B,E	E,E,E,E,E
E,E,B,E,E	E,R,E,E,E	E,E,E,B,E	E,E,W,E,E	E,E,E,E,E	E,E,B,E,E
E,E,E,E,E	E,E,E,E,E	E,E,E,E,E	E,E,E,E,B	E,E,E,E,E	E,E,E,E,E
B,E,E	B,E,E	W,E,E	B,E,E	B,E,E	B,E,E
E,E,R	E,E,E	E,E,E	E,E,B	E,E,E	E,B,E
E,E,E	E,E,R	E,E,E	E,E,E	B,E,E	E,E,E
R,E,E,B,E,E,E,E,B	E,E,E,B,E,E,E,B,E	E,E,E,W,E,E,E,B,B	E,E,E,W,E,E,E,B,E	B,E,E,W,E,B,E,B,B	E,B,E,B,E,B,E,E,B
E,E,E,E,E,E,E,B,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,B,E	E,E,E,E,E,E,E,B,E	E,E,E,E,E,E,E,B,E	E,E,B,E,E,E,E,B,E
E,E,E,E,B,E,E,E,E	E,E,E,B,B,E,E,E,E	E,E,E,B,W,E,E,E,E	E,E,E,B,W,B,E,E,E	E,E,E,B,W,E,E,E,E	E,E,E,E,B,E,E,E,E
E,E,E,E,B,E,E,B,E	E,E,B,E,B,E,E,B,E	E,E,B,E,W,E,E,W,E	E,E,B,E,W,E,E,W,E	E,E,B,E,W,E,E,W,E	B,E,E,E,B,E,E,B,E
E,B,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,B,E,E,E,E,E,E,E	E,B,E,E,E,E,E,E,E	E,B,E,E,E,E,E,E,E	E,W,E,B,E,E,E,E,E
E,E,E,E,E,E,E,E,E	E,E,E,E,R,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E
E,E,E,B,E,E,E,E,B	E,E,E,E,E,E,B,E,E	E,E,E,B,E,E,B,E,B	E,E,E,B,E,E,E,B,B	E,E,E,B,E,E,B,E,E	E,E,E,B,E,E,E,E,B
E,E,E,E,E,E,E,E,E	E,E,E,E,B,E,E,E,E	E,E,E,E,B,E,E,E,E	E,E,E,E,E,E,B,E,E	E,E,E,E,E,E,E,E,E	E,B,E,E,E,E,E,E,E
E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,E,E,E,E,E	E,E,E,E,B,E,E,B,E

Figure 22: **Spatial addition task (SAdd) with text inputs:** What is the sum of the two arrays? Ignore “R”. “E” is 0, “B” is 1 and “W” is 2. *The array colors are purely for illustration purposes.*

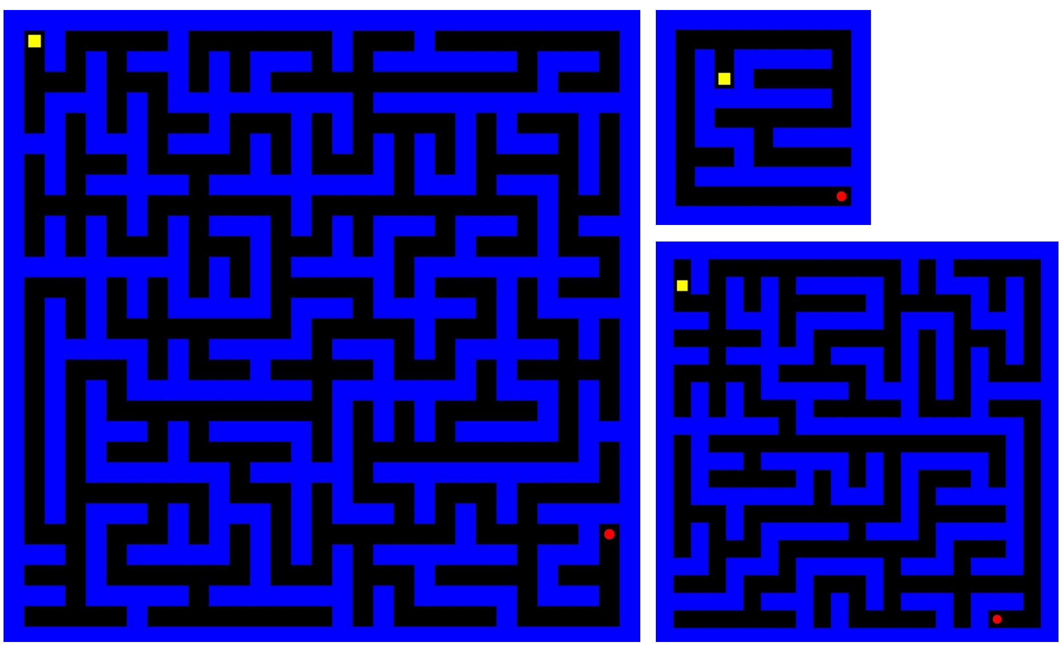

Figure 23: **Maze completion task (MCT) with visual inputs:** We illustrate examples of mazes used for the MCT task. We programmatically generate mazes of different sizes using Mazelib (Stilley, 2014). Blue cells are obstacles. Black cells are navigable space. The yellow square represents the current location. The red circle represents the goal location.

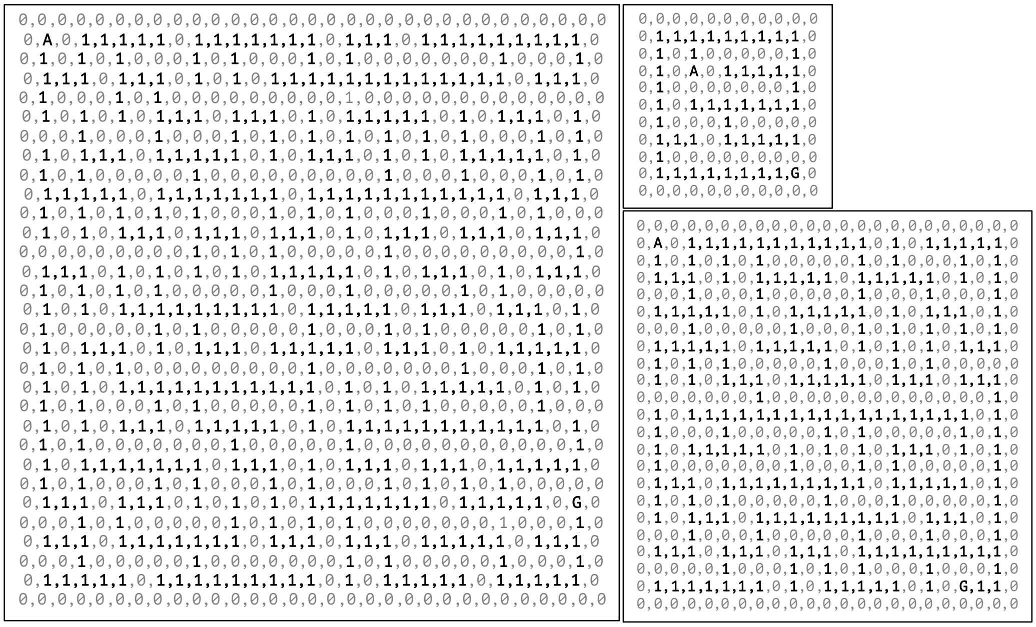

Figure 24: **Maze completion task (MCT) with text inputs:** We illustrate examples of mazes used for the MCT task. We programmatically generate mazes of different sizes using Mazelib (Stilley, 2014). 0s are obstacles. 1s are navigable space. A represents the current location. G represents the goal location. *The array colors are purely for illustration purposes.*

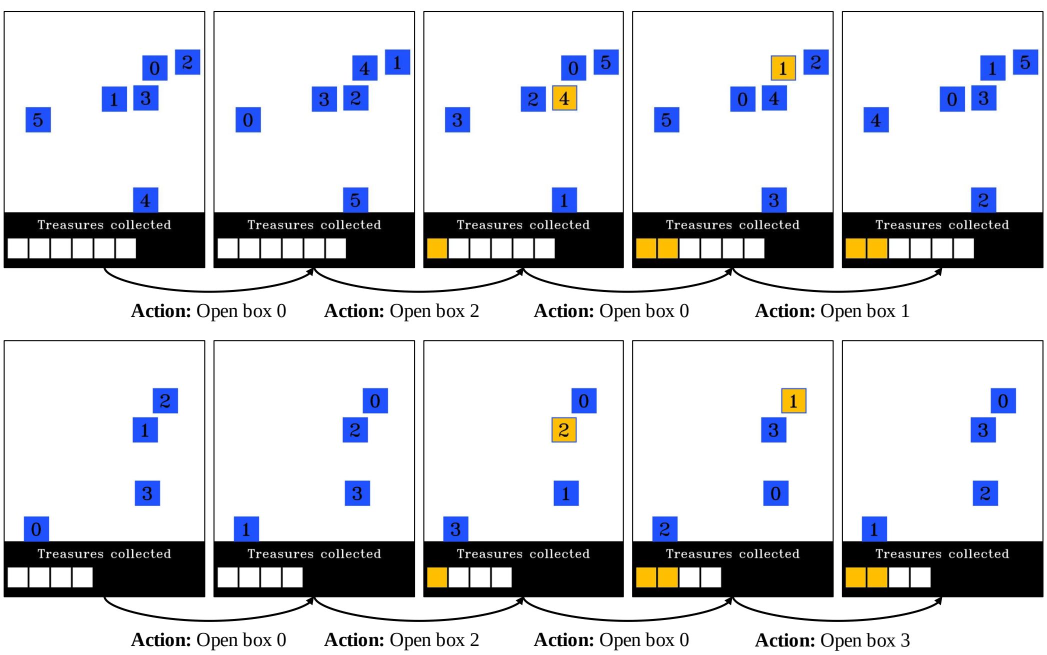

Figure 25: **Cambridge spatial working memory task (CSWM) with visual inputs:** We illustrate two game plays of the CSWM task in the two rows. In each row, we show the initial observation followed by actions taken and the resulting observations. Note how the box identities change after each step. This is intended to force models to remember boxes by their spatial locations instead of their integer identities. As treasures get collected, they are populated in the "Treasures collected" section of the game screen. When a treasure is collected, a new treasure is placed in one of the boxes where the treasure never appeared before.

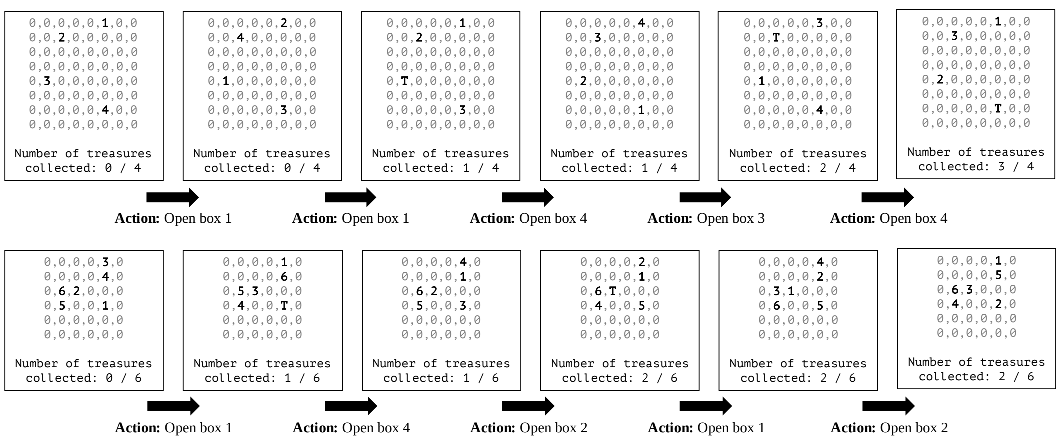

Figure 26: **Cambridge spatial working memory task (CSWM) with text inputs:** We illustrate two game plays of the CSWM task in the two rows. In each row, we show the initial observation (provided as text arrays) followed by actions taken and the resulting observations. The boxes in each array are the non-zero elements. Note how the box identities change after each step. This is intended to force models to remember boxes by their spatial locations instead of their integer identities. As treasures get collected, the "Number of treasures collected" gets incremented. When a treasure is collected, a new treasure is placed in one of the boxes where the treasure never appeared before. *The array colors are purely for illustration purposes.*

Prompt 1: Direction estimation (Ego image)

USER: You are a sentient AI system capable of visually understanding the physical world, performing spatial reasoning, remembering landmarks in the world and navigating in it. Here is a video taken in a physical enviornment as you were navigating in it. Understand the environment you are navigating in, build a map of the environment to keep track of your position as well as locations of landmarks in the environment. Landmarks are paintings of objects hung on the walls.

```
<IMAGE 1>
<IMAGE 2>
.
.
.
```

Pretend that you are standing facing the painting of a Shopping_Cart. See image below.

```
<IMAGE OF Shopping_Cart>
```

At what direction (in angles from -180 to 180 degrees) is the painting of a Lawn_Mower relative to you? See image below.

```
<IMAGE OF Lawn_Mower>
```

Here are your choices: 1) -39 2) 111 3) 81 4) -69

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```

Prompt 2: Direction estimation (BEV image)

USER: You are playing a game in a 2D world. Each image shows the immediate surroundings around you, with you at the center of the image (in yellow). The black cells are obstacles, i.e., you cannot move over them. The blue cells are navigable spaces that you can move over. Some blue cells have landmarks in them (red circles with a text label). These are important to remember. Here is a video taken in the 2D world as you were navigating in it. Understand the 2D world you are navigating in, build a map of the world to keep track of your position as well as locations of landmarks in the world.

```
<IMAGE 1>
<IMAGE 2>
.
.
.
```

You must now answer a question based on your understanding of the 2D world. Pretend that you are standing next to the landmark A. See image below.

```
<IMAGE OF A>
```

What is the angle between the line connecting your current location to the landmark A and the line connecting your current location to the landmark C? Angles range from -180 to 180 degrees. Here are your choices: 1) 156 2) 96 3) 66 4) -24 Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```

Large-scale spatial cognition

		Direction est.	Distance est.	Map sketching	Route retracing	Novel shortcuts
Ego image	# questions	150	135	30	30	30
	# videos	30	30	30	30	30
BEV image	# questions	150	135	30	30	30
	# videos	30	30	30	30	30
BEV text	# questions	150	135	30	30	30

Small-scale spatial cognition

		MRT	PTT	WLT	MPFB	JLO	SAtt	MCT	CBTT	SAdd	CSWM
Visual	# questions	172	100	50	50	50	100	45	50	50	50
	# images	139	20	300	250	51	200	-	297	300	-
Textual	# questions	40	100	-	50	50	100	45	50	50	50

Table 6: **SPACE benchmark statistics:** We show the number of questions, images, and videos for each SPACE task. For large-scale spatial cognition tasks, we have one video per environment. We generate questions and navigation tasks based on these videos. Some small-scale spatial cognition tasks have multiple images for the same question (e.g., MPFB, WLT, SAtt and CBTT), while other tasks have multiple questions for the same image (e.g., PTT, MRT). For interactive tasks like MCT, CSWM, route retracing and novel shortcuts, images are rendered conditioned on the actions taken by the agent.

Prompt 3: Direction estimation (BEV text)

USER: You are playing a game in a 2D text world. The console of the game is represented as a comma-separated text array. Obstacles are represented using 0, i.e., you cannot move over them. Navigable spaces that you can move over are represented using 1. Some navigable spaces have landmarks represented as an ascii character (A - Z). These are also navigable spaces and are just labeled with an ascii character. These landmarks are important to remember. You will always be located at the center of the array with your position highlighted using the "*" character. Here is a sequence of console screen recordings taken as you were navigating in the 2D text world. Understand the world you are navigating in, build a map of the world to keep track of your position as well as locations of landmarks in the world.

```
========== Start of console screen recordings ==========
Screen at time = 0

0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0


Screen at time = 1

0,0,0,0,0
0,0,0,0,0
0,0,*,1,1
0,0,C,1,1
0,0,0,1,1

.
.
.
========== End of console screen recordings ==========
```

You must now answer a question based on your understanding of the 2D text world. Pretend that you are standing next to the landmark B as shown below.

```
1,1,1,1,1
1,1,1,1,1
1,B,*,1,1
1,0,1,1,0
1,1,1,1,1
```

What is the angle between the line connecting your current location to the landmark B and the line connecting your current location to the landmark C? Angles range from -180 to 180 degrees. Note that you may not see landmark C in your immediate vicinity. You must use spatial knowledge from the sequence of screen recordings to locate your current position and both landmarks to answer this question.

Here are your choices: 1) -127 2) 53 3) 83 4) -97

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```

Prompt 4: Distance estimation (Ego image)

USER: You are a sentient AI system capable of visually understanding the physical world, performing spatial reasoning, remembering landmarks in the world and navigating in it. Here is a video taken in a physical enviornment as you were navigating in it. Understand the environment you are navigating in, build a map of the environment to keep track of your position as well as locations of landmarks in the environment. Landmarks are paintings of objects hung on the walls.

<IMAGE 1>
<IMAGE 2>
.
.
.

Pretend that you are standing facing the painting of a Shopping_Cart. See image below.

<IMAGE OF Shopping_Cart>

What are the euclidean distances (in meters) to each of the following landmarks from your current position?

Landmark: painting of a Guitar

<IMAGE OF Guitar>

Landmark: painting of a Horse_Cart

<IMAGE OF Horse_Cart>

Landmark: painting of a Lawn_Mower

<IMAGE OF Lawn_Mower>

Landmark: painting of a Hammer

<IMAGE OF Hammer>

Landmark: painting of a Soccer_Ball

<IMAGE OF Soccer_Ball>

Here are your choices:
1) 2.2, 4.0, 5.7, 5.3, 5.0
2) 5.7, 4.0, 2.2, 5.3, 5.0
3) 1.7, 2.0, 3.3, 2.2, 1.0
4) 4.0, 5.0, 2.2, 5.7, 5.3

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```
````

Prompt 5: Distance estimation (BEV image)

USER: You are playing a game in a 2D world. Each image shows the immediate surroundings around you, with you at the center of the image (in yellow). The black cells are obstacles, i.e., you cannot move over them. The blue cells are navigable spaces that you can move over. Some blue cells have landmarks in them (red circles with a text label). These are important to remember. Here is a video taken in the 2D world as you were navigating in it. Understand the 2D world you are navigating in, build a map of the world to keep track of your position as well as locations of landmarks in the world.

<IMAGE 1>
<IMAGE 2>
.
.
.

You must now answer a question based on your understanding of the 2D world. Pretend that you are standing on the landmark C. What are the euclidean distances (in meters) from landmark C to each of the following landmarks: B, A, N, O, Y? Assume that each grid square (white borders) is 1m x 1m in size. Here are your choices: 1) 4.5, 8.5, 11.4, 11.2, 10.0 2) 11.4, 8.5, 4.5, 11.2, 10.0 3) 10.0, 8.5, 4.5, 11.4, 11.2 4) 12.5, 6.0, 3.4, -3.5, 7.2

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```
````

Prompt 6: Distance estimation (BEV text)

USER: You are playing a game in a 2D text world. The console of the game is represented as a comma-separated text array. Obstacles are represented using 0, i.e., you cannot move over them. Navigable spaces that you can move over are represented using 1. Some navigable spaces have landmarks represented as an ascii character (A - Z). These are also navigable spaces and are just labeled with an ascii character. These landmarks are important to remember. You will always be located at the center of the array with your position highlighted using the "*" character. Here is a sequence of console screen recordings taken as you were navigating in the 2D text world. Understand the world you are navigating in, build a map of the world to keep track of your position as well as locations of landmarks in the world.

```
========== Start of console screen recordings ==========
Screen at time = 0

0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0


Screen at time = 1

0,0,0,0,0
0,0,0,0,0
0,0,*,1,1
0,0,C,1,1
0,0,0,1,1

.
.
.
========== End of console screen recordings ==========
```

You must now answer a question based on your understanding of the 2D text world. Pretend that you are standing on the landmark C. What are the euclidean distances (in meters) from landmark C to each of the following landmarks: B, A, N, O, Y? Assume that each array location is 1m x 1m in size.

Here are your choices: 1) 4.5, 8.5, 11.4, 11.2, 10.0 2) 10.0, 19.2, 16.5, 12.5, 11.4 3) 11.4, 8.5, 4.5, 11.2, 10.0 4) 11.2, 11.4, 8.5, 4.5, 10.0

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```

Prompt 7: Map sketching (Ego image)

USER: You are a sentient AI system capable of visually understanding the physical world, performing spatial reasoning, remembering landmarks in the world and navigating in it. Here is a video taken in a physical enviornment as you were navigating in it. Understand the environment you are navigating in, build a map of the environment to keep track of your position as well as locations of landmarks in the environment. Landmarks are paintings of objects hung on the walls.

```
<IMAGE 1>
<IMAGE 2>
.
.
.
```

You must sketch a map of the environment with the locations of the start, goal and landmark locations. To refresh your memory, here are the landmarks present in the environment.

Landmark: Soccer_Ball

`<IMAGE OF Soccer_Ball>`

Landmark: Shopping_Cart

`<IMAGE OF Shopping_Cart>`

Landmark: Lawn_Mower

`<IMAGE OF Lawn_Mower>`

Landmark: Horse_Cart

`<IMAGE OF Horse_Cart>`

Landmark: Guitar

`<IMAGE OF Guitar>`

Landmark: Hammer

`<IMAGE OF Hammer>`

Follow these map conventions. Your initial heading direction in the video must be along the Y axis (upward). Which of these map sketches best capture the structure of the environment?

Choice 1

`<SKETCH IMAGE OF Choice 1>`

Choice 2

`<SKETCH IMAGE OF Choice 2>`

Choice 3

`<SKETCH IMAGE OF Choice 3>`

Choice 4

`<SKETCH IMAGE OF Choice 4>`

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```
````

Prompt 8: Map sketching (BEV image)

USER: You are playing a game in a 2D world. Each image shows the immediate surroundings around you, with you at the center of the image (in yellow). The black cells are obstacles, i.e., you cannot move over them. The blue cells are navigable spaces that you can move over. Some blue cells have landmarks in them (red circles with a text label). These are important to remember. Here is a video taken in the 2D world as you were navigating in it. Understand the 2D world you are navigating in, build a map of the world to keep track of your position as well as locations of landmarks in the world.

```
<IMAGE 1>
<IMAGE 2>
.
.
.
```

You must now sketch a map of the environment with the locations of the start and landmark locations. Use your understanding of the 2D world. Which of these map sketches best capture the true structure of the 2D world?

Choice 1

```
<SKETCH IMAGE OF Choice 1>
```

Choice 2

```
<SKETCH IMAGE OF Choice 2>
```

Choice 3

```
<SKETCH IMAGE OF Choice 3>
```

Choice 4

```
<SKETCH IMAGE OF Choice 4>
```

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```
````

Prompt 9: Map sketching (BEV text)

USER: You are playing a game in a 2D text world. The console screen of the game is represented as a comma-separated text array. Obstacles are represented using 0, i.e., you cannot move over them. Navigable spaces that you can move over are represented using 1. Some navigable spaces have landmarks represented as an ascii character (A - Z). These are also navigable spaces and are just labeled with an ascii character. These landmarks are important to remember. You will always be located at the center of the array with your position highlighted using the "*" character. Here is a sequence of console screen recordings taken as you were navigating in the 2D text world. Understand the world you are navigating in, build a map of the world to keep track of your position as well as locations of landmarks in the world.

```
========== Start of console screen recordings ==========
Screen at time = 0

0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0


Screen at time = 1

0,0,0,0,0
0,0,0,0,0
0,0,*,1,1
0,0,C,1,1
0,0,0,1,1

.
.
.
========== End of console screen recordings ==========
```

You must now sketch a map of the environment with the locations of the start and landmark locations. Use your understanding of the 2D text world.

Which of these map sketches best capture the true structure of the 2D world? The map sketches are 2D arrays with markers highlighting the landmarks (ascii characters from A to Z) and the start location (marked as *). Locations in the map sketch with 0s are empty and can be ignored.

Choice 1

```
<TEXT ARRAY OF Choice 1>
```

Choice 2

```
<TEXT ARRAY OF Choice 2>
```

Choice 3

```
<TEXT ARRAY OF Choice 3>
```

Choice 4

```
<TEXT ARRAY OF Choice 4>
```

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer choice value>
}
```

Prompt 10: Route retracing (Ego image)

SYSTEM: You are a sentient living creature capable of navigating in environments, building internal spatial representations of environments, and finding goals in them. You will be shown a video of the shortest route from the initial position to the goal. You must look at the video and understand the environment structure and the route taken. Then, you will be placed in the environment at the same initial position. You must navigate from the initial position to the goal using the same route shown in the video, as quickly as possible. Below, you will find sections highlighting more details about the task. You can refer to these for more information.

OBSERVATIONS:
The images are recorded from a perspective viewpoint (i.e., egocentric or first-person). This means that you are likely to see objects from different angles, resulting in a skewed appearance of the underlying 3D objects. It is important for you to look past this skew in the appearance and percive the true shape of the object in 3D.

GOAL:
You will be provided an object goal using a text description and an image of the object. You must find the goal object in the environment by repeating the path shown in the video walkthrough. Once you find it, move close to the location of the goal and re-orient yourself to face the object.

ACTIONS:
You have four actions available.
move_forward: move forward by 0.25m along the current heading direction. It does not change the heading angle.
turn_left: decrease your heading angle by 30 degrees. It does not change the (x, y) position.
turn_right: increase your heading angle by 30 degrees. It does not change the (x, y) position.
stop: ends the current task. Issue this action only if you think you have reached the goal. If you haven't reached the goal, this action will result in a navigation failure that cannot be recovered from.

STUCK IN PLACE BEHAVIOR:
Avoid getting stuck in one place, i.e., do not alternate between left and right turns without going anywhere. You must try and move around consistently without being stuck in one place.

STOP CRITERIA:
Before executing stop, you must ensure that you've "reached" the goal correctly. To reach a goal, you have to move close enough to the wall where you see the goal, and see the object clearly in your observation in front of you.

RESPONSE FORMAT:
Respond in the following format:
Reasoning: text explanation string in one or two short sentences — provide all your explanations and inner thoughts here - avoid verbosity and be concise
Intent: state your intent in one short sentence, i.e., what you are trying to achieve
Then provide the final action to take in a json formatted string.

```
{
    "action": <action name -- must be one of
    move_forward, turn_left, turn_right, stop>
}
```

USER: Here are the sequence of frames from the walkthrough video demonstrating the route you need to take. Analyze the walkthrough to understand the movements and the maze structure. Take a note of all the details needed to help you repeat this route when navigating next. Think step by step.

```
<IMAGE 1>
<IMAGE 2>
.
.
.
```

ASSISTANT: ...

USER: Now, you must navigate to the goal. Here is the goal description and the image: Painting of a Soccer_Ball

```
<IMAGE OF Soccer_Ball>
```

USER: Here is the current observation.

```
<CURRENT IMAGE OBSERVATION>
```

ASSISTANT: ...

USER: Here is the current observation.

```
<CURRENT IMAGE OBSERVATION>
```

.
.
.

Prompt 11: Route retracing (BEV image)

SYSTEM: You are playing a game in a 2D world. You will be shown a video of the shortest route from an initial position to a goal. You must look at the video and understand the 2D world structure and the route taken. Then, you will be placed in the 2D world at the same initial position. You must navigate from the initial position to the goal using the same route shown in the video, as quickly as possible. Below, you will find sections highlighting more details about the 2D world and the task. You can refer to these for more information.

2D WORLD:
The world consists of the following.
* black cells: these are obstacles, i.e., you cannot move over them
* blue cells: these are navigable spaces, i.e., you can move over them
Some blue cells contain landmarks, which are red circles filled with a text character. These are important as they will allow you to better understand the world and locate yourself. Your position will be marked using a yellow square.

OBSERVATIONS:
The images are recorded from a birds-eye view of the 2D world. The images capture a local neighborhood surrounding your current position in the world, i.e., you will always remain at the center of the image while the world changes around you.

GOAL:
You will be asked to navigate to a goal landmark. You must find the goal in the 2D world by repeating the path shown in the video. Once you find it, move to the location of the goal till you are standing on the landmark and then execute a stop action.

ACTIONS:
You have four actions available.
up: move up by one unit cell
down: move down by one unit cell
left: move left by one unit cell
right: move right by one unit cell
stop: ends the current task. Issue this action only if you think you have reached the goal. If you haven't reached the goal, this action will result in a navigation failure that cannot be recovered from.

STOP CRITERIA:
Before executing stop, you must ensure that you've "reached" the goal correctly. To reach a goal, you have to move to the cell containing the goal landmark. Then execute the stop action.

RESPONSE FORMAT:
Respond in the following format:
Reasoning: text explanation string in one or two short sentences — provide all your explanations and inner thoughts here - avoid verbosity and be concise
Intent: state your intent in one short sentence, i.e., what you are trying to achieve
Then provide the final action to take in a json formatted string.

```
```
{
    "action": <action name -- must be one of up, down, left, right>
}
```
```

USER: Here is sequence of video frames recorded in the 2D world. This demonstrates the route you need to repeat. Analyze the video to understand the movements and the world structure. Take a note of all the details needed to help you repeat this route when navigating next. Think step by step.

```
<IMAGE 1>
<IMAGE 2>
.
.
.
```

ASSISTANT: ...

USER: Now, you must navigate to the goal based on your knowledge of the 2D world you obtained from the video. Here is the goal description: landmark Y

USER: Here is the local view of your surroundings in the 2D world. You are at the center of this view.

```
<CURRENT IMAGE OBSERVATION>
```

ASSISTANT: ...

USER: Here is the local view of your surroundings in the 2D world. You are at the center of this view.

```
<CURRENT IMAGE OBSERVATION>
.
.
.
```

Prompt 12: Route retracing (BEV text) — part 1

SYSTEM: You are playing a game in a 2D text world. The console screen of the game is represented as a comma-separated text array. You will be shown a sequence of console screen recordings that demonstrate the shortest route from an initial position to a goal. You must look at the sequence and understand the 2D text world structure and the route taken. Then, you will be placed in the 2D text world at the same initial position. You must navigate from the initial position to the goal using the same route shown in the screen recording sequence, as quickly as possible. Below, you will find sections highlighting more details about the 2D text world and the task. You can refer to these for more information.

2D TEXT WORLD:
The console of the game is represented as a comma-separated text array. Obstacles are represented using 0, i.e., you cannot move over them. Navigable spaces that you can move over are represented using 1. Some navigable spaces have landmarks represented as an ascii character (A - Z). These are also navigable spaces and are just labeled with an ascii character. These landmarks are important to remember. You will always be located at the center of the array with your position highlighted using the ”*” character.

OBSERVATIONS:
The images are recorded from a birds-eye view of the 2D world. The images capture a local neighborhood surrounding your current position in the world, i.e., you will always remain at the center of the image while the world changes around you.

GOAL:
You will be asked to navigate to a goal landmark. You must find the goal in the 2D text world by repeating the path shown in the console screen recording sequence. Once you find it, move to the location of the goal till you are standing on the landmark and then execute a stop action.

ACTIONS:
You have four actions available.
up: move up by one unit cell
down: move down by one unit cell
left: move left by one unit cell
right: move right by one unit cell
stop: ends the current task. Issue this action only if you think you have reached the goal. If you haven’t reached the goal, this action will result in a navigation failure that cannot be recovered from.

STOP CRITERIA:
Before executing stop, you must ensure that you’ve ”reached” the goal correctly. To reach a goal, you have to move to the cell containing the goal landmark. Then execute the stop action.

RESPONSE FORMAT:
Respond in the following format:
Reasoning: text explanation string in one or two short sentences — provide all your explanations and inner thoughts here - avoid verbosity and be concise Intent: state your intent in one short sentence, i.e., what you are trying to achieve Then provide the final action to take in a json formatted string.

```
{
    "action": <action name -- must be one of up, down, left, right>
}
```

USER: Here is the sequence of console screen recordings taken in the 2D text world. This demonstrates the route you need to repeat. Analyze the sequence to understand the movements and the world structure. Take a note of all the details needed to help you repeat this route when navigating next. Think step by step.

```
### Console screen recorded at time = 0
```

```
0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0
```

```
### Console screen recorded at time = 1
```

```
0,0,0,0,0
0,1,1,1,1
0,C,*,1,1
0,0,1,1,0
0,1,1,1,1
```

.
.
.

Prompt 13: Route retracing (BEV text) — part 2

ASSISTANT: ...

USER: Now, you must navigate to the goal based on your knowledge of the 2D text world you obtained from the sequence of console screen recordings. Here is the goal description: landmark Y

USER: Here is a birds-eye view of the 5x5 area surrounding your current position. You are located at the center of this view. Your position is denoted by "*".

```
0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0
```

The landmarks visible in your local context are: C. Note that the landmark locations are also navigable spaces, i.e., you can move over them. Your objective is to reach landmark: Y

ASSISTANT: ...

USER: Here is a birds-eye view of the 5x5 area surrounding your current position. You are located at the center of this view. Your position is denoted by "*".

```
0,0,0,0,0
0,1,1,1,1
0,C,*,1,1
0,0,1,1,0
0,1,1,1,1
```

The landmarks visible in your local context are: C. Note that the landmark locations are also navigable spaces, i.e., you can move over them. Your objective is to reach landmark: Y

ASSISTANT: ...

.
.
.

Prompt 14: Novel shortcuts (Ego image)

SYSTEM: You are a sentient living creature capable of navigating in environments, building internal spatial representations of environments, and finding goals in them. You will be shown a video of some route from the initial position to the goal. You must look at the video and understand the environment structure, and remember the locations of the start and the goal. The video may show a long-winded route from the start to the goal with unnecessary detours. Based on the environment structure, you must identify a faster route to the goal. Then, you will be placed in the environment at the same initial position. You must navigate to the goal using your identified shortest route as quickly as possible. Below, you will find sections highlighting more details about the task. You can refer to these for more information.

OBSERVATIONS:
The images are recorded from a perspective viewpoint (i.e., egocentric or first-person). This means that you are likely to see objects from different angles, resulting in a skewed appearance of the underlying 3D objects. It is important for you to look past this skew in the appearance and perceive the true shape of the object in 3D.

GOAL:
You will be provided an object goal using a text description and an image of the object. You must find the goal object in the environment by identifying the shortest route based on your experience from the video. Once you find the goal, move close to its location and re-orient yourself to face the object.

ACTIONS:
You have four actions available.
move_forward: move forward by 0.25m along the current heading direction. It does not change the heading angle.
turn_left: decrease your heading angle by 30 degrees. It does not change the (x, y) position.
turn_right: increase your heading angle by 30 degrees. It does not change the (x, y) position.
stop: ends the current task. Issue this action only if you think you have reached the goal. If you haven't reached the goal, this action will result in a navigation failure that cannot be recovered from.

STUCK IN PLACE BEHAVIOR:
Avoid getting stuck in one place, i.e., do not alternate between left and right turns without going anywhere. You must try and move around consistently without being stuck in one place.

STOP CRITERIA:
Before executing stop, you must ensure that you've "reached" the goal correctly. To reach a goal, you have to move the robot close enough to the wall where you see the goal, and see the object clearly in your observation in front of you.

RESPONSE FORMAT:
Respond in the following format:
Reasoning: text explanation string in one or two short sentences — provide all your explanations and inner thoughts here - avoid verbosity and be concise
Intent: state your intent in one short sentence, i.e., what you are trying to achieve
Then provide the final action to take in a json formatted string.

```
{
    "action": <action name -- must be one of move_forward, turn_left, turn_right, stop>
}
```

USER: Here are the sequence of frames from the walkthrough video demonstrating a suboptimal route from the start to some goal location. Analyze the walkthrough to understand the movements and the environment structure. Keep track of the start and goal locations, and the current location in the environment as you watch the walkthrough. Then plan a shortcut route that takes you to the goal while avoiding unnecessary detours. Think step by step.

<IMAGE 1>
<IMAGE 2>
.
.
.

ASSISTANT: ...

USER: Now, you must navigate to the goal. Here is the goal description and the image: Painting of a Soccer_Ball

<IMAGE OF Soccer_Ball>

USER: Here is the current observation.

<CURRENT IMAGE OBSERVATION>

ASSISTANT: ...

USER: Here is the current observation.

<CURRENT IMAGE OBSERVATION>

ASSISTANT: ...

.
.
.

Prompt 15: Novel shortcuts (BEV image)

SYSTEM: You are playing a game in a 2D world. You will be shown a video of some route from an initial position to a goal. You must look at the video and understand the 2D world structure and remember the locations of the start and the goal. The video may show a long-winded route from the start to the goal with unnecessary detours. Based on the world structure, you must identify a faster route to the goal. Then, you will be placed in the 2D world at the same initial position. You must navigate from the initial position to the goal using your identified shortest route as quickly as possible. Below, you will find sections highlighting more details about the 2D world and the task. You can refer to these for more information.

2D WORLD:
The world consists of the following.
* black cells: these are obstacles, i.e., you cannot move over them
* blue cells: these are navigable spaces, i.e., you can move over them
Some blue cells contain landmarks, which are red circles filled with a text character. These are important as they will allow you to better understand the world and locate yourself. Your position will be marked using a yellow square.

OBSERVATIONS:
The images are recorded from a birds-eye view of the 2D world. The images capture a local neighborhood surrounding your current position in the world, i.e., you will always remain at the center of the image while the world changes around you.

GOAL:
You will be asked to navigate to a goal landmark. You must find the goal in the 2D world by identifying the shortest path based your your experience from the video. Once you find it, move to the location of the goal till you are standing on the landmark and then execute a stop action.

ACTIONS:
You have four actions available.
up: move up by one unit cell
down: move down by one unit cell
left: move left by one unit cell
right: move right by one unit cell
stop: ends the current task. Issue this action only if you think you have reached the goal. If you haven't reached the goal, this action will result in a navigation failure that cannot be recovered from.

STOP CRITERIA:
Before executing stop, you must ensure that you've "reached" the goal correctly. To reach a goal, you have to move to the cell containing the goal landmark. Then execute the stop action.

RESPONSE FORMAT:
Respond in the following format:
Reasoning: text explanation string in one or two short sentences — provide all your explanations and inner thoughts here - avoid verbosity and be concise
Intent: state your intent in one short sentence, i.e., what you are trying to achieve
Then provide the final action to take in a json formatted string.

```
```
{
    "action": <action name -- must be one of up, down, left, right>
}
```
```

USER: Here is the sequence of video frames recorded in the 2D world. This demonstrates a suboptimal route from the start to some goal location. Analyze the video to understand the movements and the world structure. Keep track of the start and goal locations, and the current location in the world as you watch the video. Then plan a shortcut route that takes you to the goal while avoiding any unnecessary detours. Think step by step.

<IMAGE 1>
<IMAGE 2>
.
.
.

ASSISTANT: ...

USER: Now, you must navigate to the goal based on your knowledge of the 2D world you obtained from the video. Here is the goal description: landmark Y

USER: Here is the local view of your surroundings in the 2D world. You are at the center of this view.

<CURRENT IMAGE OBSERVATION>

ASSISTANT: ...

USER: Here is the local view of your surroundings in the 2D world. You are at the center of this view.

<CURRENT IMAGE OBSERVATION>

ASSISTANT: ...

.
.
.

Prompt 16: Novel shortcuts (BEV text) — part 1

SYSTEM: You are playing a game in a text 2D world. The console screen of the game is represented as a comma-separated text array. You will be shown a sequence of console screen recordings that demonstrates a route from an initial position to a goal. You must look at the sequence and understand the 2D text world structure and remember the locations of the start and the goal. The recordings may show a long-winded route from the start to the goal with unnecessary detours. Based on the world structure, you must identify a faster route to the goal. Then, you will be placed in the 2D text world at the same initial position. You must navigate from the initial position to the goal using your identified shortest route as quickly as possible. Below, you will find sections highlighting more details about the 2D text world and the task. You can refer to these for more information.

2D TEXT WORLD:
The console of the game is represented as a comma-separated text array. Obstacles are represented using 0, i.e., you cannot move over them. Navigable spaces that you can move over are represented using 1. Some navigable spaces have landmarks represented as an ascii character (A - Z). These are also navigable spaces and are just labeled with an ascii character. These landmarks are important to remember. You will always be located at the center of the array with your position highlighted using the "*" character.

OBSERVATIONS:
The images are recorded from a birds-eye view of the 2D world. The images capture a local neighborhood surrounding your current position in the world, i.e., you will always remain at the center of the image while the world changes around you.

GOAL:
You will be asked to navigate to a goal landmark. You must find the goal in the 2D text world by identifying the shortest path based your your experience from the screen recording sequence. Once you find it, move to the location of the goal till you are standing on the landmark and then execute a stop action.

ACTIONS:
You have four actions available.
up: move up by one unit cell
down: move down by one unit cell
left: move left by one unit cell
right: move right by one unit cell
stop: ends the current task. Issue this action only if you think you have reached the goal. If you haven't reached the goal, this action will result in a navigation failure that cannot be recovered from.

STOP CRITERIA:
Before executing stop, you must ensure that you've "reached" the goal correctly. To reach a goal, you have to move to the cell containing the goal landmark. Then execute the stop action.

RESPONSE FORMAT:
Respond in the following format:
Reasoning: text explanation string in one or two short sentences — provide all your explanations and inner thoughts here - avoid verbosity and be concise
Intent: state your intent in one short sentence, i.e., what you are trying to achieve
Then provide the final action to take in a json formatted string.

````
```
{
    "action": <action name -- must be one of up, down, left, right>
}
```
````

USER: Here is the sequence of console screen recordings taken in the 2D text world. This demonstrates a suboptimal route from the start to some goal location. Analyze the sequence to understand the movements and the world structure. Keep track of the start and goal locations, and the current location in the world as you study the sequence. Then plan a shortcut route that takes you to the goal while avoiding any unnecessary detours. Think step by step.

````
### Console screen recorded at time = 0

```
0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0
```

### Console screen recorded at time = 1

```
0,0,0,0,0
0,0,0,0,0
0,0,*,1,1
0,0,C,1,1
0,0,0,1,1
```

.
.
.
````

Prompt 17: Novel shortcuts (BEV text) — part 2

ASSISTANT: ...

USER: Now, you must navigate to the goal based on your knowledge of the 2D text world you obtained from the sequence of console screen recordings. Here is the goal description: landmark Y

USER: Here is a birds-eye view of the 5x5 area surrounding your current position. You are located at the center of this view. Your position is denoted by ”*”.

```
0,0,0,0,0
0,0,0,0,0
0,1,*,1,1
0,C,1,1,1
0,0,1,1,0
```

The landmarks visible in your local context are: C. Note that the landmark locations are also navigable spaces, i.e., you can move over them. Your objective is to reach landmark: Y

ASSISTANT: ...

USER: Here is a birds-eye view of the 5x5 area surrounding your current position. You are located at the center of this view. Your position is denoted by ”*”.

```
0,0,0,0,0
0,0,0,0,0
1,1,*,1,0
C,1,1,1,0
0,1,1,0,0
```

The landmarks visible in your local context are: C. Note that the landmark locations are also navigable spaces, i.e., you can move over them. Your objective is to reach landmark: Y

ASSISTANT: ...

Prompt 18: Mental rotation test (vision)

USER: Here is an image of a three-dimensional shape.

<IMAGE OF REFERENCE 3D SHAPE>

Which of these images show the same object rotated in 3D?

Choice 1

<CHOICE 1 IMAGE>

Choice 2

<CHOICE 2 IMAGE>

Choice 3

<CHOICE 3 IMAGE>

Choice 4

<CHOICE 4 IMAGE>

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```

Prompt 19: Mental rotation test (text)

USER: Here is a two-dimensional array.

```
over,none,page
such,none,free
site,none,list
```

Which of these options show the same array rotated in 2D? Note: It must only be rotated, not mirrored.

Choice 1:

```
page,free,list
none,none,none
over,such,site
```

Choice 2:

```
list,free,page
none,none,none
site,such,over
```

Choice 3:

```
page,none,over
free,none,such
list,none,site
```

Choice 4:

```
over,such,site
none,none,none
page,free,list
```

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 20: Perspective taking test (vision)

USER: Here is an image of various objects (animate and inanimate) on a two-dimensional plane.

```
<IMAGE OF OBJECTS>
```

Pretend that you are standing at the centroid of guitar and facing the centroid of bat. Visualize the world around you. At what angle (from -180 to 180 degrees) is snake located relative to you? Clockwise rotations are positive and anti-clockwise rotations are negative.
Here are your options: 1) 45 2) 85 3) 105 4) 5
Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 21: Perspective taking test (text)

USER: Here is an array of numbers representing the birds-eye view of a two-dimensional plane.

```
0,7,0,9
8,0,0,0
0,0,0,5
0,0,0,3
```

Empty locations are indicated using 0. Important locations are indicated with a number 1 - 9. Pretend that you are standing at the location 8 and facing the location 9. Visualize the world around you. At what angle (from -180 to 180 degrees) is the location 3 relative to you? Clockwise rotations are positive and anti-clockwise rotations are negative.
Here are your options: 1) 52 2) 32 3) 72 4) 112
Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 22: Water level test (vision)

USER: Here is a container filled with water.

```
<IMAGE OF FILLED WATER CONTAINER>
```

What will be the water level when it is rotated as shown here?

```
<IMAGE OF ROTATED EMPTY WATER CONTAINER>
```

Here are your choices. Which of these match the expected water level in the rotated container?

Choice 1

```
<IMAGE OF CHOICE 1>
```

Choice 2

```
<IMAGE OF CHOICE 2>
```

Choice 3

```
<IMAGE OF CHOICE 3>
```

Choice 4

```
<IMAGE OF CHOICE 4>
```

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 23: Minnesota Paper Form Board (vision)

USER: This image shows the different pieces of a puzzle.

```
<IMAGE OF PUZZLE PIECES>
```

These pieces are put together by an oracle. Which one of these four options shows what it would look like when the pieces are put together? Pay close attention to not just the final fitted shape, but also the individual pieces contained within the shape.

Choice 1

```
<IMAGE OF CHOICE 1>
```

Choice 2

```
<IMAGE OF CHOICE 2>
```

Choice 3

```
<IMAGE OF CHOICE 3>
```

Choice 4

```
<IMAGE OF CHOICE 4>
```

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 24: Minnesota Paper Form Board (text)

USER: You are playing putting together a text jigsaw puzzle. Here are the pieces, where 0 represents the interiors of the puzzle piece and 1 represents the boundary.

```
1,1,1,1,1,1
1,0,0,0,0,1
1,0,0,0,0,1
1,0,0,0,0,1
1,0,0,0,0,1
1,0,0,0,0,1
1,1,1,1,1,1

1,1,1,1,1,1,1
1,0,0,0,0,0,1
1,0,0,0,0,0,1
1,0,0,0,0,0,1
1,1,1,1,1,1,1

1,1,1,1,1,1
1,0,0,0,0,1
1,0,0,0,0,1
1,1,1,1,1,1

1,1,1,1,1
1,0,0,0,1
1,0,0,0,1
1,1,1,1,1
```

These pieces are now put together to solve the puzzle by an oracle. Which of these four options shows what it would look like when the pieces are put together?

Choice 1:

`<CHOICE 1 ARRAY>`

Choice 2:

`<CHOICE 2 ARRAY>`

Choice 3:

`<CHOICE 3 ARRAY>`

Choice 4:

`<CHOICE 4 ARRAY>`

Pay close attention to not just the final fitted shape, but also the individual pieces contained within the shape. Also, note that the puzzle pieces may need to be rotated before fitting them together.Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```

Prompt 25: Judgement of line orientations (vision)

USER: Here is an image showing two lines. Your goal is to measure the angle between the two lines.

`<IMAGE OF LINES>`

Here is a legend showing a set of reference lines numbered from 1 to 11.

`<IMAGE OF LEGEND>`

Which of the following reference line pairs match the angle between the original lines shown in the image?
1) Lines 1 and 9
2) Lines 1 and 7
3) Lines 1 and 10
4) Lines 1 and 3
Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```

Prompt 26: Judgement of line orientations (text)

USER: Here is a reference array.

```
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
2,2,2,2,2,2,0,0,0,0,0,0,0,1,1,1,1,1,1,1
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

0s mean empty space, ignore them. There are two lines made out of 1s and 2s, respectively. Your goal is to measure the angle between the two lines. Specifically, here are four choices of arrays with two lines per array. Which one of the choices has an angle between the two lines that matches the angle between lines in the reference array?

Choice 1:

```
<ARRAY OF CHOICE 1>
```

Choice 2:

```
<ARRAY OF CHOICE 2>
```

Choice 3:

```
<ARRAY OF CHOICE 3>
```

Choice 4:

```
<ARRAY OF CHOICE 4>
```

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 27: Selective attention task (vision)

USER: Here is an image of a apple. Let us call this the target.

```
<IMAGE OF TARGET>
```

Here is a grid of apple images. This contains multiple instances of apple, but not all of them are the target object. The grid is indexed from top-left to bottom-right, starting from row, column = (0, 0).

```
<IMAGE OF GRID>
```

Which of these options represent the true locations of the target object in the grid? Locations are represented as (row, column).

Choice 1. (0, 3), (1, 2), (2, 2), (3, 3)
Choice 2. (0, 1), (1, 0), (2, 0), (3, 3)
Choice 3. (0, 3), (1, 2), (2, 2), (3, 1)
Choice 4. (0, 3), (1, 2), (2, 2), (3, 1)

Think step by step. Then answer your question in the following json format.

````
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
````

Prompt 28: Selective attention task (text)

USER: Here is a grid of numbers / letters. The grid is indexed from top-left to bottom-right, starting from row, column = (0, 0).

```
f,t,o,e,r
r,e,o,t,f
f,r,o,t,e
f,e,o,r,t
o,t,r,e,f
```

Your goal is to find all occurrences of ‘e‘ in the grid. Which of these options represent the true locations of the where ‘e‘ occurs in the grid? Locations are represented as (row, column).

Choice 1. (0, 3), (1, 1), (2, 4), (3, 1), (4, 3)
Choice 2. (0, 3), (1, 1), (2, 4), (3, 1), (4, 4)
Choice 3. (0, 1), (1, 1), (2, 0), (3, 2), (4, 1)
Choice 4. (0, 2), (1, 1), (2, 4), (3, 1), (4, 3)

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```

Prompt 29: Maze completion task (vision)

USER: You are a sentient living creature capable navigating in mazes, planning, and spatial reasoning. You are playing a Pacman-style maze game. You start at some random position in the maze. You must escape the maze as quickly as possible to reach the goal. You are given the game screen that shows the following:
* maze structure - blue is obstacle space, black is navigable space. You can only move on black spaces. You cannot move through blue spaces.
* your current position - yellow square
* goal position - red circle

Below the screen, a status message might appear indicating that you collided into a wall after your previous action.

Actions available: You can take five possible actions.
* left - move left from your current position by one step
* right - move right from your current position by one step
* up - move up from your current position by one step
* down - move down from your current position by one step
* stop - issue this action only after you have reached the goal position. If you execute it prematurely, you will fail. If you do not execute it after reaching the goal, you will again fail.

Response format: Respond in the following format.

````
<text explanation string - explain your reasoning concisely>
<next, provide a json formatted output with the next action>
```
{
    "action": "<action>"
}
```
````

ASSISTANT: ...

USER: Here is the current state of the maze.

<IMAGE OF MAZE>

Think step-by-step about how to reach the goal. What action do you take next?

ASSISTANT: ...

USER: Here is the current state of the maze.

<IMAGE OF MAZE>

Think step-by-step about how to reach the goal. What action do you take next?

.
.
.

Prompt 30: Maze completion task (text)

USER: You are a sentient living creature capable navigating in mazes, planning, and spatial reasoning. You are playing a text-based maze game. You start at some random position in the maze. You must escape the maze as quickly as possible to reach the goal. You are given a 2D array representing the maze, which contains the following:
* maze structure - 0 is obstacle space, 1 is navigable space. You can only move on 1s (i.e., navigable spaces). You cannot move through 0s (i.e., obstacles).
* your current position - marked as A
* goal position - marked as G

Goal and current positions are always navigable spaces.

Actions available: You can take five possible actions.
* left - move left from your current position by one step
* right - move right from your current position by one step
* up - move up from your current position by one step
* down - move down from your current position by one step
* stop - issue this action only after you have reached the goal position. If you execute it prematurely, you will fail. If you do not execute it after reaching the goal, you will again fail.

Response format: Respond in the following format.

````
<Think step-by-step about what action to take next. Be concise.>
<next, provide a json formatted output with the next action>
```
{
    "action": "<action>"
}
```
````

ASSISTANT: ...

USER: Here is the current view of the maze.

```
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
0,A,0,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0
0,1,0,1,0,1,0,1,0,0,0,0,0,1,0,1,0,0,0,1,0,1,0
0,1,1,1,0,1,0,1,1,1,1,1,0,1,1,1,1,1,0,1,0,1,0
0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,1,0
0,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,0,1,1,1,0,1,0
0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,1,0,1,0,1,0,1,0
0,1,1,1,1,1,0,1,1,1,1,1,0,1,0,1,0,1,0,1,1,1,0
0,1,0,1,0,1,0,0,0,0,0,1,0,0,0,1,0,1,0,0,0,0,0
0,1,0,1,0,1,1,1,0,1,1,1,1,1,0,1,1,1,0,1,1,1,0
0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0
0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0
0,1,0,0,0,1,0,0,0,0,0,1,0,1,0,0,0,0,0,1,0,1,0
0,1,0,1,1,1,1,1,0,1,0,1,0,1,0,1,1,1,0,1,0,1,0
0,1,0,0,0,0,0,0,0,1,0,0,0,1,0,1,0,0,0,0,0,1,0
0,1,1,1,0,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,0,1,0
0,1,0,1,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0
0,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,0,1,1,1,0,1,0
0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,1,0
0,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,0
0,0,0,0,0,1,0,0,0,1,0,1,0,1,0,0,0,1,0,0,0,1,0
0,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,1,0,G,1,1,0
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

0 represents obstacles. 1 represents free spaces. G is the goal. A is your current position in the maze. Your current location in the maze is row, column = (1, 1). The goal location is row, column = (21, 19). Think step-by-step about how to reach the goal. What action do you take next?

ASSISTANT: ...

.
.
.

Prompt 31: Corsi block-tapping task (text)

USER: You are playing the Corsi board tapping game. The board is represented as a two dimensional array. The array contains empty locations (marked as 0) and 7 box locations (marked as B). You will be shown an ordered sequence of 6 arrays, representing a sequence of taps on the board. At each step of the sequence, one of the boxes is tapped, and it the tap is highlighted by marking the box as a T instead of B. You must remember the sequence of taps by remembering which exact boxes were tapped. Here is the sequence of arrays.

```
0,0,0,0,0,B,0
0,0,B,0,0,0,0
0,0,0,0,0,0,0
0,B,0,0,0,0,0
B,0,0,T,0,0,B
0,0,0,0,0,0,0
0,0,B,0,0,0,0

0,0,0,0,0,B,0
0,0,B,0,0,0,0
0,0,0,0,0,0,0
0,B,0,0,0,0,0
B,0,0,B,0,0,T
0,0,0,0,0,0,0
0,0,B,0,0,0,0

0,0,0,0,0,T,0
0,0,B,0,0,0,0
0,0,0,0,0,0,0
0,B,0,0,0,0,0
B,0,0,B,0,0,B
0,0,0,0,0,0,0
0,0,B,0,0,0,0

0,0,0,0,0,B,0
0,0,B,0,0,0,0
0,0,0,0,0,0,0
0,B,0,0,0,0,0
B,0,0,B,0,0,B
0,0,0,0,0,0,0
0,0,T,0,0,0,0

0,0,0,0,0,B,0
0,0,T,0,0,0,0
0,0,0,0,0,0,0
0,B,0,0,0,0,0
B,0,0,B,0,0,B
0,0,0,0,0,0,0
0,0,B,0,0,0,0

0,0,0,0,0,B,0
0,0,B,0,0,0,0
0,0,0,0,0,0,0
0,B,0,0,0,0,0
T,0,0,B,0,0,B
0,0,0,0,0,0,0
0,0,B,0,0,0,0
```

You must now identify the sequence of taps. Here is the corsi board with numbers 1 - 7 assigned on each box. Use this numbering as a reference to answer the question.

```
0,0,0,0,0,3,0
0,0,5,0,0,0,0
0,0,0,0,0,0,0
0,7,0,0,0,0,0
6,0,0,1,0,0,2
0,0,0,0,0,0,0
0,0,4,0,0,0,0
```

What is the sequence of boxes that were tapped? Here are your choices:
(1) 1, 2, 3, 4, 5, 6
(2) 1, 2, 3, 4, 6, 5
(3) 6, 5, 3, 4, 7, 1
(4) 2, 1, 5, 4, 3, 6

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```

Prompt 32: Corsi block-tapping task (vision)

USER: You are playing the Corsi board tapping game. You will be shown a video with 5 boxes (blue squares). These boxes will be tapped one at a time in a specific sequence. A tap on a box will be shown by highlighting the box in yellow. You must remember the sequence of taps by remembering which exact boxes were tapped. Here is the video.

```
<IMAGE 1>
<IMAGE 2>
<IMAGE 3>
```

You must now identify the sequence of taps. Here is an image of the corsi board with numbers on each box. Use this image as a reference to answer the question.

```
<IMAGE WITH NUMBERS ON BOXES>
```

Here are your choices:
(1) 1, 4, 0
(2) 1, 0, 4
(3) 0, 1, 4
(4) 1, 0, 2

Think step by step. Then answer your question in the following json format.

```
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
```

Prompt 33: Spatial addition task (vision)

USER: You are playing the array addition game. You have to add two arrays by following certain rules. Each array location can be empty (i.e., fully white) or filled with colored circles. Empty locations represent zeros. The colors of the circles mean specific things.
* Blue circle is a one
* Red circle is a distrction and must be ignored (i.e., it does not contribute to the array addition)
* White circle is a two

Array addition works as follows:
* sum of zeros must be a zero (i.e., an empty array cell)
* sum of one and zero (or zero and one) must be one (i.e., a blue circle)
* sum of one and one must be two (i.e., a white circle)

Here is the first array.

```
<IMAGE OF ARRAY 1>
```

Here is the second array.

```
<IMAGE OF ARRAY 2>
```

What is the sum of the two arrays? Pick from one of these four choices.

Choice 1

```
<IMAGE OF CHOICE 1>
```

Choice 2

```
<IMAGE OF CHOICE 2>
```

Choice 3

```
<IMAGE OF CHOICE 3>
```

Choice 4

```
<IMAGE OF CHOICE 4>
```

Think step by step. Then answer your question in the following json format.

```
```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```
```

Prompt 34: Spatial addition task (text)

USER: You are playing the array addition game. You have to add two arrays by following certain rules. Each array location can be filled with E, B, R or W. E represent a zero. B represents a one. W represents a two. R is a distraction and must be ignored (i.e., it does not contribute to the array addition).

Array addition works as follows:
* sum of zeros must be a zero (i.e., E)
* sum of one and zero (or zero and one) must be one (i.e., B)
* sum of one and one must be two (i.e., W)

Here is the first array.

```
E,E,R,E,E,E,E
E,E,E,E,E,B,E
E,E,E,E,E,E,E
E,E,E,E,E,E,E
B,E,B,E,E,E,E
E,E,B,E,E,B,E
E,E,E,E,E,E,E
```

Here is the second array.

```
E,E,E,E,E,E,E
E,E,E,E,E,B,E
E,E,E,E,E,E,E
E,R,E,E,E,E,E
E,E,E,E,E,B,E
E,E,B,E,E,B,E
B,E,E,E,E,E,E
```

What is the sum of the two arrays? Pick from one of these four choices.

Choice 1:

```
E,E,E,E,E,E,E
E,B,E,E,E,B,B
E,E,E,E,E,E,B
E,E,E,E,E,E,E
B,E,B,E,B,E,E
E,E,B,E,E,B,E
E,E,E,E,E,B,E
```

Choice 2:

```
E,E,E,E,E,E,E
E,E,E,E,E,W,E
E,E,E,E,E,E,E
E,E,E,E,E,E,E
B,E,B,E,E,B,E
E,E,W,E,E,W,E
B,E,E,E,E,E,E
```

Choice 3:

```
E,E,E,E,E,E,E
B,E,E,E,E,B,E
E,E,E,E,E,E,E
E,E,E,E,E,E,E
B,E,W,B,E,E,E
B,E,B,E,E,B,B
E,E,E,E,E,E,E
```

Choice 4:

```
E,E,E,E,E,E,B
E,E,E,E,E,B,E
B,E,E,E,B,E,E
E,E,E,E,E,E,E
B,B,B,E,E,E,E
E,E,B,E,E,W,E
E,E,E,E,E,E,E
```

Think step by step. Then answer your question in the following json format.

```
{
    "answer": <fill in one of 1/2/3/4 integer value>
}
```

Prompt 35: Cambridge spatial working memory (vision)

USER: You are playing the Cambridge Spatial Working Memory game. You will be shown a screen with blue boxes. A treasure is hidden in one of the blue boxes. You must identify the box containing the treasure, which is shown as an yellow square. Once you find a treasure, it will be collected and placed in the "Treasures collected" section shown below the image. A new treasure will be hidden in one of the other boxes where the treasure did not appear before. You must again find the new treasure. This process is repeated till you find all treasures placed in each of the blue boxes once. Note: The treasure will never appear in a box where it had already been placed. Each turn, there are randomly selected numbers associated with each box. These numbers are meant to aid you with communication, i.e., specify what box you want to open in that turn. However, these numbers will change after every turn. So do NOT associate boxes with numbers over the long term. The number identity of a box can change any time. Therefore, you must remember the boxes based on their spatial positions and not the numbers.

RESPONSE FORMAT:
Think step-by-step about where the treasure might be based on your past actions. After that, indicate the box you want to open in the following json format:

```
{
    "action": <box integer index>
}
```

ASSISTANT: ...

USER: Here is the current state of the game. You must find the next treasure. Note that the numbers of the boxes have changed, but the box locations are fixed. Decide which box you want to open next, and then use the number associated with the box as the action.

`<IMAGE OF SCREEN>`

ASSISTANT: ...

Prompt 36: Cambridge spatial working memory (text)

USER: You are playing the Cambridge Spatial Working Memory game. You will be shown an array with integers. 0 represents empty locations. Locations numbered 1 - 9 represent boxes. A treasure is hidden in one of the boxes. You must identify the box containing the treasure. Once you find a treasure, the location will be momentarily shown as a "T" indicating that the treasure was found. The treasure is then collected and a new treasure will be hidden in one of the other boxes where the treasure did not appear before. You must then find the new treasure. This process is repeated till you find all treasures placed in each of the boxes once. Note: The treasure will never appear in a box where it had already been placed.
While the boxes are represented using integers from 1 - 9, the true identity of the box is its location (row, column) in the array. The box location is always fixed (i.e., the boxes will not move and the number of boxes will not change). However, each turn, the integer id associated with the box will change randomly. These integer ids are meant to aid you with communication, i.e., specify what box you want to open in that turn. However, these numbers will change after every turn. So do NOT associate boxes with numbers over the long term. The number identity of a box can change any time. Therefore, you must remember the boxes based on their spatial positions and not the numbers.

RESPONSE FORMAT:
Think step-by-step about where the treasure might be based on your past actions. After that, indicate the box you want to open in the following json format:

```
{
    "action": <box integer index>
}
```

ASSISTANT: ...

USER: Here is the current view of the board. You must find the next treasure. Note that the numbers of the boxes have changed, but the box locations are fixed. Decide which box location you want to open next. Then provide the number associated with the box as the action.

```
0,0,0,0,0,0,0
0,0,0,0,0,3,0
0,0,0,0,0,0,0
0,0,0,0,0,0,1
0,0,0,0,0,0,0
0,0,0,0,0,0,0
0,0,0,2,0,0,0
```

Number of treasures collected: 0 / 3

ASSISTANT: ...