Spatial Representations in Multimodal AI Systems

Scott O. Murray
University of Washington
Department of Psychology
somurray@uw.edu

Bridget Leonard
University of Washington
Department of Psychology
bll313@uw.edu

Abstract

This study details how spatial information is represented within a multimodal AI system (GPT-4-turbo, "GPT-4v"), leveraging established methodologies from human cognitive science research. Our investigation shows both rich underlying spatial comprehension but also uncovers notable limitations. We found that the structure of spatial representation in GPT-4v is predominantly propositional, diverging from the analog-like representations that are characteristic of human and animal spatial cognition. This discrepancy becomes particularly evident in tasks requiring spatial manipulation or perspective shifts, where GPT-4v falls short. Our analysis aims to bridge the gap between AI and human cognition, highlighting critical areas for future research and development in multimodal intelligence.

1. Introduction

Spatial cognition—a cornerstone of human intelligence—enables us to navigate our environment, understand relationships between objects, and perform complex manipulations. However, as we push the boundaries of multimodal AI, questions arise about the fidelity of these systems' spatial representations compared to humans. We draw upon established approaches from cognitive science to probe the representations of multimodal AI. To preview, multimodal AI can form what appear to be very sophisticated spatial representations. However, they differ substantially from those of humans and result in enhanced performance in some domains and reduced performance in others.

1.1. Rich understanding but a failure to manipulate

Our study utilized GPT-4-vision (GPT-4v; "gpt-4-turbo-2024-04-09" via OpenAI's API) to conduct a series of experiments designed to elucidate the system's spatial understanding from visual stimuli. GPT-4v demonstrates a remarkable ability to extract spatial information and infer potential functions from simplified visual images. In Figure 1A, GPT-4v accurately describes a stick figure next to a large rectangle and a smaller square, suggesting a narrative of a person facing a barrier with an object or goal represented by the smaller square—perhaps

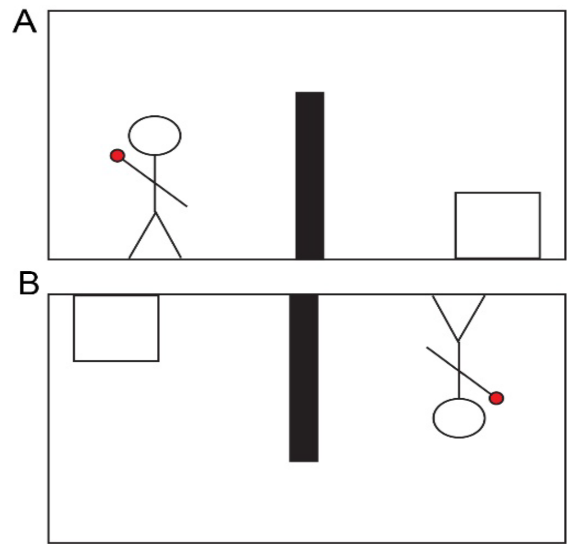

Figure 1. GPT4-v's response to "what's going on in this image?": (**A**) "*In this simple, abstract image, we see a stick figure with a red circular object (perhaps representing a ball) near its hand, standing next to a tall black rectangle (which could be interpreted as a wall or a barrier) and facing a small square or box placed at a distance on the right side. This setup might represent a scenario where the stick figure is attempting to throw or move the red ball over the black barrier to reach the square on the other side, possibly resembling a game or a challenge.*" (**B**) "*This image appears to be a simple, abstract representation, perhaps depicting a conceptual art or a minimalist graphic design. It consists of basic geometric shapes and a few lines: A large black vertical rectangle, probably representing a barrier or partition. A square outlined in black near the top left, having a similar style to the rectangle but detached and smaller. On the right side, there is a contraption-like structure comprising two lines that converge into a point, from which a circle and a smaller red circle connected by a line hang. This part might represent some sort of pendulum or mechanical device, though it is depicted in a stylized, simplified manner.*"

a target in a game. Overall, the description GPT-4v provides demonstrates its capacity for not just recognition but also contextual interpretation.

However, the fidelity of this spatial understanding is challenged when the visual input is altered. As demonstrated in Figure 1B, when the same image is inverted, GPT-4v's interpretation shifts drastically. It no longer perceives the figures as entities with potential narratives but reduces them to abstract shapes and lines, a pendulum, or line hang, losing the previously inferred context entirely. This is unlike human perception, which while sensitive to image orientation (as noted with the classic example of face recognition, [1]), can usually maintain recognition and functional understanding despite such transformations. This point is emphasized when GPT-4v was given the same upside-down image and prompted to "*imagine rotating this image upside-down*." Instead of reverting to anything similar to the original interpretation, it offered a description focused on geometric shapes and their arrangement without the earlier ascribed functionality: "*The image shows a minimalist abstract composition consisting of geometric shapes and lines. There's a large vertical black bar on the right, a square outline to the left, and a circular object connected by lines, with a smaller red circle attached off to one side by a diagonal line. If the image were flipped upside down, the vertical and horizontal positioning of these shapes would be inverted, but the abstract nature of the design means the general perception of the image wouldn't dramatically change in terms of recognizability or function—it would still appear as a collection of intriguing geometric shapes*."

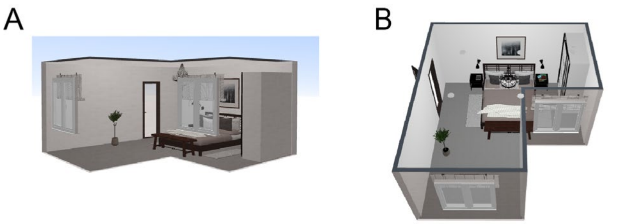

Figure 2. (**A** and **B**) A rendered scene of a bedroom from two different views. When prompted with only, "I have two images, can you compare them?" GPT4-v was able to recognize they are the same room: "The first image appears to be a three-dimensional rendering of a room from an angle that gives a slightly elevated side view. It shows part of the room with a bed, windows with curtains, a plant, and a picture on the wall. The perspective allows for a view of the room's layout, including the walls and door, but it's limited to one side of the room. The second image presents a top-down view of what seems to be the same room. This perspective shows the room's entire layout, including the placement of the bed, a desk with a computer, a bookshelf beside the bed, and the location of the windows and doors. It provides a more comprehensive view of the room arrangement and the relationship between different pieces of furniture. In summary, both images depict the same room, but from different angles, offering complementary views."

It was this initial observation – this inability to "imagine" the upside-down image properly rotated – that prompted us to explore the spatial reasoning abilities of GPT-4v in more detail.

1.2. Viewpoint invariance

Biological visual systems exhibit a robust capacity for viewpoint invariance, which allows for the recognition of objects and scenes from various perspectives. This ability is indispensable for navigation and interaction within our three-dimensional world. The struggles of GPT-4v with the inverted image presented in Figure 1 imply a potential shortfall in achieving this aspect of spatial cognition compared to humans.

We begin with an anecdotal example that appears to demonstrate robust viewpoint invariance. We analyzed GPT-4v's response to a complex scene of a 3D-rendered bedroom from two distinct viewpoints, depicted in Figures 2A and 2B. GPT-4v provided a detailed description of the scenes, recognizing them as two perspectives of the same room. Such a response could suggest an advanced capability to manipulate spatial information, similar to the mental flexibility demonstrated by humans and other animals in recognizing consistency within spatial transformations. On the other hand, the response could be achieved through comparing a list of linguistic labels for each room (e.g., the rooms contents and a description of their placement) and recognizing a high degree of overlap in the descriptions for each room.

To experimentally explore viewpoint invariance, we used abstract 3D figures with rotations about the vertical axis. These figures are reminiscent of those used in classic mental rotation studies [2], which are designed to assess the ability to process and manipulate spatial representations with minimal reliance on linguistic descriptions. The three object shapes that we used (in their 0-degree "reference" orientation) are shown in Figure 3A. GPT-4v was given two images to evaluate, one of the reference images (for example, Object 1 – deg 0) and a test image, either the same or different object in one of 10 different rotations around the vertical axis, ranging from 0 to 90 degrees, in steps of 10 degrees. The images were accompanied with the prompt, "*Examine the two images and decide whether they have the same three-dimensional shape.*" Twenty different iterations of each comparison were made to arrive at a measure of "percent of the time the two objects are labeled 'same.'" Thus, perfect performance would be 100% when the same object is used as the test and 0% when a different object is used.

When using Object 1 (0 degree) as the reference and the rotated versions of Object 1 as the test object (Figure 3B, solid line), nearly all orientations were correctly labeled "same". At first, this might suggest a high degree of rotation invariance. However, it might also simply reflect a strong bias to respond “same” for a comparison of any two abstract 3D shapes. Indeed, using a similar (3-armed) yet clearly distinct comparison object (Object 2), revealed a strong “respond same” bias (dashed line, Figure 3B; “correct” would be all values at 0%). These results show a surprising level of difficulty in simply distinguishing 3D structure from the objects.

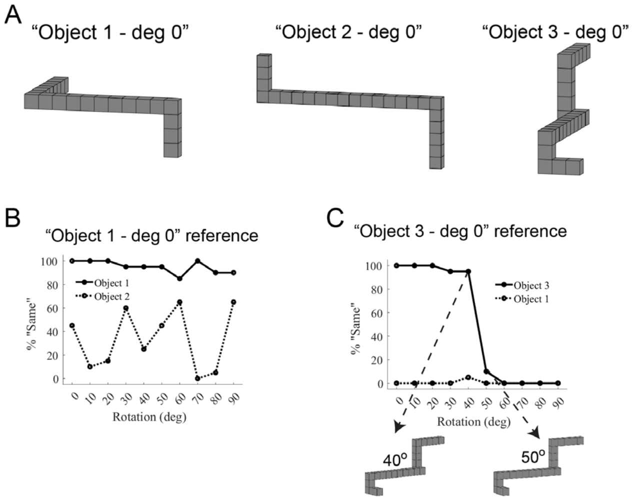

Figure 3. (**A**) Examples of the three objects in their “reference” (0-deg) orientation. (**B**) Performance when Object 1 was the reference (solid line = comparisons to rotated versions of Object 1; dashed line = comparisons to rotated versions of Object 2). (**C**) Performance when Object 3 was the reference. GPT-4v successfully discriminated all comparisons to Object 1 (dashed line near zero). However, it failed to recognize rotated versions of the Object 3 as “same” beyond a 40 deg rotation. Surprisingly there is a categorical difference in behavior between a 40-deg and 50-deg when, visually, these two objects appear very similar.

Using a more complex figure (Object 3) also revealed behavior inconsistent with mental rotation (Figure 3C). Up to 40-degree rotations of the Object 3 test object (i.e., same shape, different orientations), GPT-4v accurately responded “same” with same-choice levels at nearly 100%. Surprisingly, responses categorically switched at 50-degree rotations with “same”-choice percentages falling to nearly 0%. This categorical change in behavior is not reflected in image properties – in fact, through visual inspection it is difficult to recognize any difference between the 40- and 50-degree rotation images. To investigate the potential source of the categorical shift in behavior, we allowed GPT-4v full textual responses in a follow-up set of queries. There was no obvious discernable difference in the description strategies for 40-degree and 50-degree rotation objects.

As a final comparison, GPT-4v performed nearly perfectly in distinguishing Object 3 from all orientations of Object 1 (Figure 3C, dashed line) demonstrating that when there is sufficient image-level difference, GPT-4v can distinguish between abstract 3D objects.

Overall, we found no consistent evidence that GPT-4v is able to achieve viewpoint invariance. While it does respond “same” to 3D rotations of the same object, this likely represents a general inability to perceptually distinguish 3D objects of similar complexity (e.g., Object 1 vs. Object 2). In addition, it shows categorically different behavior across small changes in rotation which does not align with the analog-like representational structure of biological systems.

1.3. Perspective taking

The capacity for perspective taking is an integral component of human spatial reasoning, allowing us to conceptualize spatial relationships from different points of view. For example, consider Figure 2B's rendered room; humans can effortlessly answer questions like, “*If you walked into the room through the door, would the plant be to the left or right of you?*” We inherently adopt the incoming perspective and can navigate mentally to the target.

However, GPT-4v struggles with such perspective-taking tasks, performing at no better than chance levels. For example, in Figure 2A, GPT-4v incorrectly answered “left” on 9/10 trials, indicating a strong image-based bias in determining the location of the plant. For Figure 2B, GPT-4v answered “left” on 5/10 trials. Attempts to rephrase the question—such as, “...would you turn to the left or right to reach the plant?”—do not enhance GPT-4v’s performance. This is unexpected, particularly given GPT-4v’s earlier recognition of multiple views of the same room, suggesting a superficial level of spatial understanding.

Figure 4. Prompt: “A person is traveling down a road as depicted in this image. From the person's perspective, is the fire hydrant to the left or right of them?” (**A**) Performance, correct out of 50 iterations = 12% (**B**) 94%.

This concept is further illustrated in Figure 4, which presents a simplified schematic for perspective-taking evaluation. Not surprisingly, GPT-4v can adeptly describe the scenario depicted: a person walking down a street, providing rich and detailed context. Yet, when asked from the person's perspective if the fire hydrant is to their left or right, GPT-4v's response fails, leaning on the 2D image-based spatial arrangement of the person and object rather than adopting the person's perspective. GPT-4v responded “right” (i.e., the incorrect answer) on 88% of 50 trials using the image in Figure 4A and “right” (i.e., the correct answer) on 94% of the trials using the image in Figure 4B. This indicates that GPT-4v is largely dependent on visual representation rather than true perspective-taking. Interestingly, even when the image orientation and person's perspective align (as in the person walking away from the viewer, Figure 4B), GPT-4v still shows some errors. This may reflect a bias introduced by the training data where prompts that specify "from the person’s perspective" typically signal a deviation from the straightforward visual cues. Thus, although GPT-4v performs well when the perspectives align, the linguistic framing in the prompts appears to subtly influence its responses, underscoring a limitation in its ability to fully interpret perspective-based queries. Overall, this suggests that GPT-4v is unable to move beyond the literal representation of the image to engage in true perspective taking.

Figure 5

A similar challenge is presented in Figure 5, where GPT-4v is prompted to determine the direction that a person is looking from the person’s perspective: “*is the person looking to **their left or to their right**?*”. GPT-4v incorrectly infers the person’s gaze direction based on the 2D orientation of the image, for example, stating, “*The person in the image is looking to their right*,” again reflecting a bias towards the explicit image content rather than the implied perspective. This was experimentally tested in 30 distinct face images similar to Figure 5, all with clear eye gaze direction evenly distributed between leftward and rightward. GPT-4v had two attempts per image and had an overall accuracy of 45%. In addition, responses changed between attempts for 50% of the images underscoring the random nature of the responses.

Overall, these findings reveal a significant gap in GPT-4v’s image understanding. The inability to perform perspective taking compromises GPT-4v's understanding of spatial relationships in a context-dependent manner. This emphasizes a critical difference between current AI spatial representation and the inherently more flexible and context-sensitive processes characteristic of human cognition.

1.4. Discussion

This study demonstrates a fundamental difference in the spatial reasoning capabilities of GPT-4v compared with human cognition. GPT-4v excels at extracting detailed, narrative descriptions from images, demonstrating sophisticated perceptual reasoning. However, it struggles with representational manipulation, such as mental rotation tasks and perspective shifting, indicating a lack of an embodied, manipulable spatial representation. This reliance on propositional reasoning highlights both strengths and limitations. While it allows for strong performance in well-defined, logically deducible tasks, it fails to capture the intuitive and flexible thinking that characterizes human interaction with space. Particularly, in perspective-taking tasks, GPT-4v cannot move beyond the literal 2D image data to adopt a more context-dependent understanding.

The study also highlights how AI systems, despite advancements, differ significantly from human cognition in internal processing methods. Current benchmarks for spatial reasoning in AI [3], including the latest versions of Embodied Question Answering benchmarks [4], reveal that multimodal foundation models like GPT-4v significantly underperform compared to humans, especially in tasks requiring spatial manipulation or perspective-taking.

As AI continues to develop, it remains to be seen whether these systems can be trained or redesigned to genuinely understand and manipulate spatial representations analogously to humans, or if such capabilities are inherently biological, tied to physical experience. The implications of this research extend into the future trajectory of AI development. It may necessitate a paradigm shift from purely computational models to ones that integrate sensorimotor experiences, perhaps through advanced robotics or virtual simulations that allow AI to 'experience' space as humans do.

References

[1] Yin, R. K. (1969). Looking at upside-down faces. *Journal of experimental psychology*, *81*(1), 141.
[2] Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701-703.
[3] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-10).
[4] Majumdar, A., et al. (2024) OpenEQA: Emodied Question Answering in the Era of Foundation Models. https://open-eqa.github.io/assets/pdfs/paper.pdf