Why Multimodal Models Struggle with Spatial Reasoning: Insights from Human Cognition

Bridget Leonard (bll313@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Kristin Woodard (woodkm@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Scott O. Murray (somurray@uw.edu)
Department of Psychology, University of Washington, Seattle, WA 98195 USA

Abstract

Multimodal models excel in tasks requiring semantic integration of language and vision but struggle with spatial cognition. Using a visual perspective-taking task inspired by cognitive science, we find these models fail when the image and reference view differ, reflecting spatial cognition comparable to a two-year-old child. To explore these disparities further, we analyze internal representations using a human action fMRI dataset and voxelwise encoding models, revealing key differences between AI and human spatial encoding. This work provides new benchmarks and insights into bridging artificial and biological cognition.

Keywords: multimodal AI; spatial cognition; perspective-taking; fMRI; voxelwise encoding models

Introduction

Transformer-based large language models (LLMs) have revolutionized the field of artificial intelligence with their ability to exhibit human-like intelligence, excelling across a range of tasks. Many of these models incorporate vision capabilities, such as CLIP (Radford et al., 2021), enabling them to extract relevant information from visual inputs and integrate it into their text outputs. The performance of multimodal LLMs (MLMs) on challenging benchmarks has seen exponential growth, yet one critical aspect of cognition remains underdeveloped: spatial cognition.

With their foundations rooted in natural language supervision, MLMs succeed at grasping many complex concepts, tasks, and skills. However, as we propose in this paper, language alone may not be a sufficient carrier of spatial information. Although MLM representations are difficult to interpret directly, cognitive science and neuroscience offer valuable tools for probing 'black boxes' to reveal representational structure. In this paper, we draw on behavioral experiments and neuroimaging methods from cognitive science to investigate the spatial reasoning abilities of multimodal AI systems.

We propose that multimodal AI (MLMs) has effectively fused the functions of the human language network and the ventral visual stream, equipping it to handle complex tasks involving semantic and visual information. However, these models fail to emulate the dorsal stream's role in visual-motor integration, limiting their ability to use visual input for spatially guided action. This paper explores how this gap in spatial reasoning manifests in MLMs, why it exists, and how we can begin to bridge the divide between artificial and biological spatial cognition.

Background and Related Work

Two primary limitations appear within AI spatial cognition literature: 1) linguistic reasoning can inflate performance on spatial benchmarks, and 2) benchmark scores can be hard to interpret when models perform poorly. For example, text-only GPT-4 performs surprisingly well on Meta's openEQA spatial understanding task (Majumdar et al., 2024), suggesting that many real-world questions about visual scenes can be solved linguistically. Additionally, the limited improvement observed in the multimodal GPT-4v indicates that vision models add little to spatial reasoning beyond what language alone can infer.

On the BLINK benchmark (Fu et al., 2024), which focuses specifically on visual perception and includes spatial cognition tasks like relative depth and multi-view reasoning, GPT-4v performed only marginally better than random guessing and far below human accuracy. This highlights the significant limitations of MLMs in visuospatial tasks and suggests substantial advancements are needed for real-world reliability. However, pinpointing the reasons behind these failures remains challenging, as they cannot always be attributed to the absence of specific cognitive processes.

Visual Perspective-Taking

Here we focus on an established skill in cognitive science that reflects spatial cognition in a precise manner. *Visual perspective-taking*, or the ability to mentally simulate a viewpoint other than one's own, is a critical aspect of spatial cognition. It allows us to understand the relationship between objects and how we might have to manipulate a scene to align with our perspective, which is essential for tasks like navigation and social interaction.

In the human developmental literature, perspective-taking has been stratified into two levels. Level 1 refers to knowing that a person may be able to see something another person does not, and it appears fully developed by the age of two (Moll & Tomasello, 2006). A common Level 1 task might ask if an object is viewable (or positioned to the front or back) of a person or avatar in a scene. Level 2 refers to the ability to represent how a scene would look from a different perspective, often measured by having subjects assess the spatial relationship between objects. Although success on some simple Level 2 tasks is first seen around age 4 (Newcombe

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

& Huttenlocher, 1992), Level 2 perspective-taking continues to develop into middle childhood (Surtees & Apperly, 2012) and even into young adulthood (Dumontheil, Apperly, & Blakemore, 2010). Mental rotation underlies Level 2 tasks, as shown by increased response times with greater angular differences, unlike Level 1 tasks, which are angle-invariant (Surtees, Apperly, & Samson, 2013).

Cognitive Science-Inspired Benchmark: Perspective-Taking Task

Leveraging the distinction between Level 1 and Level 2 perspective-taking (Surtees et al., 2013), we created a small perspective-taking benchmark that assesses multimodal model capabilities across three tasks: Level 1, Level 2 with spatial judgments, and Level 2 with visual judgments. Level 1 trials also serve as a perceptual control to verify that models can accurately detect basic visual attributes, such as the person's facing direction and object position, before assessing higher-level spatial reasoning in Level 2 trials.

This benchmark aims to address gaps in current AI spatial cognition measures by increasing process specificity, limiting language-based solutions, and offering straightforward comparisons to human cognition.

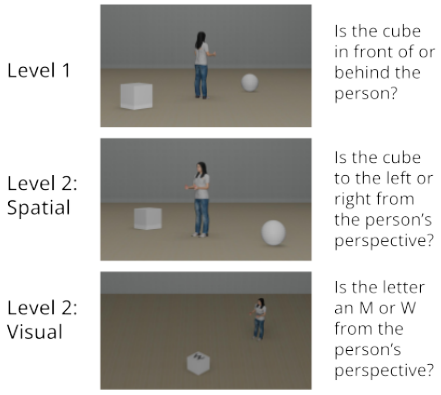


Figure 1: Examples of benchmark items.

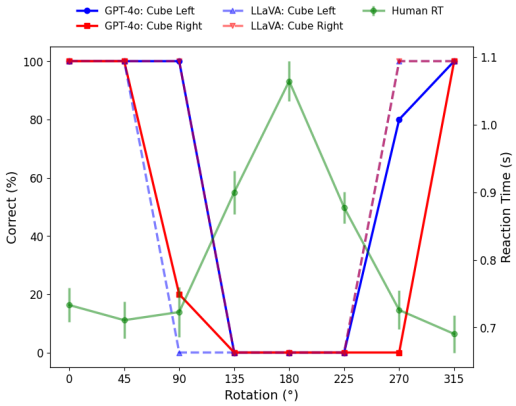
Our experimental design was inspired by previous studies that evaluated viewpoint dependence using targets like toy photographers (Frick, Möhring, & Newcombe, 2014) and avatars with blocks (Surtees et al., 2013). We used an avatar as a target and different stimuli, either cubes with numbers and letters or cubes and spheres, to investigate the influence of visual and spatial judgments on model performance. Each task consisted of 16 trial types, featuring images at 8 different angles (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°) with 2 response options for each task (e.g., cube in front or behind, 6/9 or M/W on the cube, and cube left or right). Examples of the stimuli are shown in Figure 1. All images, prompts, and evaluation scripts for the perspective-taking benchmark are available at `https://github.com/bridgetleonard2/PerspectiveTaking`

Results

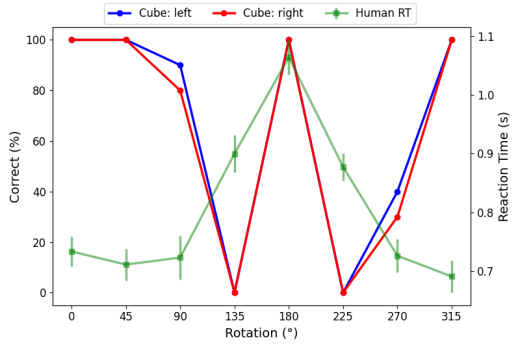
To examine the current state of perspective-taking in multimodal AI, we administered our benchmark to GPT-4o ("gpt-4o-2024-05-13" via OpenAI's API). Ten iterations of each image were passed through the model to calculate the percentage of correct responses as an accuracy measure.

As previously mentioned, human response times increase on perspective-taking tasks as the angular difference between the target and observer increases (Surtees et al., 2013). We administered the same task to a small number of human participants, part of a larger, IRB-approved study, and replicated this effect with both our stimuli types, finding a bell-shaped curve in the relationship between response time and angle. Human response times peaked when the target required a full mental rotation (180°), as seen in the green lines in Figure 2.

GPT-4o performed perfectly on Level 1 trials, confirming perceptual competence, but failed Level 2 tasks when humans engage mental rotation, beginning around a 90° angular difference (Figure 2a). We observed this behavior with both visual and spatial task stimuli.



(a) Without chain of thought



(b) With chain of thought

Figure 2: Level 2 spatial task performance. (a) Both GPT-4o and LLaVA exhibit strong egocentric biases, with sharp performance drops at 90° and 270° rotations. (b) GPT-4o shows improved spatial reasoning, aligning more closely with human-like performance.

To further examine how language might support spatial reasoning, we used chain-of-thought (CoT) prompting for the Level 2 task. After experimenting with several variations, we selected the following prompt based on its consistent success across trials:

Analyze this image step by step to determine if the cube is to the person's left or right, from the person's perspective. First, identify the direction the person is looking relative to the camera. Second, determine if the cube is to the left or right, relative to the camera. Third, if the person is facing the camera, then from their perspective, the cube is to the inverse of the camera's left or right. If the person is facing away from the camera, then the cube is on the same side as seen from the camera. Respond with whether the cube is to the person's left or right.

GPT-4o performance significantly improved with CoT prompting on 180° stimuli (Figure 2b). However, this linguistic strategy did not improve the model's ability to handle intermediate rotations between 90° and 180°. This suggests that while language can convey some level of spatial information, it lacks the precision required for human-level spatial cognition. This demonstration of surface-level perspective-taking abilities can partially explain how multimodal models achieve high performance on certain spatial benchmarks.

The Level 2 perspective-taking task reveals a key limitation in multimodal models' spatial reasoning abilities. While GPT-4o's performance declines on tasks typically solved by humans using mental rotation, this does not necessarily mean the model cannot perform such operations. Instead, it suggests that GPT-4o approaches these tasks using a different strategy, relying on image-based pattern recognition rather than simulating spatial transformations. This became more evident when we tested GPT-4o with visually rotated letter/number stimuli (e.g., M/W or 6/9), using open-ended prompts that did not specify which character to identify. The model often responded with "E" or "0" for inputs rotated by approximately $\sim$90°. These findings raise an important question: what internal representations are driving this performance, and how do they compare to the distinct representational structures of the human ventral and dorsal streams, particularly the dorsal stream's role in spatial reasoning?

Internal Representations in Models vs. Humans

Voxelwise encoding models require large, naturalistic datasets from a single participant, limiting options for spatial reasoning analyses. The best available option we know of is the Human Action Dataset (HAD) (Zhou, Gong, Dai, Wen, & Zhen, 2023), which is a large-scale fMRI dataset with voxelwise activations for 180 action categories. The videos in HAD depict dynamic, embodied actions involving body orientation, movement, and object interaction, offering spatially structured input suitable for probing dorsal stream processing. Moreover, the dataset includes predefined ROI masks for early visual (EV), ventral (VS), dorsal (DS), and lateral (LS) visual streams, allowing comparison across functionally distinct areas. However, one potential limiting aspect of HAD is that it is not directly a spatial cognition fMRI dataset so results are focused more on comparing model features to the dorsal and ventral streams rather than specifically investigating internal representations of spatial concepts. All analyses use data from subject 1 (sub-01).

To generate corresponding model features, we input video frames into the open-source multimodal model LLaVA 1.5 (13B) (Liu, Li, Li, & Lee, 2024), which performed comparably to GPT-4o on our perspective-taking task (Figure 1a), despite its smaller size. Features were extracted from three intermediate layers, corresponding to the vision tower, multimodal projector, and language model stages.

For each video, features from all frames were averaged to create a movie-level feature vector. These movie-level features were further averaged across videos corresponding to the same action category to generate category-level feature vectors. As a result, for each of the 180 action categories, we obtained both human brain data and model feature data, enabling direct comparison of how spatial and action-related information is represented in the two systems.

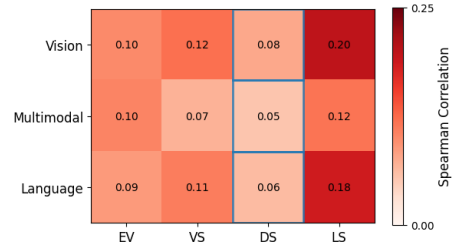
After ROI segmentation and feature extraction, we obtained fMRI betas from four ROIs (early visual, ventral, dorsal, and lateral streams) and model features from three model layers (vision tower, multimodal projector, and language model), each representing the 180 action categories.

Representational Similarity Analysis

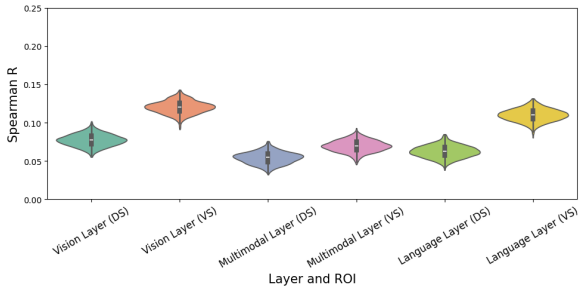
We began our analysis by calculating representation dissimilarity matrices (RDM) for each ROI and model layer, producing seven 180x180 matrices. Each RDM captures the pairwise dissimilarities between video stimuli based on voxel activity or model features, providing insight into how actions are organized within each representational space. We compared these matrices using Spearman correlation, which measures item-level relationships, and Procrustes disparity, which applies more complex transformations to assess structural similarity between matrices.

RDM correlations revealed relatively lower similarity between model features and the DS compared to other visual areas, particularly the VS (Figure 3a). This pattern was consistent across all three model layers. To assess the reliability of this effect, we used a bootstrapping approach to compare RSA correlations between DS and VS. These differences were statistically significant in all layers ($p < .0001$; Figure 3b), suggesting that model representations are more closely aligned with ventral stream organization than with the spatial and motor-relevant representations of the dorsal stream. Procrustes analysis confirmed this pattern, revealing higher structural disparity between model features and DS compared to other visual ROIs (Figure 4a). Bootstrapped comparisons again showed significant differences between DS and VS disparities across all layers ($p < .0001$; Figure 4b).

Although RSA and Procrustes differences are small, they may still capture meaningful distinctions in feature tuning across ROIs. For example, the lateral stream is more involved

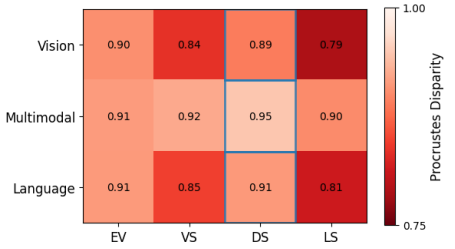


(a) Pairwise Spearman correlations

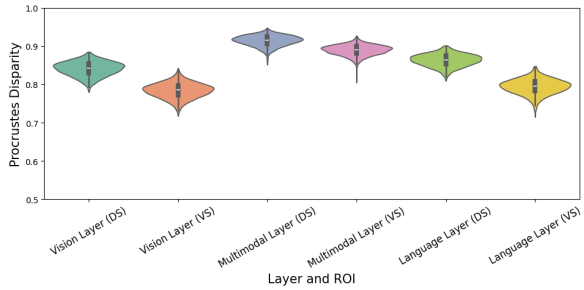


(b) Bootstrapped RDM correlations (DS/VS only)

Figure 3: RDM correlations between layers and ROIs. (a) Spearman correlations show stronger alignment between model features and ventral/lateral streams than with the dorsal stream. (b) Bootstrap results confirm significantly lower correlations for dorsal vs. ventral RDMs across layers ($p < .0001$).



(a) Pairwise Procrustes Disparity



(b) Bootstrapped RDS Procrustes Disparity (DS/VS only)

Figure 4: Procrustes Disparities between layers and ROIs. (a) Dorsal stream RDMs show higher disparity (lower similarity) with model features than ventral/lateral streams. (b) Bootstrap comparisons confirm significantly higher DS disparities across layers ($p < .0001$). *Note: Color scale reversed for comparison with Figure 3.*

in semantic processing, where the dorsal stream specializes in spatial information, offering a functional explanation for the observed differences. The lower similarity with early visual areas may also reflect architectural differences: unlike CNNs, LLaVA's transformer-based vision encoder likely prioritizes higher-order features over low-level cues like edges, resulting in closer alignment with later visual areas. These findings suggest that model features underrepresent spatial information processed in the dorsal stream, potentially explaining poor performance on spatial reasoning tasks. This may stem from the loss of spatial specificity when visual features are projected into language space; while the language model compensates semantically, its representations remain misaligned with spatial neural systems.

Voxelwise Encoding Models

To further investigate internal representations, we trained voxelwise encoding models to map the relationship between model features and voxelwise brain activity (la Tour, Eickenberg, Nunez-Elizalde, & Gallant, 2022). This approach allows us to evaluate how well information encoded in model features can predict the neural activity elicited by the same stimulus, offering a window into representational alignment between artificial and biological systems.

Given the limited number of samples in our dataset (180 action categories), we trained the encoding model on movie-level fMRI betas and corresponding features. For the selected HAD participant (sub-01), this yielded 720 training samples from individual video clips. For validation, we tested on the averaged category-level betas (180 samples). Although the validation set was not fully independent, our focus was on representational alignment rather than predictive accuracy.

We trained three separate encoding models using features from the vision tower, multimodal projection layer, and language model, respectively. We then evaluated validation performance across four ROIs (early visual, ventral, dorsal, and lateral streams), producing R^2 values for 12 conditions (3 layers × 4 ROIs). R^2 scores increased across layers (Figure 5), and dorsal stream performance was surprisingly comparable to the ventral stream, despite RSA suggesting weaker alignment.

To assess whether these differences were meaningful, we compared the distributions of R^2 scores between dorsal and ventral stream voxels in each layer. While all comparisons were statistically significant ($p < .0001$), effect sizes were small (Cohen's $d = 0.09$–0.30), and the distributions largely overlapped (Figure 6). These results suggest subtle but consistent shifts in model predictivity across regions, likely detectable due to the large number of voxels analyzed.

This finding contradicts our initial expectation and the results from the RSA but the dorsal stream's reliable prediction performance offers a more optimistic view. Encoding models perform a more complex transformation compared to RSA

or Procrustes, enabling them to "discover" relationships between features and brain activity that simpler similarity analyses might overlook.

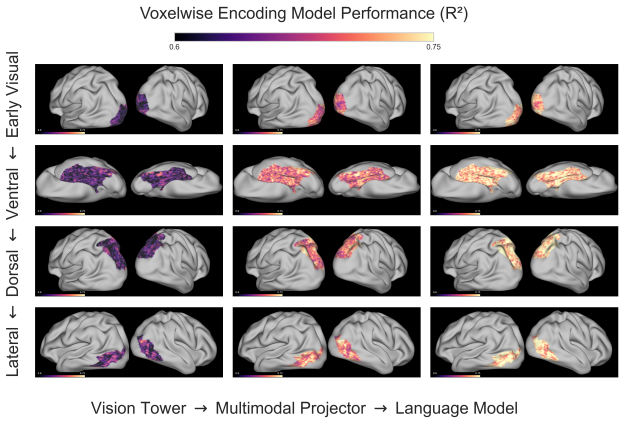


Figure 5: Voxelwise encoding model performance across hierarchies. Each brain map shows voxelwise R^2 for one ROI × model layer combination. Encoding improves with model depth, peaking in higher-order visual regions.

Encoding models essentially search for latent structures within the model's features that align with voxel activity, including features that might not dominate the model's output but are still present. While RSA and Procrustes analyses suggest weak structural alignment with dorsal stream representations, the encoding models' ability to recover predictive features implies that spatial information may still be present, although not prominently utilized, in these multimodal systems.

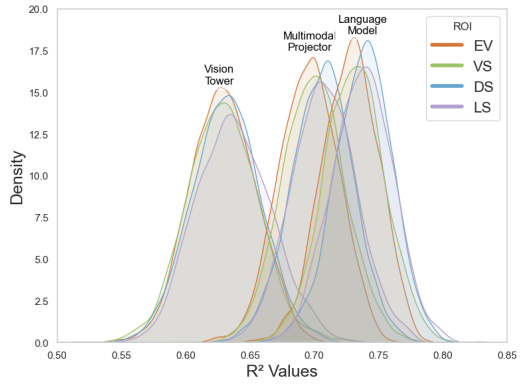


Figure 6: Figure 6: Distribution of voxelwise encoding model R^2 performance, showing increased accuracy with model depth. DS overlaps with VS and LS, suggesting relevant features are present but underutilized.

Encoding Model Weights

To further examine how different model features contribute to voxelwise encoding, we analyzed the learned feature weights from each encoding model. Specifically, we asked whether the dorsal stream relies on distinct subsets of model features compared to the more visual (early visual, lateral) and semantic (ventral) streams. Each encoding model (trained using features from a different layer of the multimodal model) produces a coefficient matrix that linearly maps features to voxel activity. These matrices have shape *(features* × *voxels)*, where each value represents the weight assigned to a feature for predicting activity in a specific voxel.

To assess regional patterns, we applied ROI masks to extract voxel subsets corresponding to each of the four visual streams. We then averaged the absolute feature weights across voxels within each ROI, resulting in a single vector of feature importance values per ROI. We focused on two analyses: (1) identifying the most relevant features for each ROI using either a top-n or threshold-based filtering approach, and (2) quantifying how many of these important features were shared across ROI pairs. This allowed us to test the hypothesis that the dorsal stream emphasizes a distinct subset of features, particularly those supporting spatial and action-relevant processing.

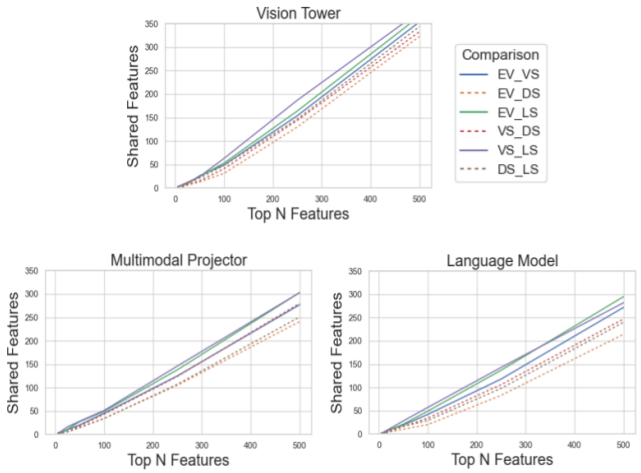


Figure 7: Top-*n* feature overlap between ROIs. Shared top-*n* features between ROI pairs. Dorsal stream comparisons (dashed) consistently show lower overlap than non-DS pairs across layers.

Top-*n* Approach To investigate whether the dorsal stream relies on distinct model features, we identified the top-*n* features with the highest weights for each ROI and measured how many of these were shared across ROI pairs. Since model features likely encode different aspects of visual or semantic information, we expected lower feature overlap between functionally distinct regions, particularly the dorsal stream, which is thought to prioritize spatial features. By varying n, this analysis allowed us to explore how the scale of feature importance influences shared feature distributions across regions.

As n increased, the dorsal stream consistently shared fewer

top-weighted features with other ROIs compared to the ventral stream, a pattern that appeared stable across all model layers (Figure 7). This difference was particularly evident in comparisons with early visual and lateral streams and was statistically significant or trending in all cases ($p < .05$ or marginal). When collapsing across all DS-involved and non-DS comparisons, dorsal stream pairings exhibited significantly lower feature overlap across all layers ($p < .05$), supporting the hypothesis that the dorsal stream draws on a more distinct subset of model features than other visual regions.

Thresholding Approach Since the top-n approach assumes that all ROIs rank feature importance similarly, we also applied a thresholding method to account for potential differences in feature weight distributions. Some ROIs may rely heavily on a few dominant features, while others may spread importance more evenly across many features. To capture this variability, we selected features exceeding a fixed weight threshold for each ROI and computed the percentage of shared features across all ROI pairs. Because the number of features retained varied by region, similarity was measured as the proportion of overlapping features. To prevent small sets from skewing results, comparisons were normalized: if one ROI retained only a single feature that was also present in another ROI, the overlap was set to 100%. Threshold ranges were tuned per layer to ensure meaningful comparisons, particularly for the vision tower, which required a distinct threshold range due to different weight distributions.

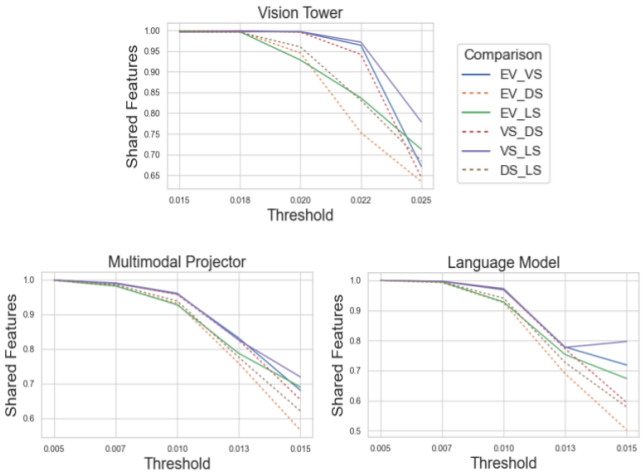


Figure 8: Threshold-based feature overall between ROIs. Proportion of shared features between ROIs after weight thresholding. Dorsal comparisons (dashed) show lower overlap, though not statistically significant.

In both the multimodal and language layers, comparisons involving the dorsal stream again tended to show lower overlap than other ROI pairs (see dashed lines in Figure 8), particularly at higher thresholds where retained features are more likely to be meaningful. However, unlike the top-n results, none of these differences reached statistical significance in any model layer. Greater variability in the number and distribution of high-weight features across ROIs may limit the stability and sensitivity of threshold-based comparisons.

These findings support our earlier interpretation of the encoding results: the features that best predict activity in the dorsal stream differ from those used by other visual areas. While all encoding models performed well across ROIs, the features driving this performance varied, with the dorsal stream showing the most distinct pattern. This suggests that spatially relevant information may be embedded within model features, but not naturally emphasized during standard model operation.

This preliminary analysis highlights the potential of weight-based interpretability for uncovering region-specific feature use. Future work could build on this approach using techniques like sparse autoencoders to isolate and amplify spatial components, offering a path toward improving multimodal model performance on spatial reasoning tasks.

Discussion

This study reveals critical limitations in the spatial reasoning capabilities of multimodal models, highlighting a stark contrast with human cognitive processes. While these models excel at tasks requiring the semantic integration of language and vision, they struggle with spatial tasks such as perspective-taking and visuospatial reasoning. In particular, models fail on perspective-taking tasks that require mental rotation, a cognitive process that relies on motor planning computations performed outside the ventral visual stream and language regions. Furthermore, when examining the internal representations of human actions, model features show greater similarity to representations in the ventral visual stream, while under-representing the dorsal stream, which is essential for vision-for-action and spatial understanding.

Interestingly, encoding models reveal that spatial information may exist in model features, but remains underutilized during standard inference. Further investigation of these relevant feature components could not only enhance spatial reasoning capabilities in models but also uncover new insights into spatial cognition more broadly. Exploring spatial cognition in humans, particularly within the perspective-taking network, could reveal critical computations that underpin spatial reasoning, offering valuable guidance for the development of future multimodal AI systems.

Leveraging cognitive science-inspired benchmarks and investigating internal representations can identify the shortcomings of MLMs and guide future advancements. The integration of spatially relevant computations, such as those found in the dorsal stream and perspective-taking network, holds promise for enhancing model performance on spatial tasks. Bridging the divide between biological and artificial intelligence may improve AI systems while simultaneously deepening our understanding of spatial cognition and its underlying neural mechanisms.

References

Dumontheil, I., Apperly, I., & Blakemore, S. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*(2), 331-338.

Frick, A., Möhring, W., & Newcombe, N. (2014). Picturing perspectives: development of perspective-taking abilities in 4- to 8-year-olds. *Frontiers in Psychology*, *5*.

Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., ... Krishna, R. (2024). *Blink: Multimodal large language models can see but not perceive.*

la Tour, T. D., Eickenberg, M., Nunez-Elizalde, A., & Gallant, J. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. doi: https://doi.org/10.1016/j.neuroimage.2022.119728

Liu, H., Li, C., Li, Y., & Lee, Y. (2024). *Improved baselines with visual instruction tuning.* Retrieved from `https://arxiv.org/abs/2310.03744`

Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., ... Rajeswaran, A. (2024). Openeqa: Embodied question answering in the era of foundation models. In *Conference on computer vision and pattern recognition (cvpr).*

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603-613.

Newcombe, N., & Huttenlocher, J. (1992). Children's early ability to solve perspective-taking problems. *Developmental psychology*, *28*(4), 635.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, *abs/2103.00020*.

Surtees, A., & Apperly, I. (2012). Egocentrism and automatic perspective taking in children and adults. *Child development*, *83*(2), 452–460.

Surtees, A., Apperly, I., & Samson, D. (2013). Similarities and differences in visual and spatial perspective-taking processes. *Cognition*, *129*(2), 426-438.

Zhou, M., Gong, Z., Dai, Y., Wen, Y., & Zhen, Z. (2023). *"a large-scale fmri dataset for human action recognition"*. OpenNeuro. doi: doi:10.18112/openneuro.ds004488.v1.1.0