Evaluating Vision Language Models: Challenges in Symmetry and Size Perception

Salvador Robles
The University of Texas at Austin
salvadorrh@utexas.edu

Eshan Balachandar
The University of Texas at Austin
eshan@cs.utexas.edu

Abstract

Vision-Language Models (VLMs) have shown remarkable generalization across a wide range of multimodal tasks. However, despite their impressive performance, they exhibit consistent failures in spatial reasoning tasks, particularly those involving symmetry and size comparison.

In this work, we conduct a targeted evaluation of state-of-the-art VLMs on simple yet fundamental spatial understanding benchmarks. Our experiments reveal that current models frequently misidentify left-right orientation and struggle with relative size perception, even under controlled conditions.

We evaluate and compare the performance of models including LLaVA-1.5 (7B and 13B), Claude-3.5 Sonnet, and GPT-4o on these tasks. Our findings highlight the need for new training objectives or architectures that better incorporate geometric and spatial relationships.

1 Introduction

Vision-Language Models (VLMs) like LLaVA, InstructBLIP, and GPT-4o have made huge progress in tasks such as Visual Question Answering (VQA), image captioning, and general visual reasoning. However, despite their impressive generalization, there are still simple tasks where they surprisingly fall short. Prior work, including Eyes Wide Shut (Tong et al., 2024) and Winoground (Thrush et al., 2022), has already pointed out that VLMs can struggle with spatial understanding.

Most VLMs today are trained using contrastive learning objectives or instruction tuning over large multimodal datasets. These training approaches help the models develop strong generalization skills across a wide range of tasks. However, they often emphasize matching high-level semantics (e.g., "a dog on a couch") rather than fine-grained spatial details. As a result, models may miss small but important visual differences, such as which side of an image an object appears on, or which item is slightly larger. This makes them prone to errors in tasks that require precise spatial understanding.

Recent work like Do Vision-Language Models Represent Space and How? (Zhang et al., 2025) shows that VLMs often struggle when asked to reason about spatial relationships, especially when there's some ambiguity in the image. For example, terms like "left" or "behind" can be interpreted differently depending on whether the model uses an egocentric (viewer-based) or object-centric perspective. In this project, we focus on investigating two specific types of spatial reasoning: symmetry and size perception.

For symmetry, we explore what happens when we flip an image horizontally and ask the same questions as before. Ideally, the answers should remain consistent, but we often find that model outputs change or degrade. This raises the question: are these models actually learning spatial relationships, or are they relying on shortcuts or biases in the training data?

For size perception, we generate synthetic images where the ground truth about object sizes is known. We then test whether VLMs can answer simple questions like "Which object is bigger?" Again, we find that many models make mistakes even in controlled settings.

We are interested in whether some models do better than others, and what this tells us about their architecture or training objective. Our research questions are:

- **RQ1:** Are some model architectures consistently better at spatial reasoning tasks like symmetry and size perception?

- **RQ2:** How well can VLMs judge relative size in synthetic but controlled images?

2 Benchmark Design

Our goal is to evaluate the ability of Vision-Language Models (VLMs) to perform basic spatial reasoning tasks. While these models excel at image captioning and high-level visual descriptions, they often struggle with grounded spatial relationships. We focus on two fundamental benchmarks that probe different aspects of visual spatial understanding: **symmetry** and **size perception**.

2.1 Symmetry Reasoning

Our goal is to test whether Vision-Language Models (VLMs) can handle basic spatial reasoning tasks, especially involving symmetry and left–right relationships. To do this, we design a series of small, focused tasks using real-world and synthetic variations of images from the COCO 2017 validation set. Each task is centered around how a model behaves when an image is flipped horizontally, a simple transformation that should maintain the objects in the scene but alter their spatial relationships.

- **Symmetry Detection:** Determine whether an image is visually symmetrical.
- **Object List Consistency:** Verify whether the model's object detections are stable across horizontal flips.
- **Left/Right Reasoning:** Check whether spatial relationships (e.g., "the dog is to the left of the couch") are updated appropriately after flipping the image.

To further stress-test models, we vary the prompt framing across two perspectives:

- **Viewer-Centric Prompts:** The model is asked to respond as a human viewer would.
- **Ego-Centric Prompts:** The model is instructed to imagine itself as one of the objects in the scene.

This lets us evaluate not just spatial accuracy but also frame-of-reference ambiguity, which has been mentioned in recent works such as (Zhang et al., 2025).

These tasks probe not just the model's visual capabilities, but also its implicit assumptions about viewpoint, grounding, and spatial language.

2.2 Size Perception

Our second benchmark isolates size comparison in controlled synthetic images. Each image contains two geometric objects (e.g., a green circle and a red square) of varying sizes, with known ground truth attributes. This setup lets us precisely evaluate whether a model can identify the larger object using questions like:

- "Which object is larger: the one on the left or the one on the right?"
- "The larger object has which color?"
- "Which object is larger: the square or the circle?"

By removing the confounds of real-world image noise, these tests reveal whether models actually perceive spatial magnitude or are relying on shallow cues like color and label associations.

Together, these two benchmarks provide complementary lenses into how—and whether—VLMs understand spatial geometry.

3 Models

To evaluate how well different Vision-Language Models handle symmetry and spatial reasoning, we tested a range of models, from open-source to commercial state-of-the-art systems:

1. **LLaVA-1.5-7B:** This is a popular open-source model that combines a vision encoder with a language model. It performs well on general VQA tasks and is relatively lightweight, which makes it easy to run for large-scale evaluation.
2. **LLaVA-1.5-13B:** A larger version of the 7B model, this one usually gives better performance thanks to its increased capacity. It's still based on the same architecture and training objective but captures more visual and textual patterns.
3. **GPT-4o:** This is OpenAI's most recent vision-language model, and it currently leads many multimodal benchmarks. While the training details are not public, we suspect it follows a multi-stage pipeline—likely involving some object detection system followed by a reasoning module. We accessed GPT-4o through OpenAI's API.

4. **Claude 3.5 Sonnet:** After observing that many models struggled with left-right consistency and spatial relationships, we tested Claude 3.5 Sonnet, which is claimed to be more robust in those areas. It's a commercial model from Anthropic that supports vision input and is designed for stronger reasoning performance.

4 Symmetry Benchmarking

4.1 Symmetry dataset: COCO

For our symmetry experiments, we use a subset of the COCO 2017 validation set, which contains around 5,000 images. COCO offers a wide variety of real-world scenes with diverse objects and high-quality ground-truth annotations. This makes it a great choice for evaluating spatial reasoning in Vision-Language Models, especially since the images are "in the wild" and more representative of the kind of data these models are likely to encounter.

Due to computational limits, we focus only on the validation set. In our setup, we take a single COCO image, ask a question about it, and then flip the image horizontally and ask the same question again. This simple transformation helps us test whether the model's answers remain consistent or whether flipping introduces confusion, bias, or failure. Since many VLMs are trained on COCO (or datasets with similar distributions), this also helps us explore whether there is any data leakage or memorization that causes the model to prefer one orientation over the other.

Importantly, flipping an image should not change its core content. A bus is still a bus, a person is still a person. What does change are the spatial relationships. For example, if a dog is on the left side of a couch in the original image, then after flipping, the dog should now be on the right. We use this as a basic test: does the model remain consistent and updates the spatial orientations?

Figure 1: Example of image being flipped horizontally.

4.2 Symmetry Left/Right dataset

For our left–right reasoning task, we picked about **1,000 images** from the COCO 2017 validation set. These were the ones where we could clearly find two unique objects that were far enough apart horizontally.

We used the COCO annotations to make sure we could tell which object was on the left and which was on the right. This filtered set gave us cleaner examples to test how well models understand spatial relationships.

4.3 Symmetry Detection

This is a simple sanity-check task. First, we take a regular COCO image and ask the model:

> "Is this image symmetrical?"

Since COCO images are naturally captured in the wild, they are never perfectly symmetrical, so the answer is No.

Next, we synthetically generate a perfectly symmetrical version by taking the left half of the image and mirroring it to the right. We present this new image to the model and ask the same question. In this case, the expected answer is Yes, since the image is now intentionally symmetrical.

This task helps us verify if the model has a basic visual understanding of symmetry and whether it responds appropriately to clearly mirrored content.

4.4 Object List Consistency

In this task, we test whether models are consistent in the way they perceive and describe objects before and after a flip.

We prompt the model to:

> "List all the objects you can see in this image."

Then we flip the image horizontally and ask the same question again. Ideally, the model should list the same objects, as the flip shouldn't change what's being present in the scene.

To evaluate this, we compute the Jaccard similarity between the object lists before and after the flip. We also keep track of how many objects were added or deleted when the image was flipped as to calculate consistency.

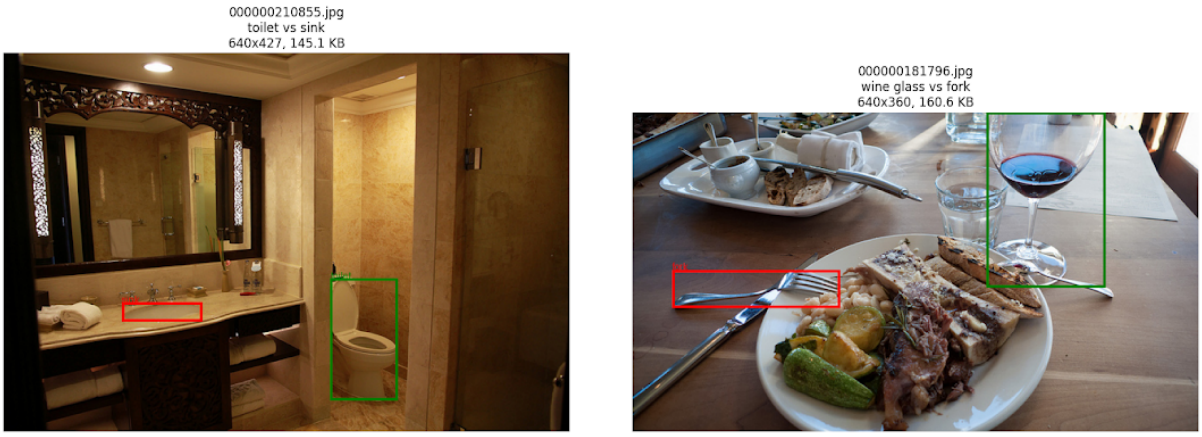

Figure 2: Is the sink to the left or to right of the toilet? Is the wine of glass to the left or to the right of the fork?

4.5 Left/Right Consistency under Image Flips

Using our left/right symmetry dataset, where we have two clear distinct objects that are clearly separated horizontally, we prompt the model with a spatial question, such as:

> "Is the dog to the left or right of the couch?"

Then, we flip the image and ask the same question again. A spatially-aware model should answer with the opposite direction the second time. If the original answer was "left," the flipped answer should be "right," and vice versa.

We evaluate the model's spatial reasoning under both viewer-centric and ego-centric prompts:

- **Viewer-centric:** "From your point of view, is the dog to the left or right of the couch?"
- **Ego/Agentic-centric:** "If you are the couch in this scene, is the dog to your left or right?"

Our goal in evaluating this Vision-Language Model is to figure out if there is any consistency issue, as well as proposing a new Ego/Agentic prompt for spatial relationship reasoning. We track the accuracy, flip-consistency, and any biases toward one direction.

5 Symmetry Tasks Results

We now evaluate our four Vision-Language Models on the three tasks we described earlier.

5.1 Symmetry Detection

In this task, we asked the models whether an image is symmetrical. When we give them perfectly symmetrical images (which we created ourselves), all models perform extremely well—most reach 100% accuracy.

However, when we use real-world images that are *not* symmetrical, the models often fail. In fact, most models default to saying "Yes, it is completely symmetrical" even when the image clearly isn't. For example, GPT-4o labeled every image as symmetrical, regardless of the actual content.

As shown in Figure 3, LLaVA-13B stands out as the best at catching non-symmetrical images, with around 48% accuracy. That's still far from perfect, but it's noticeably better than the rest. Claude does slightly better than the others too, it doesn't always default to "yes," and gives about 95% accuracy on symmetrical images rather than 100%, showing a bit more balance.

Overall, this highlights that detecting symmetry, especially the lack of it, is still a big challenge for these models.

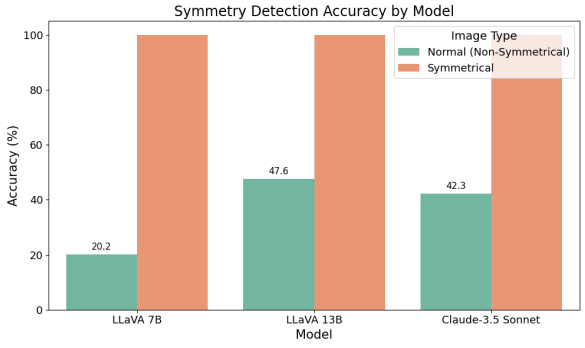

Figure 3: Symmetry Detection: Comparison between LLaVA-7B/13B and Claude-3.5 Sonnet for images that are symmetrical and non-symmetrical.

5.2 Object List Consistency

In this task, we looked at how consistent models are when listing objects in an image, before and after flipping it horizontally. The setup is simple: show the model an image, ask it to list all the objects it sees, then flip the image and ask the same question again. Ideally, we want the object lists to be as similar as possible.

What we found is that flipping the image often throws the models off, the lists change quite a bit. Even though the content of the image stays the same, the models often detect different objects.

Interestingly, LLaVA-13B listed fewer objects on average than LLaVA-7B, but the objects it found were more consistent and relevant. LLaVA-7B tends to repeat itself or list the same object multiple times, which inflates its numbers but doesn't help with quality. Because of this, LLaVA-13B ended up being the most robust and consistent overall in this task.

Meanwhile, LLaVA-7B and GPT-4o had pretty similar performance in terms of average objects and overlap, but still trailed behind 13B when it came to clean and stable object recognition after flipping.

5.3 Left/Right Consistency under Image Flips

In this task, we tested how well models handle left/right spatial reasoning when an image is flipped. We used our curated subset of 1,000 COCO images, where we know the exact ground truth of where one object is in relation to another.

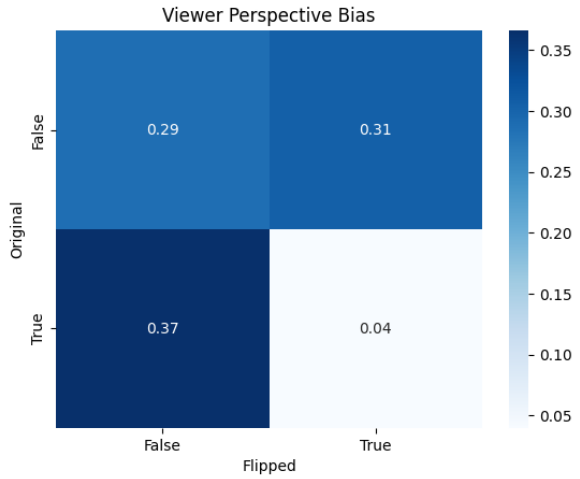

Figure 4: Bias found for LLaVA7B on viewer-centric perspective, big differnce between 0.31 and 0.37.

To evaluate this, we introduced a new metric: consistency. This checks whether the model gives logically consistent answers before and after the image is flipped. For example, if an object is to the left in the original image, it should be to the right after the flip. We consider it consistent when the model either gets both answers right, or both wrong. Inconsistent answers like saying "left" both times reveal a failure to account for the change in image orientation.

We also looked at differences between the two types of prompts: viewer-centric (looking at the image from the outside) and egocentric (pretending to be the object in the image). If the viewer-centric ground truth says "Object A is to the left of Object B," then from an egocentric point of view, it should become "right."

In terms of raw accuracy, LLaVA-7B (73.5% original image and 73.6% for mirrored image) surprisingly outperformed the larger LLaVA-13B model. We think this could be because smaller models are less prone to overfitting, and may generalize better in this type of task. However, when it comes to consistency under image flips, LLaVA-13B does better (74.6%), it seems to handle spatial transformations more robustly, even if it's slightly less accurate overall.

GPT-4o, on the other hand, struggled quite a bit here. It performed noticeably worse than the LLaVA models in both accuracy and consistency, showing that spatial reasoning under image transformations remains a real challenge even for top-tier models.

6 Size Perception Benchmarking

To evaluate spatial reasoning in Vision-Language Models (VLMs), we introduce a synthetic benchmark focused on size perception. Unlike natural image datasets, this benchmark provides full control over visual variables, enabling precise analysis of model behavior under different types of visual symmetry and contrast.

6.1 Dataset Construction

Each image consists of two geometric shapes, placed side-by-side on a white canvas: one on the left, one on the right. Shapes are either `circle` or `square`, with size determined by radius or side length, and colors drawn from a small fixed set (e.g., `red`, `green`, `blue`). The size difference between the two objects is always non-trivial (≥ 20 pixels) to avoid ambiguity.

All images are procedurally generated using sim-

Model	Avg. #Objects (Original)	Avg. #Objects (Flipped)	Avg. Shared	Avg. Jaccard
LLaVA-7B	**7.07**	**6.81**	3.64	0.470
LLaVA-13B	3.02	2.61	1.72	**0.660**
GPT-4o	6.79	6.89	**3.89**	0.460

Table 1: Object list consistency across flipped images. We report the average number of objects detected in the original and flipped images, average number of shared objects, and Jaccard similarity. Best scores are highlighted in bold.

Model	View	Original Accuracy (%)	Flipped Accuracy (%)	Consistency (%)
LLaVA-13B	Ego	72.2	72.6	**74.6**
LLaVA-13B	Viewer	34.3	35.2	34.5
LLaVA-7B	Ego	**73.5**	**73.6**	68.1
LLaVA-7B	Viewer	35.8	36.9	35.3
GPT-4o	Ego	59.6	59.4	45.8
GPT-4o	Viewer	30.7	28.3	49.8
Claude-3.5 Sonnet	Ego	72.2	71.9	63.7
Claude-3.5 Sonnet	Viewer	25.2	26.0	61.6

Table 2: Evaluation of each Vision-Language Model on the left–right spatial reasoning task, comparing performance across different model architectures and point-of-view prompts (egocentric vs. viewer-centric).

ple rendering logic and include exact annotations for the shape, size, and color of each object. A total of **720** such images are created, spanning three key experimental conditions:

- **Same Shape, Same Color:** Objects are identical in appearance except for size (e.g., two green circles).
- **Same Shape, Different Color:** Shape is the same, but color provides an additional cue (e.g., a small blue circle and a large green circle).
- **Different Shape and Color:** Shape and color vary independently (e.g., a small green circle vs. a large red square).

6.2 Sample Images

Figure 5 illustrates one example from each of the three dataset categories described above. These examples help visualize the range of variation in our benchmark while maintaining the strict control needed for diagnostic evaluation.

6.3 Experimental Protocol

To evaluate each model's understanding of relative size, we prompt it with natural-language questions about the synthetic images. For each image, we ask up to five questions, depending on the experimental condition:

- `What color is the object on the left?`
- `What color is the object on the right?`

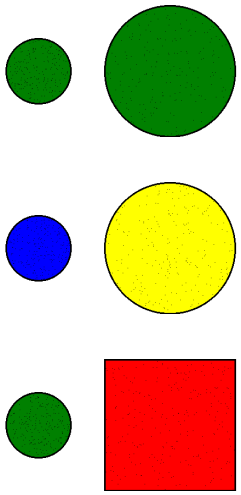

Figure 5: Examples from the synthetic size benchmark. Top: same shape and color. Middle: same shape, different color. Bottom: different shape and color.

- `Which object is larger: the one on the left or the one on the right?`
- `The larger object has which color?`
- `Which object is larger: the square or the circle?`

Each response is normalized using a simple string-matching function that selects the first valid label (e.g., "`left`", "`green`", "`circle`") mentioned in the model's answer. These are then compared to the image's ground truth annotations. This automatic scoring process ensures high throughput and consistency across models and image-question pairs.

Condition	Model	Size Comparison (Left vs Right)	Color Identification Accuracy	Shape Identification Accuracy (Left / Right)	Size Comparison (By Shape)
Same Shape, Same Color	LLaVA-7B	**95/160**	–	–	–
	LLaVA-13B	81/160	–	–	–
	GPT-4o	**160/160**	–	–	–
Same Shape, Different Color	LLaVA-7B	376/480	475/480	–	382/480
	LLaVA-13B	376/480	475/480	–	382/480
	GPT-4o	**480/480**	**480/480**	–	**480/480**
Different Shape and Color	LLaVA-7B	64/80	–	75/80, 78/80	78/80
	LLaVA-13B	30/80	–	**80/80**, 73/80	61/80
	GPT-4o	**80/80**	–	**80/80**, **80/80**	**69/80**

Table 3: Performance of vision-language models on controlled size perception tasks. "Size Comparison (Left vs Right)" tests spatial reasoning; "Color Identification" and "Shape Identification" test recognition; "Size Comparison (By Shape)" asks which object (e.g., square vs. circle) is larger. GPT-4o consistently achieves near-perfect or perfect accuracy.

We evaluate three classes of questions:

1. **Recognition**: identifying the color or shape of each object.

2. **Relative Size (by position)**: determining whether the left or right object is larger.

3. **Relative Size (by label)**: determining which shape or color corresponds to the larger object.

For all experiments, we evaluate three models: LLaVA-1.5-7B, LLaVA-1.5-13B, and GPT-4o. All models are run in a sequential question-answering mode, where each question is posed independently and does not rely on prior context.

7 Size Perception Results

Table 3 reports the performance of each model across the three experimental conditions. We break down results by question type: object recognition (color or shape), spatial comparison (left vs. right), and comparative reasoning by label (e.g., "Which shape is larger?").

GPT-4o achieves near-perfect accuracy across all question types and benchmark conditions, including the most difficult setting where both objects are visually identical in shape and color, and only size differentiates them. This suggests that GPT-4o possesses a robust, grounded understanding of visual magnitude. While the model is closed-source and its internal architecture is not fully disclosed, we hypothesize that its strong performance may be due in part to a dedicated object detection or segmentation module in its image-processing pipeline, which enables it to accurately parse object boundaries and compare relative areas.

In contrast, LLaVA-1.5-7B consistently outperforms LLaVA-1.5-13B in spatial comparison tasks, especially under low-cue conditions. This counterintuitive trend, where a smaller model exhibits better spatial reasoning—suggests that increasing model size alone does not guarantee improved perceptual grounding. Neither LLaVA variant performs reliably when color and shape cues are removed, indicating a fundamental weakness in size perception. These findings echo concerns raised by the "Eyes Wide Shut" benchmark (Tong et al., 2024), which highlighted that CLIP-based models often fail to encode fine-grained spatial relationships and visual detail.

8 Future Work

Our benchmark opens several avenues for future exploration. First, the current setup focuses on 2D spatial reasoning. Extending this work to 3D environments, such as simulated robotics or embodied agents in virtual worlds, would allow us to test whether VLMs can reason about occlusion, depth, and perspective, capabilities that are critical for real-world applications.

Second, while the symmetry benchmark was used strictly for evaluation, it can also be used for targeted training. Fine-tuning or instruction-tuning models like LLaVA directly on these size comparison tasks may help diagnose whether their architectural limitations are due to lack of data or inherent biases in their training regime.

Finally, we propose a probing study: extract CLIP image embeddings for each image in the benchmark and train a simple regression model to predict the size of the larger object or the relative size difference. If regression achieves high performance, this would indicate that the information is present in the embeddings but not being utilized effectively downstream, mirroring similar findings in probing language models.

Each of these directions offers potential to push VLMs beyond high-level captioning and classification, toward more grounded, perceptual intelligence.

9 Conclusion

Overall, our results suggest that while foundation VLMs have made progress in general visual understanding, basic perceptual faculties like size comparison remain underdeveloped in open-source models trained on large-scale contrastive or instruction-tuning objectives.

This limitation becomes especially apparent when models are tasked with reasoning about fine-grained spatial relationships, such as identifying left-right orientation under image transformations or detecting symmetry. Despite performing well on benchmarks like VQA or captioning, these models often default to biased or inconsistent outputs when the spatial layout of a scene is altered. Our experiments with flipped images, controlled synthetic scenes, and object list consistency tests reveal that current training pipelines may not sufficiently expose models to spatial variations or teach them to reason geometrically.

This suggests that incorporating spatial priors or explicit spatial reasoning objectives into the training process may be necessary for improving robustness in real-world applications that rely on physical understanding.

10 Acknowledgments

The team would like to thank Dr. Raymond Mooney for his guidance through out this class project and for his valuable feedback.

Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. *Preprint*, arXiv:2204.03162.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *Preprint*, arXiv:2401.06209.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. 2025. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. *Preprint*, arXiv:2410.17385.