

Choosing ‘Right’ from Wrong: A Closer Look at Selection Bias in Spatial Multiple-Choice Questions in Large Multimodal Models

Giselle Zeno
MIT Lincoln Laboratory
giselle.zeno@ll.mit.edu

Nour Jedidi
MIT Lincoln Laboratory
nour.jedidi@ll.mit.edu

Steven R. Gomez
MIT Lincoln Laboratory
steven.gomez@ll.mit.edu

Abstract

This paper examines the limitations of a probing method for determining whether Large Multimodal Models (LMMs) actually understand spatial concepts in image inputs, which may be important for users calibrating their trust during visual-question answering. A typical approach is to prompt these models to choose the correct matching text caption from multiple-choice options. However, similar prompting with Large Language Models (LLMs) has been shown to give biased results as an artifact of the multiple-choice question (MCQ) format.

We explore the extent to which this bias persists for spatial understanding tasks in LMMs, which take visual inputs alongside the MCQ text prompt, and go beyond recent work in characterizing this selection bias for current models. First, we demonstrate how image-text alignment benchmarks can be used to formulate a question for the LMM to reveal potential bias. Then, we use eight LMM architectures to perform this task using image-caption sets from the What’s Up benchmark. We find that this multiple-choice selection bias exists in LMMs even with state of the art models and under different methods of analysis. Finally, we discuss future steps, as well as alternatives to MCQs for probing spatial understanding in LMMs that could minimize bias introduced by the prompt format.

1. Introduction

Recently, Large Multimodal Models (LMMs), such as GPT-4o and LLaVa family [14] models, have gained popularity as visual question-answering tools, since they are generative architectures capable of processing both visual and text inputs. LMMs are transforming the way AI assistants help people understand images and text, and it stands that users ought to be able to test basic spatial knowledge, such as whether a model can discern concepts like ‘left’ from ‘right’. We may ask, *What is the best way to do that probing?* As this paper explores, recent research has shown

that some of the most intuitive ways of testing a generative model’s skill are prone to bias, and in this work we reproduce and look more deeply at some of those findings.

There is a growing body of datasets and benchmarks that are helping evaluators characterize the abilities of large pre-trained models. In the text-only setting, multiple-choice questions (MCQs) have been used as a straight-forward probing method to assess the capabilities of Large Language Models (LLMs), and a number of MCQ text benchmarks have been proposed. However, modern LLMs have been shown by to suffer from *selection bias* where they prefer specific option order indices or option labels as answers (e.g., option ‘A’), and thus are not robust against option permutations in MCQs [30]. In fact, the authors Zheng et al. suggest that selection bias is an inherent behavioral bias of LLMs and it cannot be addressed by simple prompt engineering [30].

There may be implications of this for testing vision models, like LMMs. Many Vision-Language Compositionality (VLQ) benchmarks have been formulated as image-to-text retrieval problems, where vision-language models are given as input an image and some text descriptions (candidate captions or answers), and the model is tasked with selecting the matching or correct response for the given image. These tasks are functionally multimodal MCQs. In contrastively-trained cross-modal vision-language models (e.g., CLIP [18]), which embed images and text into a common vector space based on semantics, the answer “selection” can be determined by finding the text candidate from the answer options whose embedding is closest numerically to the image embedding. LMMs, on the other hand, process MCQs end-to-end as combined text and image prompts, and generate a text answer—as a human would—rather than only compute embeddings. Here, LMMs use language models to generate text after processing both the text inputs and image encodings that are produced earlier in the LMM architecture. But since LMMs depend on language models, which are known to exhibit selection bias, we should be cautious about relying on MCQs for evaluation without further characterization of how this bias manifests in LMM

MCQ tasks. These models should rely on the visual input to ground their answer to the captioning prompts, but earlier findings suggest that vision-language models are also vulnerable to bias and option permutations [32].

Our goal in this paper is to build on these earlier studies by looking more closely at how LMMs confuse options even on visually-unambiguous MCQs related to basic spatial relationships (i.e., *does low visual-reasoning difficulty sufficiently mitigate selection bias?*), and what directions remain for controlled ways of evaluating spatial understanding in LMMs. The contributions of our work include:

- a quantitative evaluation that finds evidence that LMMs are susceptible to positional bias in spatial MCQ tasks, with varying biases observed across different LMMs (even those with larger model sizes), and
- a discussion of new directions for investigating bias in the next generation of large pre-trained models, as well as next steps for understanding LMM spatial-reasoning skills without relying on the MCQ format.

The remainder of this paper is organized as follows: In Section 2, we describe related work on using MCQs as an assessment for LLMs (and biases discovered), along with other benchmarks and probing methods for reasoning in these models. In Section 3, we present a formative study to spot-check whether modern LMM architectures are still susceptible to the same biases found earlier, which led to newly identified research questions. We detail a follow-on comparative study across LMMs in Section 4, summarize our findings in Section 5, and outline new research directions in Section 6.

2. Related Work

Multiple-choice question answering with Large Language Models. How should one go about prompting an LLM to answer a multiple-choice question in the first place? In fact, studies have looked at alternative prompting schemes to understand their benefits and costs. Robinson et al. contrasted *cloze prompting* versus *multiple-choice prompting* (MCP), which differ in how the likelihoods of generated outputs are used to ‘answer’ the question [20]. In cloze prompting, each candidate answer’s likelihood by the LLM is compared, and the most likely one is treated as the answer. But as the authors note, this process blurs the likelihood of the produced sentence with the accuracy of the response to the actual question, which can lead to problems (e.g., the most accurate response may be unusual in the language). They also found that different normalization approaches to the probabilities compared in cloze prompting led to different answers. Instead, the authors argue for MCP, where the prompt binds option ID symbols to a set of option labels (*A*, *B*, *C*, etc.) and they treat the LLM’s answer as the option associated with the highest-likelihood option label. Research building on MCP has discovered new challenges,

however.

Imposing a list of options in a MCQ prompt has been shown to reveal biases related to this construction. Recent work by Zheng et al. [30] found that LLMs are not reliable MCQ selectors, demonstrating that some models are biased to particular selections independent of the correctness of those captions. For the open-source LLMs the researchers evaluated, they prompted the model with the multiple-choice question, accessed the probabilities of option ID tokens, and chose the most-likely option as the ‘answer’. In the case of GPT, where these probabilities were not available, they compared the first token generated with the gold answer. The authors found that both *position bias*—where LLMs may prefer answers because of their positions in a sequence of options—and *token bias*—where LLMs may prefer, for example, ‘A’ over ‘B’—both occur and in some cases may interact. Another contribution of this work was a debiasing method called PriDe that estimates the selection bias on each option through a sampling process, and then removes the weight of that bias when decoding the likelihoods of the options for the model’s answer in the remaining tests [30].

Initial evidence that the presence of visual context alone is not a solution to selection bias. Zong et al. followed up on this work and demonstrated that the selection bias is also present when testing different vision-language models on visual benchmarks [32]. Furthermore, they highlighted the challenge of debiasing model responses, showing that PriDe is only effective when applied to MCQs in their original order; it cannot address using permutations of options for a single question to score a model’s accuracy (which, of course, should not confuse a robust question-answering system). LLMs were previously shown to be vulnerable to permuting the order of the options [17], and even the specific choice of option ID symbols (e.g., ‘1’ versus ‘i’) can influence the accuracy or interact with specific permutations of the options. Zong et al. conclude that “*not only are the [LLMs evaluated] not invariant to permutation, but the specific way in which they are not invariant is dependent on the choice of symbols*” [32].

Finally, several ideas have emerged to help evaluators interpret how large models respond to MCQs (with MCP in particular). Wang et al. found that LLMs might actually be generating the “least incorrect” option rather than the definitive correct one [23]. A suggestion in their work is to augment datasets by changing MCQs into a set of True/False questions about the individual options, among other adaptations. Another idea is to have models generate an explanation before answering, which may prime the context so the model can give a more accurate option selection [17].

Probing vision-language models for compositional reasoning. Recently, a variety of benchmarks have been developed with the goal of probing vision-language mod-



Figure 1. Two examples of image-caption pairs from Winoground [22]. The emphasized words in each caption can be swapped to describe an alternate paired image in the data set, as shown. These include swapping both object concepts (Figure 1a) and prepositions (Figure 1b).

els for their compositional reasoning capabilities. Notable datasets and work in this space include Winoground [22], the ARO benchmark [28], SUGARCREPE [9], and BiVLC [16]. Datasets that focus on spatial reasoning in scenes include What’s Up [13], COCO-spatial [22], and GQA-spatial [22].

A goal in many of these benchmarks is to create an image-caption alignment task that can only be solved if the model is capable of distinguishing the correct caption from alternative incorrect ones. Two examples from Winoground are shown in Figure 1, where alternative captions have an identical *bag of words* but their arrangement differs and they align to distinct images. For our purposes of testing spatial reasoning in LMMs with benchmarks like these, we consider selecting the image and then using the alternative captions (one being accurate) as multiple-choice options.

Focusing on spatial relationships. The ability of LMMs to generate arbitrary outputs opens up other ways to probe the spatial understanding of these models. For example, Rajabi and Kořecká developed an extension to the What’s Up benchmarks called GSR-Bench [19], which added finer-grain annotations like bounding boxes around the image objects and additional depth relations. One use case for this benchmark is to prompt models to output 2D coordinates for grounded regions of interest. This opens the door to new metrics, especially how best to give partial credit for the outputs that indicate some spatial understanding (e.g., quantifying overlaps between AI and human-curated responses on spatial tasks [4]).

Other tasks, like map interpretation or maze navigation, may test fundamental spatial reasoning abilities in different kinds of large pretrained models. Wang et al. proposed SpatialEval, a novel evaluation suite for spatial-related understanding with multimodal models and LLMs (given text-based image descriptions) that spans several task types [24]. An interesting finding in their experiments was a number of LMMs performed tasks better when given only a text description of the visual content, rather than the original image plus the description, which suggests that image encodings may be harder to reason about for LMMs than a well-curated text description, and that having both distracts

from—rather than improves—the text reasoning. Our experiments differ in that we focus on spatial reasoning in images that do not have a faithful, accompanying text description available to the model at reasoning time. In fact, one of our goal’s of evaluating spatial understanding capabilities in LMMs might be to determine their fitness for automated image captioning in vision systems.

3. Preliminary Study: Are Current LMMs Biased on Compositional MCQs?

Our starting point for investigating whether current LMMs still exhibit selection bias was to perform a formative experiment using the popular Winoground benchmark [22]. The authors of Winoground evaluated the benchmark against embedding-based vision-language models like CLIP, but a key question is how LMMs do at this compositional reasoning task, given questions about the ease of discerning the word swaps based on sentence likelihood (e.g., [9]) and known positional bias in LLMs. Note that for the remainder of this paper, we label LMM architectures that contain LLM components using a comma to separate the visual and language components (e.g., “LLaVa-1.6, Mistral-7B”).

3.1. Experiment Setup

Data In Winoground, each example has two images (I_0, I_1) with corresponding captions (C_0, C_1). The two captions contain the same words but in a different order. We use the subset of *relation* swaps, that reorders elements (verbs, adjectives, prepositions, and/or adverbs) that take nouns (objects) as semantic arguments. In this set, some relations are spatial in nature (prepositions) while others include different kinds of modifiers. In our study, permutations of the order of options are tested for each query image.

Model In this preliminary experiment, we used LLaVa-1.6, Mistral-7B with a greedy generation parameter.

Prompt Since there are two alternative captions for each image in Winoground, two prompts per image are evaluated where the option order is swapped. Each prompt has the form: “Which caption is correct: (a) <caption 1> or (b) <caption 2>?”

	Selected option		
		(a)	(b)
	Correct option	(a) 0.85	(b) 0.15
		(b) 0.58	0.42

Table 1. Results from spot-checking LLaVa-1.6-Mistral-7B on MCQ image-caption alignment tasks using Winoground [22]

3.2. Findings

Table 1 shows the selection results for the model on our MCQ tasks. For LLaVA-1.6, Mistral-7B, the model selects the correct option overall around 64% of the time. But when option (b) is the correct answer, the model selects this option only 42% of the time, worse than random chance (50%). Overall, the model chooses option (a) in 71% of the examples, exhibiting a bias towards the first option (χ^2 test, p -value $\ll 10^{-4}$). Therefore, we find that positional bias is still a problem for this LMM. This raises additional questions for further investigation:

- *With more distractor options and LMM architectures, can we find support for the claim that positional bias can appear beyond the first option in the MCQ?* One finding by Zheng et al. in their LLM MCQ testing was that positional bias was prevalent across models, but not necessarily in the same option index [30].
- *Is selection bias present when a task’s answer is more obvious to humans and less ambiguous in the visual context?* We expect the subset of Winoground examples focusing on prepositional swaps (see Figure 1b) is a lower bar of difficulty to solve as MCQ tasks, since these examples generally contain overt spatial differences between the image pairs, and may be less dependent on one’s interpretation compared to other relation swaps (e.g., Figure 1a). A test set focusing only on these spatial relations might cause LMMs to exhibit less selection bias.
- *Do we see any effect on selection bias when the option ID symbol is changed from (a, b) to another enumeration?* This would provide additional evidence for or against the spurious symbol correlation found by a recent study [32].

In the next section, we describe our follow-on experiments aimed at answering these questions.

4. Study: Comparing LMMs on a Spatial MCQ Test

The preliminary experiment led us to hone in on selection bias that might exist in spatial-understanding MCQs tasks for LMMs. By focusing on basic spatial relationships in the MCQs, we want to test selection bias not just with a visual context but where the task is as unambiguous as possible. In a sense, we are interested in knowing if making the MCQ task as simple and evident as possible (to a human) reduces the selection bias earlier studies have found. We make the

following hypotheses:

- H1:** LMMs will not exhibit significant selection bias (position or ID based) identifying basic spatial relationships in an image through MCQs. Having access to additional vision context will mitigate the effect of MCQ selection bias on basic spatial tasks.
- H2a:** If selection bias does exist, LMMs with larger parameter sizes exhibit less-biased spatial understanding answers compared to smaller models.
- H2b:** If selection bias does exist, it is more attributable to positional bias than the specific option ID symbols used, as seen in earlier studies with LLMs.

4.1. Models

Table 2 lists the LMM architectures, language models (LLMs), and model sizes (parameters) used in our experiments. We compare various LMMs to contrast differences, if any, across architectures. To investigate the effect of LLMs on selection bias in LMMs, we evaluate the same architecture with different language models. We also vary language model size to determine if the number of parameters has an effect on the option-selection bias observed.

Architecture	Language Model	Model Size [†]
LLaVA-1.6 [15]	Mistral [12]	7B
	Vicuna-1.5 [31]	7B, 13B
	Llama-3 [1]	8B
Llama-3.2-Vision [1]	Llama-3.1 [1]	11B
InstructBLIP [6]	Flan-T5-XXL [5]	11B
Qwen2-VL [25]	Qwen2 [27]	7B
Qwen2.5-VL [3]	Qwen2.5 [3]	3B, 7B

Table 2. Large Multimodal Model (LMM) architectures tested in our study. [†]In this table, model size is given as an approximate value that corresponds to the number of parameters in each of the model variations we tested, generally dominated by the language model component.

4.2. Dataset

We chose to use the What’s Up image-caption dataset for testing spatial relations [13]. Each example is a scene containing two objects and includes four images (I_0, I_1, I_2, I_3) with corresponding captions (C_0, C_1, C_2, C_3). Each set of images changes the underlying preposition that describes the spatial relationship between two objects (e.g., one set of images contains a mug “on”, “under”, “left of”, and “right of” a table). Two other visual MCQ datasets were considered for this experiment, but were not used. GSR-Bench [19] extends What’s Up, but its localization annotations/task (e.g., prompting for bounding regions of objects of interest) does not lend itself to an obvious MCQ formulation. As we saw in our preliminary study, Winoground [22]

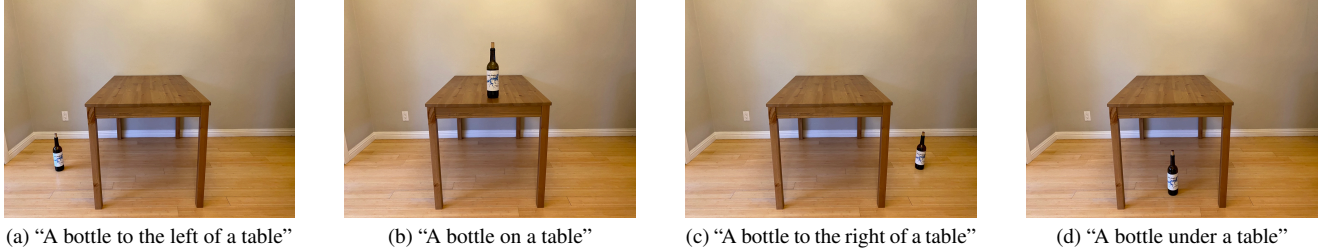


Figure 2. Image-caption pairs from the What’s Up benchmark [13], showing distinct spatial configurations between two common objects.

has image-caption examples that contain compositional swaps, but only a subset of them focus on aligning spatial/prepositional differences, which we aim to explore here. Furthermore, we are interested in seeing how models select option labels over a range of order-permutations, but Winoground only has two choices per image compared to four in What’s Up.

4.3. Methods

Our experimental setup is as follows: for a given image I_k , the model is prompted to select the single correct caption C_k from four options. Each prompt has the form: *Given the image, which caption below is correct?*, then the four options are listed with the option label (e.g., ‘A’) in parentheses coming before the option text; then the model is asked to complete “The correct answer is (”.

Each test scene has four image variations in the What’s Up dataset (one object is moved into four distinct positions relative to another), and we formulate a set of test prompts to pair with each image that includes all order permutations of the caption options (i.e., 24 permutations for each of the four image variations of a scene, 96 in total). In addition, each of these prompts is applied using three sets of option ID symbols that reflect typical enumeration schemes for MCQs: A/B/C/D, a/b/c/d, and 1/2/3/4.

We determine a model’s selection for a MCQ prompt by parsing the generated response text and matching it against the answer options. Another way of determining the model’s selection is to find the most likely option-token logit at the start of the LMM’s response generation, but—since these likelihoods are typically hidden in LMM outputs—this approach may not be available to an evaluator in closed models/APIs. In initial exploration we obtained comparable results, so we leave this as future work.

Analysis Approach We first investigate if the selection bias observed in LLMs is present in LMMs. Related work tested for selection bias by comparing the number of times that the LLM selected each MCQ option label A/B/C/D against an expected uniform distribution [30]. To test this in our results, we will let X and Y represent the distribution of correct options (uniform for a balanced dataset) and the distribution of options selected by the model, respectively. A hypothesis test only on the distribution of selected

options Y would ignore any associations between both variables. We instead consider the joint distribution of X and Y , which allows us to evaluate if the model is equally biased to a particular option across all alternate options. Therefore, we perform additional hypothesis tests to determine if any selection bias found is input dependent.

Other work has shown that LLMs tend to select the “least incorrect” option when they choose incorrectly [23]. In the dataset we use, all other captions besides the correct one are objectively false for the paired image (no meaningful “least incorrect”), so we do not explore that finding in our study. Instead, we investigate if and how LMMs exhibit patterns when they choose incorrectly. See the Supplementary Material for more details on these hypothesis tests and results.

5. Results

Summary of Findings

- **Evidence of selection bias on this task (H1 not supported):** On all models, the proportion of times each option was selected, for each of the option ID symbols, was statistically nonuniform (Table 3, χ^2 Test with $p \ll 10^{-4}$). These results show that these LMMs also suffer from selection bias.
- We also found that the selection bias towards different options varies depending on the input: the variables X and Y , for the correct option and selected option, are not independent (Table S1 in Supplementary Material, independence test with $p \ll 10^{-4}$).
- **Some models exhibited more bias than others, but not necessarily due to model size (H2a not well supported).** For example, Qwen-2.5-VL is notably less biased on the A/B/C/D tests (Figure 3) compared to InstructBLIP and Qwen-2-VL architectures, but the difference in confusion matrices between Qwen-2.5-VL 3B and the larger Qwen-2.5-VL 7B is minimal. In the LLaVA-Vicuna series, we also observed that the lower-capacity language model actually outperformed the higher-capacity 13B model in some cases.
- **The positional bias appeared mostly consistent across the sets of option ID symbols (some support for H2b).** However, we did observe some variability in the confusion matrices corresponding to the different symbol sets we tested (a/b/c/d, A/B/C/D, and 1/2/3/4).

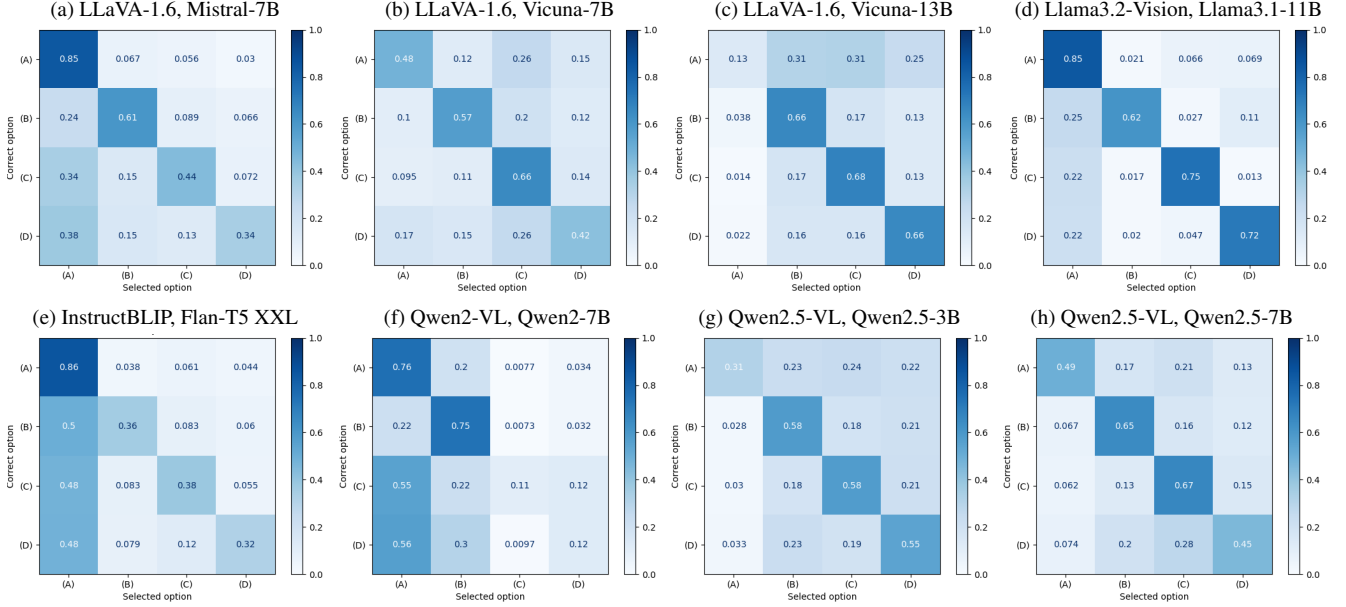


Figure 3. Confusion Matrices (row-normalized) for MCQ (A/B/C/D), What’s Up dataset.

Architecture LLM		$\bar{p}_0$	$\bar{p}_1$	$\bar{p}_2$	$\bar{p}_3$
LLaVA-1.6	Mistral-7B	0.43	0.25	0.18	0.15
	Vicuna-7B	0.24	0.31	0.30	0.14
	Vicuna-13B	0.08	0.37	0.31	0.24
	Llama3-8B	0.10	0.37	0.31	0.22
Llama3.2-Vision	Llama3.1	0.41	0.18	0.19	0.22
InstructBLIP	Flan-T5 XXL	0.47	0.24	0.14	0.16
Qwen2-VL	Qwen2-7B	0.35	0.29	0.19	0.16
Qwen2.5-VL	Qwen2.5-3B	0.16	0.31	0.29	0.24
	Qwen2.5-7B	0.17	0.27	0.33	0.23

Table 3. Average proportion of times each option was selected in the MCQ experiments with pre-filled prompt on the What’s Up dataset. Zero-based indices for p correspond to the various option ID symbols a/b/c/d, A/B/C/D, and 1/2/3/4. On all models, the proportion of times p_j that each option label was selected was statistically non-uniform (χ^2 Test, $p \ll 10^{-4}$), for all option ID symbols. Full results in Table S2 of the Supplementary Material.

Interpreting Confusion Matrices Figure 3 and Figure 4 show the MCQ confusion matrices on the What’s Up dataset for task sets with option labels A/B/C/D and 1/2/3/4 respectively. Additional results on option ID symbols a/b/c/d, along with a smaller LLaVa-Llama configuration (LLaVA-1.6, Llama3-8B) across the option ID symbols, are available in the Supplementary Material. Results for option ID symbols a/b/c/d and A/B/C/D are consistent. All confusion matrices are normalized by rows. The main diagonal values

in the confusion matrices correspond to accuracy for each option label; since all ordering permutations are tested, the overall accuracy of each model is the average of its diagonal values. In matrices with higher values below the main diagonal than above it, the model is more likely to select options that appear before the correct choice. If the non-diagonal values are homogeneous and symmetric, then the model selects the same incorrect answer regardless of its position and corresponding option label in the list of choices.

LLaVA-1.6, Mistral-7B In Figures 3a and 4a, we can observe that the model is most biased towards the first option, with average probability $\bar{p}_j = 0.43$ (Table 3). However, the conditional probabilities of Y (selected option) given X (correct option) in the first column are not uniform (i.e., the bias towards the first option is not the same).

LLaVA-1.6, Vicuna series We observe that Vicuna-7B is biased towards the third and second options in Figure 3b (C and B) and the second and first options in Figure 4b (options 2 and 1). Vicuna-13B is most biased towards the second and third options in Figures 3c and 4c (B and C, 2 and 3). Interestingly, it is also most likely to get the answer wrong when the correct choice is the first option.

Llama3.2-Vision, Llama3 In all of our tests, this is the model that appears to place most of its weight on the diagonal (i.e. higher accuracy). Despite having that high accuracy, it shows bias toward the first option under each of the option ID symbols variations (Figures 3d and 4d).

InstructBLIP, Flan-T5 XXL In Figures 3e and 4e, we see this model is most biased towards selecting the first option (options A and 1), but also to the second option in Figure 4e (option 2).

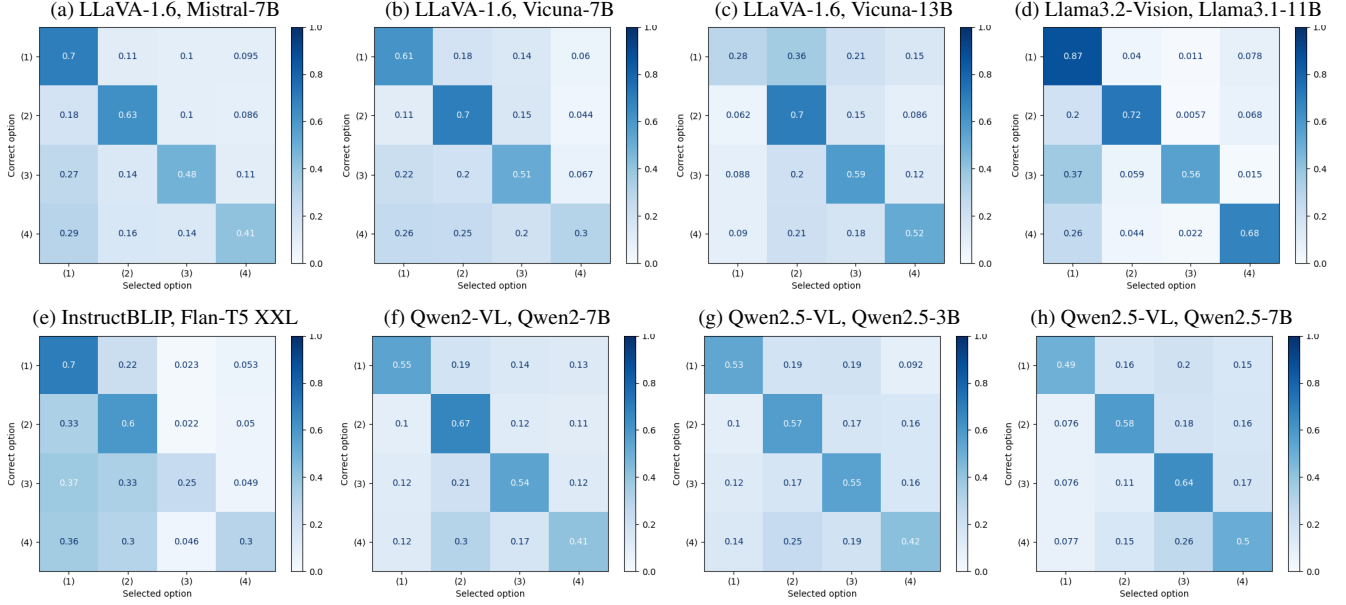


Figure 4. Confusion Matrices (row-normalized) for MCQ (1/2/3/4), What's Up dataset.

Qwen Series In Figure 3f, we observe that Qwen2-VL 7B is biased towards options A and B. Meanwhile, in Figure 4f looks to be biased towards option 2. For Qwen2.5-VL 3B, the model looks to have a stronger bias towards the last three options (B, C and D) in Figure 3g. In Figure 4g, the model is biased towards options 2 and 3. Overall, the model is most biased towards the second and third options. For Qwen2.5-VL 7B, the model is overall most biased towards the third options (C and 3) followed by the second options (B and 2) in Figures 3h and 4h. Compared to Qwen2-VL 7B, the bias to the first option is also reduced.

Model Size In the cases where we evaluated different model-size variations of the same architecture (LLaVA-1.6, Vicuna series and Qwen2.5-VL series), the confusion matrices of the larger models are very close to their smaller counterparts, with some interesting differences. Comparing LLaVa-1.6, Vicuna-7B with LLaVa-1.6, Vicuna-13B (in both Figures 3 and 4), we see that the lower-capacity language model actually outperforms the higher-capacity 13B model when the correct option is in the first position. For Qwen2.5-VL, the larger language model component (7B) shows slightly more uniformity off the diagonal compared to the 3B model. With A/B/C/D options (Figure 3), the smaller model selects option A less than other options in general.

6. Discussion and Future Work

In general, we did not find support for the hypothesis that LMMs—with a visual context accompanying the basic spatial multiple-choice question (MCQ) prompt—are less prone to selection bias compared to previous findings (H1).

Zong et al. [32] noted that the presence of visual context in the MCQ tasks “may intuitively add a layer of resilience” to bias vulnerabilities, but it did not mitigate the option-permutation attacks explored in their work; similarly, we find that the visual context is not a sufficient selection-bias mitigation even in simple spatial MCQs.

Furthermore, we found that LMMs with larger parameter sizes also struggle to give unbiased responses to MCQs that are aimed at revealing spatial reasoning capabilities (H2a). In other words, as newer, higher-capacity LMMs are released, we should not necessarily expect this type of prompting to become more reliable. We found some support for H2b, as positional biases were generally consistent in models across the sets of option ID symbols, although there were some examples of interesting differences attributable to the symbol sets (e.g., contrasting Figure 3f and Figure 4f).

In the remainder of this section, we discuss open directions in this area of research.

More controlled comparative evaluations. A limitation in our current study is that we did not exhaustively compare LMM architectures utilizing the same LLM components and model sizes. The goal in doing so would be to better pinpoint why some LMMs exhibit more bias than others on the What's Up tasks, and potentially to better predict when how new architectures perform with respect to bias.

Evaluation with Large Reasoning Models. Recently there has been a surge of interest in Large Reasoning Models (LRMs), such as OpenAI's o1 [11] and o3, Deepseek's

R1 [8], and Gemini 2.0 Flash Thinking [7]. Through generation of reasoning tokens (i.e., an extended chain-of-thought) prior to giving its final answer, these models have shown strong performance across a wide range of reasoning benchmarks. In future work, we aim to extend our analysis by investigating how LMMs that leverage LRMs as their backbone language model affect the kinds of biases present in answering multiple-choice questions. An advantage to this approach is that we can use the reasoning chain to better understand how the model makes its decisions and identify how its “thinking” may differ based on the positioning of the correct answer. We believe this can potentially enable a deeper understanding of what causes positional biases.

Attempt to debias, or drop MCQs for evaluating vision-language alignment? Given that current-generation LMMs are still prone to selection bias on spatial MCQs and that debiasing methods like PriDe may not be robust to different ways of writing the same test [32], we argue for exploring other means of testing an LMM’s spatial reasoning without the MCQ format. One alternative approach is to prompt the LMM for more complete, freeform scene descriptions of images, and then parse these outputs into scene representations that are comparable with ground-truth annotations about the spatial relationships visible. Recent work has looked at recovering and comparing structures between modalities like text and images. For example, Multimodal Named Entity Recognition (MNER) and Multimodal Relation Extraction (MRE) are information retrieval tasks that aim to recognize entities and extract relations among them using information from multiple modalities, such as text and images [10, 26]. Dependency-based RE methods (e.g., parsing sentences using syntax dependency trees) have shown to be successful, and can be applied in the task of MRE, as shown by Zheng et al. [29]. They propose the MNRE dataset, a knowledge graph containing 30k named entities and 23 relations types [29]. However, their work does not cover spatial relations between entities, which is an opportunity for extending this work in the future. For their multimodal representation, they build the visual graphs with a pretrained scene graph generator [21] and fine-tune a BERT encoder with separator tokens for entities. These kinds of scene graphs could be helpful outputs in a spatial-understanding evaluation pipeline for LMMs.

Developing metrics for probing methods like scene graph elicitation and extraction is less straightforward than looking at accuracy or selection bias in MCQs. One related metric is SPICE, for Semantic Propositional Image Caption Evaluation [2]. Although the authors in that work create scene graphs from the image captions, their metric uses sets of entity-relation and entity-attribute tuples (similar to n-grams) and calculates F1-score on the two sets of tuples (the ground truth and predicted caption). We think using seman-

tic embeddings to compare ground truth annotations against extracted scene structures—rather than tuples bound to entity/relation/attribute words—could be a more flexible path forward, but more exploration of this idea is needed.

7. Conclusion

Our study contributes evidence that Large Multimodal Models (LMMs), even with visual contexts where spatial relationships are visible and unambiguous, are biased in answering multiple-choice questions (MCQs) about these scenes. Our findings reinforce earlier results that show selection bias in LLMs. We argue that this bias (i.e., preferring certain option positions) is difficult to overcome and structural in the LLM, and as a result, LMM architectures which contain language models are also prone to this bias. Probing models with MCQs that are expected to be “easy” for the model *a priori*—or relying on the improved capabilities of new LMM architectures—are not realistic solutions to avoid this selection bias.

Our analyses demonstrated different statistics over confusion matrices to take a closer look at this MCQ bias. An exciting open opportunity is to study how next-generation Large Reasoning Models, whose inputs and outputs will be similar to LMMs but also produce reasoning tokens, perform on these kinds of tests and whether enumerating that reasoning fixes the bias or helps evaluators better understand why it occurs. We also see areas to improve probing spatial understanding in LMMs that avoid the MCQ format altogether, instead focusing on prompting and parsing more free-form spatial descriptive outputs into structures that are comparable with scene annotations for evaluation purposes.

Acknowledgements

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering. © 2025 Massachusetts Institute of Technology. Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

References

- [1] AI@Meta. Llama 3 model card. 2024. 4

- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 8
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. 4
- [4] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022. 3
- [5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. 4
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*, pages 49250–49267, 2023. 4
- [7] Google DeepMind. Gemini 2.0 Flash Thinking. <https://deepmind.google/technologies/gemini/flash-thinking/>. Accessed: 2025-04-11. 8
- [8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8
- [9] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36:31096–31116, 2023. 3
- [10] Xuming Hu. Multimodal Named Entity Recognition and Relation Extraction with Retrieval-Augmented Strategy. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3488–3488, Taipei Taiwan, 2023. ACM. 8
- [11] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 7
- [12] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. 4
- [13] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What’s “up” with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175. Association for Computational Linguistics, 2023. 3, 4, 5, 1
- [14] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1
- [15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 4
- [16] Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. BiVLC: Extending Vision-Language Compositionality Evaluation with Text-to-Image Retrieval. In *Advances in Neural Information Processing Systems*, pages 101880–101904, 2024. 3
- [17] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, 2024. 2
- [18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1
- [19] Navid Rajabi and Jana Kosecka. Gsr-bench: A benchmark for grounded spatial reasoning evaluation via multimodal llms. *arXiv preprint arXiv:2406.13246*, 2024. 3, 4
- [20] Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [21] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased Scene Graph Generation From Biased Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3713–3722, 2020. 8
- [22] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visiolinguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 3, 4
- [23] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. LLMs may perform MCQA by selecting the least incorrect option. In *Proceedings of the 31st International Conference on Computational Linguistics*,

pages 5852–5862, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. [2](#), [5](#)

- [24] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024. [3](#)
- [25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024. [4](#)
- [26] Xinyu Wang, Jiong Cai, Yong Jiang, Pengjun Xie, Kewei Tu, and Wei Lu. Named Entity and Relation Extraction with Multi-Modal Retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5925–5936. Association for Computational Linguistics, 2022. [8](#)
- [27] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024. [4](#)
- [28] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*. [3](#)
- [29] Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. Multimodal Relation Extraction with Efficient Graph Alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306, Virtual Event China, 2021. ACM. [8](#)
- [30] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, 2024. [1](#), [2](#), [4](#), [5](#)
- [31] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. [4](#)
- [32] Yongshuo Zong, Tingyang Yu, Ruchika Chavhan, Bingchen Zhao, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. In

Proceedings of the 41st International Conference on Machine Learning, pages 62892–62913, Vienna, Austria, 2024. JMLR.org. [2](#), [4](#), [7](#), [8](#)