COS 597B Course Project: Do Large Multimodal Models Possess Human-Like Geometrical Intuition?

Anirudh Ajith
Department of Computer Science
Princeton University

Abstract

Recent Large Multimodal Models (LMMs) have shown impressive results on visual question-answering (VQA) benchmarks demonstrating a high degree of understanding of visual and textual inputs. However, this work demonstrates that there is a significant gap between the mechanisms underlying these systems and those behind human visual perception. I demonstrate through a series of experiments involving prompting LMMs to answer questions about simple geometric properties of synthetic images, that these models do not possess the same geometric intuitions that are innate in humans. Their inability to identify geometric properties that would prove trivial for humans, when contrasted against their remarkable performance on VQA benchmarks consisting of more natural inputs reveals that their inductive biases, and the abstractions they form differ from those underlying human cognition.

Introduction

The ability to understand visual stimuli to make sense of one's environment is a key aspect of human cognition. This innate skill that even infants possess plays a crucial role in how we navigate the world around us. Humans show the ability to parse not only visual stimuli of natural origin, but also more abstract shapes, patterns and their geometrical attributes.

Recent months have seen the rise of powerful deep-learning based Large Multimodal Models (LMMs) that are pretrained on massive amounts of visual and textual inputs scraped from the internet. The recent public releases of instruction-tuned variants of these models such as LLaVA (H. Liu, Li, Wu, & Lee, 2023) and InstructBLIP (Dai et al., 2023) make it possible to non-invasively study the degree to which they understand various visual and textual inputs. While these models have demonstrated remarkable performance on visual question-answering benchmarks such as VQAv2 (Goyal, Khot, Summers-Stay, Batra, & Parikh, 2017) and MMBench (Y. Liu et al., 2023), more fine-grained analyses into the mechanisms behind these abilities have not yet been thoroughly conducted.

In this work, I perform a series of experiments testing the abilities of 2 popular contemporary LMMs to answer questions about simple geometrical properties of synthetically generated images. In particular, the experiments I perform testing their understanding of 1) the relative sizes of entities represented in an image, 2) the multiplicities of these represented entities, and 3) geometric properties such as angles and symmetry. I carefully select tasks that would be intuitively solvable by humans with no formal training, and most likely even by infants.

My findings demonstrate that contemporary LMMs are incapable of solving most of these tasks reliably, indicating that the mechanisms underlying their understanding of visual stimuli must differ significantly from those that drive human visual understanding.

Background

Visual perception in humans

Humans possess an innate ability to parse and interpret geometric abstractions despite evolutionary pressures only explicitly incentivizing accurate understanding of visual stimuli from the natural world. This capability, encompassing the ability to comprehend spatial relationships, perform size comparisons, assess symmetry and other abstract properties, reflects a fundamental aspect of human cognition.

Prior studies have shown that the ability to detect regularities in presented shapes appears innate in several human populations around the world including uneducated members of the Namibian Himba tribe and French toddlers (Sablé-Meyer et al., 2021). Some work (Placi, 2023; Sablé-Meyer et al., 2021) has also shown that these intuitions about Euclidean geometry seem innate to humans and appear absent from other animals such as baboons that can still adeptly parse visual stimuli from the natural world. (Spelke & Kinzler, 2007) which discusses the core knowledge systems in human cognitive development, including the understanding of geometry, suggests that humans are born with an inherent ability to perceive and understand geometric information, which is further honed through interaction with the environment. This concept is supported by research showing that even infants have a basic grasp of geometric principles, as seen in their ability to discriminate between different shapes and spatial layouts (Izard, Sann, Spelke, & Streri, 2009).

Large Multimodal Models

Recent years have seen the rise of deep-learning based AI systems called "foundation models" which are trained using large volumes of data scraped from the internet and show impressive (sometimes even human-level) performance on many kinds of tasks that were classically challenging for AI systems (Brown et al., 2020). Such abilities have prompted efforts into studying these foundation models not only through

a lens of machine-learning interpretability, but also using tools from cognitive psychology (Binz & Schulz, 2023). It has been suggested that the performance of these models on tasks, albeit occasionally astonishing, correlate strongly with the frequency with which these tasks are represented on the internet and do not indicate general problem-solving ability (McCoy, Yao, Friedman, Hardy, & Griffiths, 2023) or human levels of understanding.

Recent research into artificial intelligence has given rise to a class of deep-learning based systems known as Large Multimodal Models (LMMs). In particular, recent months have seen the public release of instruction-tuned LMMs such as GPT-4V (OpenAI, 2023), LLaVA (H. Liu et al., 2023), InstructBLIP (Dai et al., 2023) and Qwen (Bai et al., 2023) which can process both image and language data and can be prompted using interactive chat-based interfaces. These releases enable exciting opportunities to study LMMs using tools from cognitive psychology and interpretability.

Approach

In this work, I exploit the ability to prompt contemporary instruction-tuned LMMs using arbitrary visual and textual inputs to assess their understanding of the geometric properties of synthetically generated images. Specifically, while prior work performing assessments of image understanding in vision models may have been forced to resort to invasive methods involving studying activations and latent spaces, I adopt a more grounded approach by posing carefully chosen questions imposed in the form of textual prompts directly to the LMMs. Such an investigation has the added advantage of accurately representing models' abilities in practical user-interaction scenarios.

I performed experiments to evaluate the geometric understanding of LMMs along 4 axes:

1. The ability to ordinally compare lengths or areas of lines and simple shapes in input images.

2. The ability to correctly count the number of lines or shapes depicted in input images.

3. The ability to correctly predict the types of angles depicted in input images.

4. The ability to correctly identify line and rotational symmetries in input images.

In order to perform these analyses, I created 11 datasets of simple synthetic images consisting of line segments, circles, squares and triangles of various sizes such that each dataset helps test the models along one of the aforementioned axes. Each experiment involved prompting the LMMs with each of the images in the relevant datasets along with an appropriate question about the image asking the model to perform a classification.

Methods

In my experiments, I restricted myself to studying the abilities of 2 mainstream open-source instruction-tuned LMMs called LLaVA-1.5 7B (H. Liu et al., 2023) and InstructBLIP 7B (Dai et al., 2023) (each consisting of 7 billion parameters). Although both of these model families include a larger 13 billion parameter variant, those were too large for me to use without quantization under my compute-constratints.[1] For each dataset, I created square images of side length 800 pixels showing black, blue, red or green figures against a white background using the `matplotlib` (Hunter, 2007) Python package. Each experiment involved prompting the LMMs with a single image and asking a question relevant to the experiment being conducted. I used a greedy decoding strategy to obtain model predictions in all experiments to minimize experimental stochasticity and evaluate the model's most confident prediction. I provide the prompts I use for all the experiments I conduct in Table 1.

Size-comparison ability

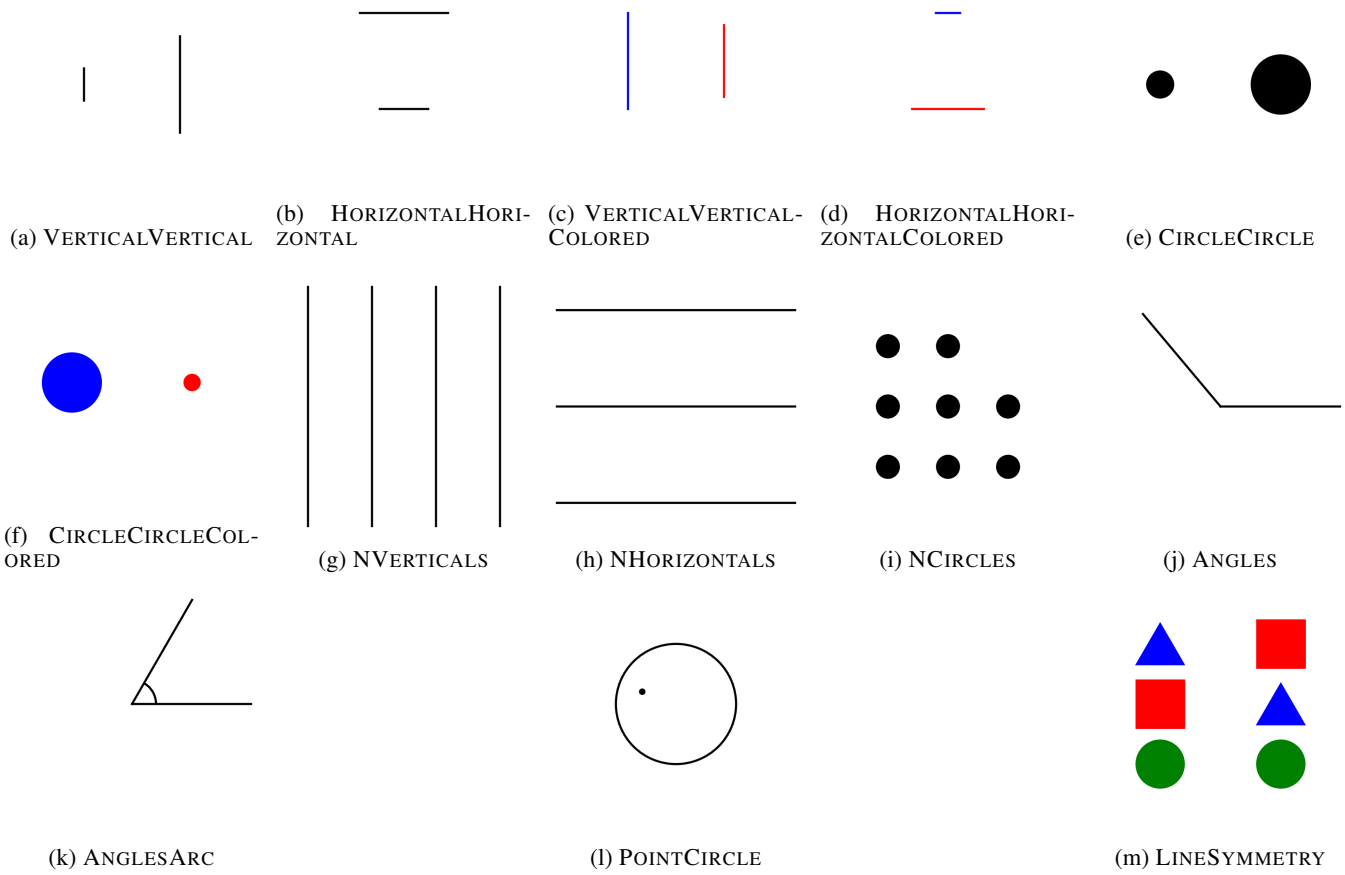
I created the VERTICALVERTICAL and HORIZONTALHORIZONTAL datasets consisting respectively of pairs of solid black vertical and horizontal line segments. The former consisted of parallel vertical line segments of (possibly) differing lengths while the latter consisted of analogously parallel horizontal line segment pairs. In order to ascertain the effect of color on models' abilities to perform these tasks, I also created similar VERTICALVERTICALCOLORED and HORIZONTALHORIZONTALCOLORED datasets where these each line segment was differently colored. The COLORED datasets allow us to assess if these models are able to more accurately answer size-comparison questions when the line-segments are referred to by color, rather than by their location in the image. In addition, I also created the datasets CIRCLECIRCLE and CIRCLECIRCLECOLORED consisting of pairs of differently sized circles to understand if the models showed a higher degree of understanding of simple geometrical shapes, than that of line segment lengths. These experiments involved prompting the LMMs with queries asking them to identify which of the two alternatives was longer/larger.

For each of the VERTICALVERTICAL, HORIZONTALHORIZONTAL, VERTICALVERTICALCOLORED and HORIZONTALHORIZONTALCOLORED datasets, I allowed the lengths of the line segments to vary uniformly from 80 pixels long to 320 pixels long creating 100 images in each dataset. I used the colors blue and red in the COLORED datasets. The circles in the CIRCLECIRCLE and CIRCLECIRCLECOLORED datasets varied in radius from 40 to 200 pixels and occupied the left and right halves of the image.

Counting ability

I created datasets NVERTICALS, NHORIZONTALS and NCIRCLES consisting of various numbers of parallel verti-

[1] I have seen much discussion in forum posts on the internet, that quantization especially hurts performance in multimodal models.



(a) VERTICALVERTICAL (b) HORIZONTALHORIZONTAL (c) VERTICALVERTICALCOLORED (d) HORIZONTALHORIZONTALCOLORED (e) CIRCLECIRCLE

(f) CIRCLECIRCLECOLORED (g) NVERTICALS (h) NHORIZONTALS (i) NCIRCLES (j) ANGLES

(k) ANGLESARC (l) POINTCIRCLE (m) LINESYMMETRY

Figure 1: Examples of images in synthetic datasets I create to assess models' ability to perform size-comparison, counting, angle-identification, and symmetry-identification.

cal lines, parallel horizontal lines and equally sized circles arranged uniformly over the image. This experiment involved prompting the LMMs to count the number of distinct lines or circles they were able to identify in the image.

NVERTICALS and NHORIZONTALS consisted of equally spaced parallel line segments occupying the full extent of the image. NCIRCLES consisted of 10 images of equally spaced solid black circles of radius 80 pixels.

Angle-identification ability

I created the ANGLES dataset of images, each consisting of two solid black line segments that are inclined at a specific angle to each other. I also created the ANGLESARC dataset where a circular arc was used disambiguate the included angle being referred to by the prompt. In these experiments, I prompted the models to identify each represented angle as either acute, right or obtuse.

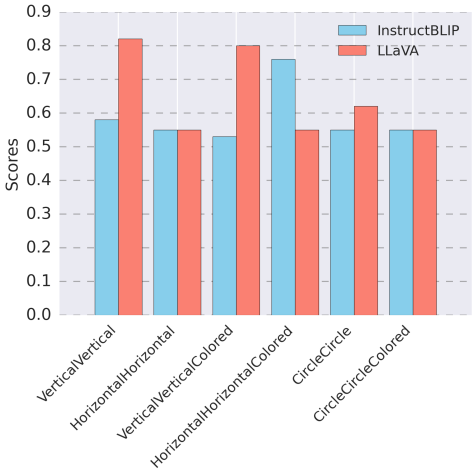
ANGLES and ANGLESARC consisted of angles represented at 10 degree increments from 0° to 180°.

Symmetry-identification ability

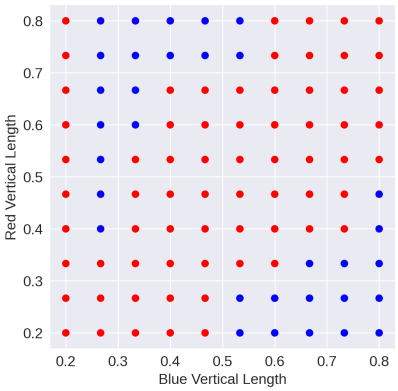
I evaluated models' abilities to identify line symmetry by creating a dataset called LINESYMMETRY consisting of sets of colored shapes (circles, triangles, squares) stacked atop one another on both the left and right halves of the image. These images were such that the stack on the left could either be identical or distinct from that on the right, thereby corresponding to images that respectively either possessed or did not possess line symmetry along a vertical axis. This experiment involved prompting the models to correctly classify the images as either symmetric or asymmetric.

I adopted a more implicit approach for evaluating the models' perception of rotational symmetry. I created a dataset called POINTCIRCLE consisting of images with a circular outline and a point placed somewhere on the image. In these experiments, I prompted the models to say whether the point was at the center of the circle. If the models could correctly perceive rotational symmetry, they would answer affirmatively only when the point was indeed at the circle's center.

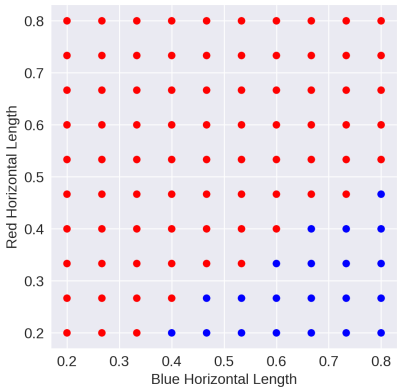
POINTCIRCLE consisted of 91 images of a centered circu-



(a) Accuracies



(b) InstructBLIP predictions on VERTICALVERTICALCOLORED



(c) InstructBLIP predictions on HORIZONTALHORIZONTALCOLORED

Figure 2: Results from experiments studying size-comparison ability. The models usually seem incapable of correctly identifying the larger entity. However, they seem to show the ability to tell when when one object is much larger than the other, even if they do not retain the larger object's identity.

lar outline of radius 200 pixels and the location of the point laid at a uniformly chosen radial distance of 0 to 360 pixels from the circles center and a uniformly angle with the horizontal. LINESYMMETRY consisted of stacks of colored shapes including all possible stacks of size 1 and 2, 50 randomly sampled asymmetric stacks of size 3, and 50 randomly sampled symmetric stacks of size 3.

Results

Size-comparison ability

Both InstructBLIP and LLaVA only show near random chance accuracy on most size-comparison datasets (Figure 2a). However, there are a few outlier experiments (such as InstructBLIP on HORIZONTALHORIZONTALCOLORED and LLaVA on VERTICALVERTICAL and VERTICALVERTICALCOLORED) where these models unexpectedly show > 0.75 classification accuracy. Empirically, I notice that the predictions in most experiments where the models show near-random performance qualitatively resemble Figure 2b where the model's prediction appears conditioned on the absolute difference between the line segment lengths, rather than whether one of them is indeed longer/larger. The predictions obtained in the outlier experiments typically resemble Figure 2c where the model sticks to predicting a single label unless the evidence to the contrary is extremely compelling.

Counting ability

Figure 3 shows that these models do indeed show the ability to obtain approximate counts of the number of represented objects. Although not completely accurate, the models' predictions correlate with the ground truth on all three counting tasks: NVERTICALS, NHORIZONTALS and NCIRCLES.

Angle-identification ability

As seen in Figure 4, both models perform exceedingly poorly on ANGLES and ANGLESARC. Their predictions do not even appear to correlate with the ground truth labels. LLaVA fails completely on both tasks predicting 'acute' for all images in the datasets. InstructBLIP appears heavily biased towards predicting 'right' even at very small or very large angles.

Symmetry-identification ability

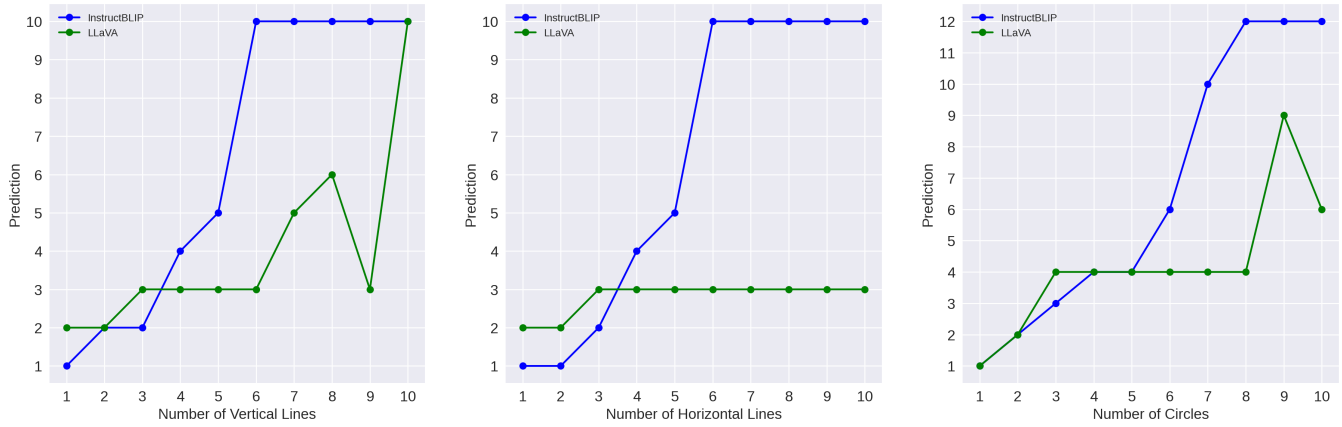
It is clear from Figure 5a that both models show near random-chance performance on LINESYMMETRY at all 3 stack sizes. Empirically, I observe that the models tend to predict that the figure is symmetric in the vast majority of cases (even when the figure is actually asymmetrical).

On the POINTCIRCLE task, the models fail spectacularly at assessing whether the point lies at the center of the circle. However, InstructBLIP's predictions shows an interesting correlation with whether the point lies in the interior of the circle (Figure 5b). On the other hand, Figure 5c shows that LLaVA does not show any understanding of the geometry of these figures and consistently predicts 'yes'.

Discussion

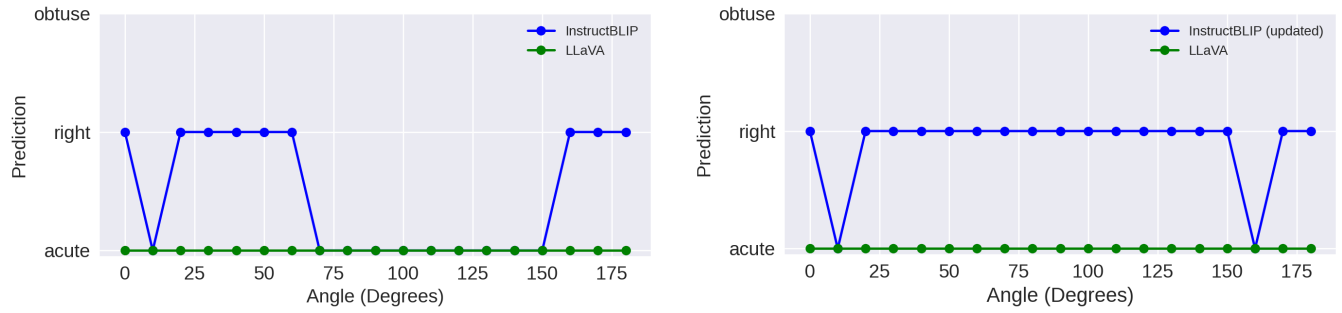
Both LLaVA and InstructBLIP fail to perform well on the majority of tasks I study.

They usually display only near-random chance accuracy on size-comparison tasks. However, they do display some non-trivial understanding of geometry in these situations as evidenced by predictions such as those represented in Figure 2b. There is clearly an approximate ability to ascertain if one entity is **significantly** larger than the other even though do not always retain the identity of the larger entity. It is possible that the inability to perform well on these tasks despite this stems from the loss of positional information from the



(a) NVERTICALS (b) NHORIZONTALS (c) NCIRCLES

Figure 3: Results on counting ability datasets. The models show a reasonably good ability to count the number of represented objects.



(a) ANGLES (b) ANGLESARC

Figure 4: Results on angle identification datasets. The models fail dramatically at categorizing an angles as 'acute', 'right' or 'obtuse'.

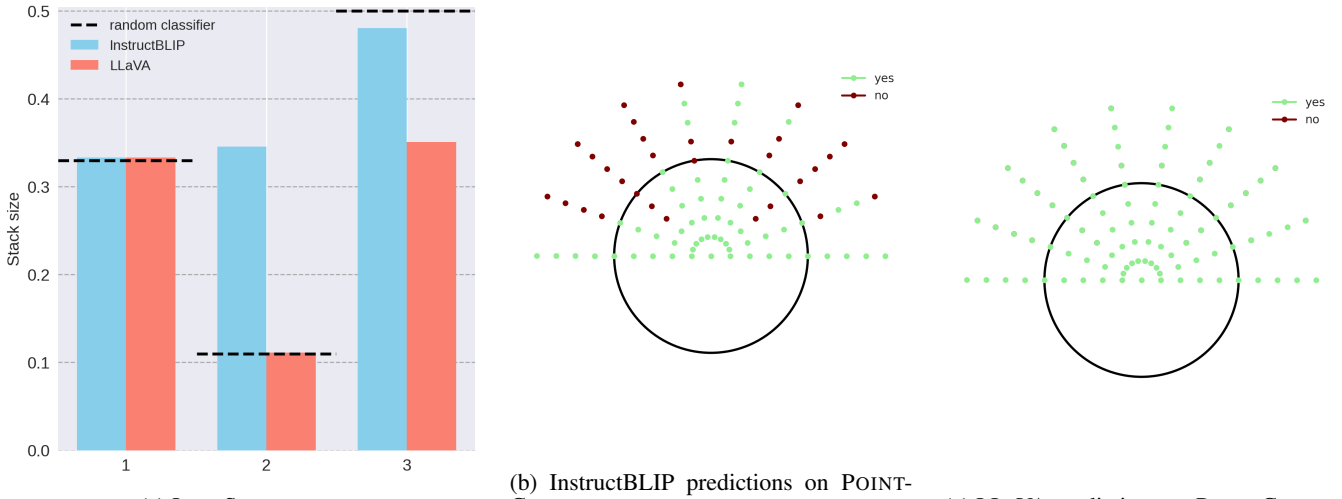
visual inputs within the models. It is also evident that they are typically unable to maintain the link between represented entities and their colors, as evidenced by low accuracies even on the VERTICALVERTICALCOLORED, HORIZONTALHORIZONTALCOLORED and CIRCLECIRCLECOLORED tasks.

Their performance on the tasks involving counting entities (NVERTICALS, NHORIZONTALS and NCIRCLES) is surprisingly good, with their predictions correlating well with the ground truth labels in all 3 experiments. It is possible that their ability to count multiplicities of natural entities like birds (as often stated in internet image captions) has generalized to enable them to count abstract geometrical constructions too.

These models however, completely fail at predicting whether angles are 'acute', 'right' or 'obtuse' as seen in Figure 4. These models seem heavily biased towards predicting 'acute' and 'right' and never predict 'obtuse' in my experiments. It is likely that this is due to more frequent occurrences of 'acute' and 'right' angles being depicted in online discourse, as compared to 'obtuse' angles. In any case, the models demonstrate no understanding of the relative sizes of angles.

The models also fail dramatically at assessing symmetry as evidenced by Figure 5. The observation that they tend to predict that images are symmetrical even when they aren't can be explained by the fact that most instances of symmetry being discussed online would most likely be positive instances. I expect that it would be rare for online discourse about asymmetry to be as common since most real-world images are asymmetrical making this an uninteresting attribute of theirs to bring up. Although the models also fail to correctly assess rotational symmetry in the POINTCIRCLE task, they show an interesting ability to approximately discern the interior of a circle from its exterior. This reveals that they indeed possess some understanding of these geometrical concepts.

In summary, InstructBLIP and LLaVA only demonstrate near random-chance accuracy in most of the geometric understanding tasks I pose. However, there are some inklings of geometric understanding that have emerged (specifically, the effects depicted in Figures 2b, 3 and 5b). While these multimodal models as a whole do not possess the robust geometric understanding that humans innately do, these inklings



(a) LINESYMMETRY

(b) InstructBLIP predictions on POINTCIRCLE

(c) LLaVA predictions on POINTCIRCLE

Figure 5: Results from experiments studying symmetry detection. The models fail to correctly assess symmetry but sometimes show some understanding of other geometric properties.

hint that this may stem from insufficient alignment between the parts of their architectures devoted to visual and verbal processing. I hypothesize that their remarkable performance on standard VQA benchmarks indicates that the representations computed by these modules are well-aligned for natural (visual and textual) inputs since these resemble their pretraining data. However, abstract images and questions like the ones I pose in this work likely constitute a minority in their pretraining data. This could potentially explain their seeming tendency to perform interesting geometrical computations which differ from what the ones they were asked to. If this were true, it may be possible to get these LMMs to show high performance on these synthetic tasks with a small amount of additional fine-tuning on such synthetic inputs. This is an interesting avenue for a followup study.

Conclusion

I find that contemporary pretrained LMMs of the scale I study do not possess the same geometric intuitions as humans. This is interesting since humans too predominantly encounter only natural visual and verbal stimuli over the course of their cognitive development. The evidence in gathered in this study points to the conclusion that the inductive biases and underlying abstractions driving human visual perception differ significantly from those that exist in contemporary LMMs.

References

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., … Zhou, J. (2023). *Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.*

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. doi: 10.1073/pnas.2218523120

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). *Language models are few-shot learners.*

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., … Hoi, S. (2023). *Instructblip: Towards general-purpose vision-language models with instruction tuning.*

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017, July). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009, June). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences*, *106*(25), 10382–10385. Retrieved from http://dx.doi.org/10.1073/pnas.0812142106 doi: 10.1073/pnas.0812142106

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In *Neurips.*

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., … Lin, D. (2023). *Mmbench: Is your multi-modal model an all-around player?*

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of autoregression: Understanding large language models through the problem they are trained to solve.*

OpenAI. (2023). *Gpt-4v(ision) system card.* Retrieved from https://cdn.openai.com/papers/GPTV_System_Card.pdf

Placi, S. (2023, March). Sensitivity to geometry in humans and other animals. *In&V*, *1*(1), 33–54.

Dataset	Prompt
VERTICALVERTICAL	Which of these two lines is longer? Left or right?
HORIZONTALHORIZONTAL	Which of these two lines is longer? Top or bottom?
CIRCLECIRCLE	Which of the two circles in this image is bigger? Left or right?
VERTICALVERTICALCOLORED	Which line is longer in this image? The blue one or the red one?
HORIZONTALHORIZONTALCOLORED	Which line is longer in this image? The blue one or the red one?
CIRCLECIRCLECOLORED	Which of the two circles in this image is bigger? Red or blue?
NVERTICALS	How many vertical black lines do you see in this image?
NHORIZONTALS	How many horizontal black lines do you see in this image?
NCIRCLES	How many circles do you see in this image?
ANGLES	Is this acute, right, or obtuse?
ANGLESARC	Is this acute, right, or obtuse?
LINESYMMETRY	Is this figure symmetrical?
POINTCIRCLE	Is the point at the center of the circle?

Table 1: The prompts I use for conducting each experiment.

Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021, April). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proc. Natl. Acad. Sci. U. S. A.*, *118*(16), e2023123118.

Spelke, E. S., & Kinzler, K. D. (2007, January). Core knowledge. *Developmental Science*, *10*(1), 89–96. Retrieved from http://dx.doi.org/10.1111/j.1467-7687.2007.00569.x doi: 10.1111/j.1467-7687.2007.00569.x