Vision language models have difficulty recognizing virtual objects

Tyler Tran, Sangeet Khemlani, J. Gregory Trafton
{tyler.k.tran.civ, sangeet.s.khemlani.civ, greg.j.trafton.civ}@us.navy.mil
Navy Center for Applied Research in Artificial Intelligence
US Naval Research Laboratory
Washington, DC 20175 USA

Abstract

Vision language models (VLMs) are AI systems paired with both language and vision encoders to process multimodal input. They are capable of performing complex semantic tasks such as automatic captioning, but it remains an open question about how well they comprehend the visuospatial properties of scenes depicted in the images they process. We argue that descriptions of virtual objects – objects that are not visually represented in an image – can help test scene comprehension in these AI systems. For example, an image that depicts a person standing under a tree can be paired with the following prompt: `imagine that a kite is stuck in the tree`*. VLMs that comprehend the scene should update their representations and reason sensibly about the spatial relations between all three objects. We describe systematic evaluations of state-of-the-art VLMs and show that their ability to process virtual objects is inadequate.*

1. Introduction

The ability to imagine is what permits humans to reason beyond what they perceive [2, 33]: they can mentally rotate images of 3D objects to imagine them in different configurations [34]; they can animate the components of pulley systems and other physical devices [3, 17]; they can imagine traversals over maps, diagrams, and architectural drawings to extract relational information [21, 36]; they can imagine novel structures by mentally combining images of parts [14]. One theorist argues that nearly all forms of human perception engage imagination in some way [5]; another argues that human imagination serves the central generative functions of permitting creativity, hypothetical reasoning, and counterfactual analysis [1].

New advances in AI have produced systems that appear to possess human-like imaginative abilities: for instance, vision language models (VLMs), which are systems built on pre-trained transformers architectures and coupled with vision encoders, can process imagery and text simultaneously [38] and some researchers have investigated how they can be used to generate imaginary scenes [41] and configurations of objects [32]. These models are trained on webscale image caption corpora to extract visuospatial information from out-of-distribution images, yielding possible human-like performance on a large swath of visual tasks such as image tagging, automatic captioning, and autonomous driving [12, 19, 40]. Researchers debate the extent to which VLMs engage in robust spatial scene understanding [6, 8–10, 15, 16], especially given that they exhibit aberrant behavior that humans don't produce [29–31].

We argue that any general purpose scene processing system should be capable of visual imagination, i.e., imagining how an image would change given new information. Imaginative processing is particularly necessary for VLMs, which are built for "generalist" purposes [39] and vaunted for their versatility [20], since they can process text and imagery concurrently. If a VLM cannot perform a variety of rudimentary imaginative tasks on input imagery, it suggests that the model cannot encode the structure of a scene in a robust and productive way.

Consider a simple case of an image that depicts a person standing under a tree canopy. A VLM may be fed the following instruction: **`Imagine that a kite is stuck in the tree`**. In this situation, is the kite above or below the man? The kite is a *virtual object*, i.e., an object within a scene that is described but not depicted. Humans have no difficulty incorporating the new information to update their mental representations of the scene, and to thereby update their understanding of the relations between the two visual objects and the one virtual object. The answer should be equally trivial for VLMs: they should respond that the kite is *above* the man.

As we show, prompts concerning virtual objects can help test the multimodal capacities of VLMs and similar machine learning approaches. We describe an evaluation study of different VLMs and their capacity to recognize mentioned virtual objects in a scene. We first describe the

Abbreviation	Present tense	Past tense
"act"	**Act as though** there **is** a C next to the A; what items **are** on the table?	**Act as though** there **was** a C next to the A, what items **would be** on the table?
"assume"	**Assume** there **is** a C next to the A; what items **are** on the table?	**Assume** there **was** a C next to the A; what items **would be** on the table?
"consider"	**Consider that** there **is** a C next to the A, what items **are** on the table?	**Consider that** there **was** a C next to the A, what items **would be** on the table?
"if"	What items **are** on the table **if** there **is** a C cup next to the A?	What items **would be** on the table **if** there **was** a C cup next to the A?
"imagine"	**Imagine** there **is** a C next to the A; what items **are** on the table?	**Imagine** there **was** a C next to the A, what items **would be** on the table?
"pretend"	**Pretend** there **is** a C next to the A; what items **are** on the table?	**Pretend** there **was** a C next to the A; what items **would be** on the table?
"suppose"	**Suppose** there **is** a C next to the A; what items **are** on the table?	**Suppose** there **was** a C next to the A; what items **would be** on the table?

Table 1. Prompts used to evaluate virtual object recognition. The evaluation study varied the tense of these prompts (present vs. past) as well as whether they provided a numerical cue or not.

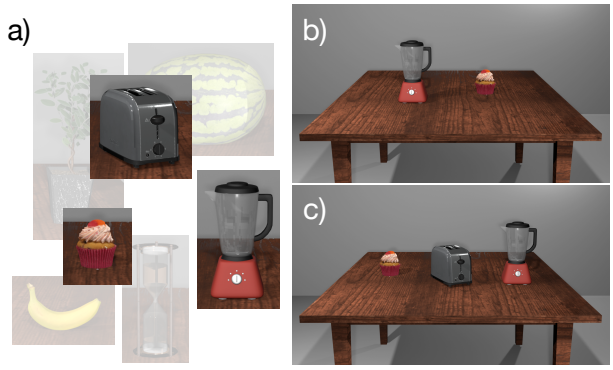

Figure 1. Examples from the TABLETEST dataset, which includes 64 individual objects (a) in various 2-object (b) and 3-object configurations (c).

dataset and the battery of prompts we used to benchmark virtual object recognition, and then describe the results of those evaluations – which reveal inadequate virtual object recognition for all VLMs under investigation.

2. Benchmarking methodology for testing virtual object recognition

TABLETEST is a dataset of synthetic imagery for investigating relational recognition and reasoning in VLMs [22]. The dataset consists of images of 1-3 objects arranged on a table next to one another (see Figure 1); it uses 64 objects from the Objaverse dataset of annotated 3D objects [13], hand-scaled to ensure sensible object sizes. It contains 4,032 2-object images and 250K 3-object images, constructed by creating all possible spatial configurations of the 64 objects.

We identified three candidate VLMs for benchmarking virtual object recognition based on the following criteria: they were recently released (post-2022), freely available, and capable of out-of-the-box, single-shot identification of the 64 objects in textscTableTest. Architecture sthat matched those criteria included: Idefics2 [27] (8B parameters), InstructBlip-Vicuna (7B parameters), which builds atop the BLIP architecture [28], and Llama 3.2 [35] (11B parameters).

The prompts used to assess virtual object recognition were minimal in nature, which allowed for systematic comparison and variation. Each prompt stipulated a hypothetical scenario that related a virtual object to one of the objects depicted in the image, e.g., **Imagine there is a banana next to the cupcake...**, and then queried for a list of the objects on the table. Successful responses are those – and only those – that list all three objects in any order.

Our evaluation study systematically manipulated the prompts along three dimensions: the manner in which the hypothetical was described; the tense of the prompt (past or present); and the manner in which the list of objects was queried. In theory, each of these dimensions and their variations should have no demonstrable effect on performance. We explain the study's manipulations further:

1. *Prompt variations.* We subjected each VLM to 7 prompt formulations that differed in the words used to create virtual objects. For instances, prompts queried VLMs to "assume" or "suppose" that a virtual object was next to an object depicted in the image (see Table 1). The names of virtual objects were those used to describe objects in the TABLETEST dataset, and were randomized for each evaluation.
2. *Tense.* Each of the prompt variations were formulated

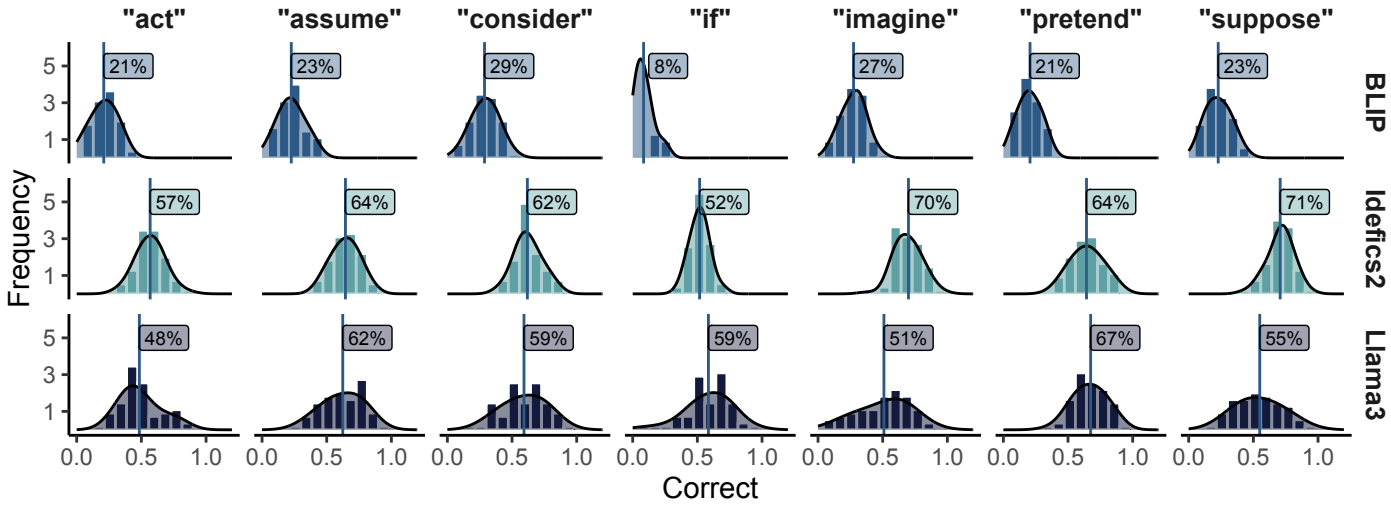

Figure 2. Proportions of accuracy from evaluations of 7 separate prompts concerning virtual object recognition in 2-object images from TABLETEST. Lines in each panel depict overall accuracies, density plots depict performance distributions across TABLETEST's 64 objects, and bars depict histograms across those objects, as organized by whether the object served as the leftmost object in images. Humanlike performance estimated at ceiling (accuracy = 1.0).

in English using either present tense or past tense (see Table 1). Since VLMs are often trained on image caption corpora, we hypothesized incidental asymmetries in how captions are described in those corpora, and that tense could have a significant effect on performance.

3. *Numerical cues.* Half of the prompts provided on evaluations queried for a list of items in a neutral way, e.g., "...what items are on the table?" The other half of the prompts provided a numerical cue, e.g., "...what three items are on the table?" We hypothesized that numerical cues should boost performance on this task.

In sum, we conducted an evaluation study in which we paired each 2-object image in TABLETEST with one of 7 different kinds of prompts $\times$ present- and past-tense versions of those prompts $\times$ queries that used numerical cues or not, yielding a total of 112,896 queries. Each of these queries were subjected to the 3 state-of-the-art VLMs. We used a fixed random seed for each evaluation and kept the temperature at 0 to ensure replicability.

3. Evaluation study results

Our evaluation study revealed that the VLMs under investigation produced inadequate virtual object recognition behavior for all of the 7 prompt formulations. Figure 2 plots the accuracies as a function of the 7 prompts and the different VLMs. In aggregate, Idefics2 produced 63% correct responses, Llama3 produced 57% correct responses, and BLIP produced 22% correct responses. We calculated mean accuracies for the different objects in TABLETEST and subjected them to nonparametric analyses to assess whether these differences were statistically reliable. They revealed significant differences in performance between the three VLMs (Friedman test, χ^2 = 99.08, $p < .001$). Likewise, the different prompts produced statistically reliable differences in accuracy (Friedman test, χ^2 = 120.91 $p < .001$); the "pretend" prompt produced the best performance (51% accuracy) and the "if" produced the worst (40% accuracy). As Figure 2 shows, the most aberrant pattern was BLIP's performance on the "if" prompt, which yielded only 8% correct responses.

The different tenses of the prompts affected performance: prompts in the past tense were more accurate than those in the present tense (51% vs. 44%; Wilcoxon test, $z = 6.84, p < .001$; see Figure 3, top panel). One tentative reason for this difference may be that the captions used to train VLMs – from websites and newspapers – may use past tense descriptions more often than present tense descriptions. Or it may be because past tense descriptions in corpora are incidentally longer and more concrete. Research into the data used to train these systems is necessary to evaluate these claims.

As hypothesized, numerical cues ("...what three items...") reliably boosted the performance of VLMs (62% correct with a numerical cue vs. 32% correct without; Wilcoxon test, $z = 6.96, p < .001$). This boost may be be-

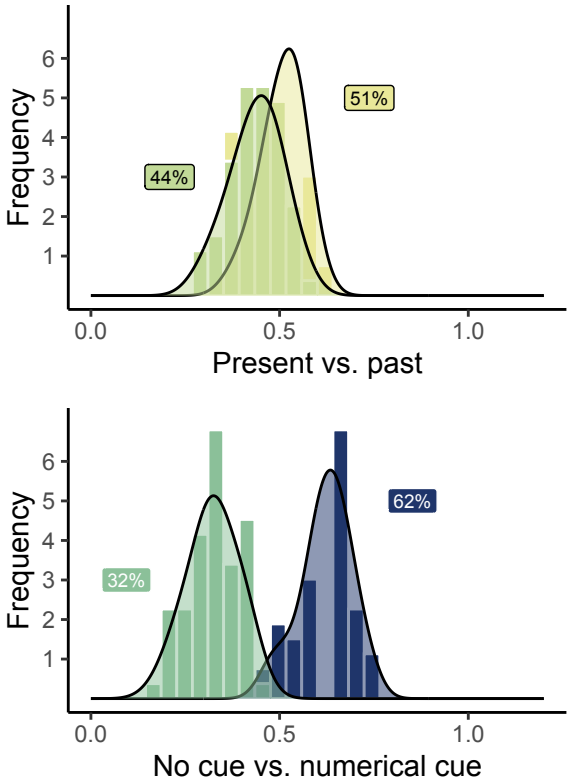

Figure 3. Proportions of accuracy from evaluations of **present** vs. **past** versions of prompts (top panel) and for **no cue** vs. **numerical cue** formulations; density plots depict performance distributions across 64 objects, and bars show accuracy histograms for those objects, as organized by the leftmost object in images. Humanlike performance estimated at ceiling (accuracy = 1.0).

cause the cue helps the VLM consider a virtual object and its relations in a scene; or it may be for altogether trivial reasons, e.g., the prompt mentions a number and the VLM attempts to produce a response that matches any and all nouns depicted in either the scene or described in the prompt.

We underscore that all of the prompts we used have correct – and trivial – answers: a VLM is correct if it can describe both the depicted objects in the image and the virtual object in the prompt. And suboptimal performance should not be perturbed by irrelevant factors, such as the tense of the prompt; indeed, none of the factors we tested should have affected whether a VLM can detect virtual objects. The results therefore suggests significant limitations in the capacity for VLMs to engage in hypothetical reasoning about objects not depicted in imagery.

4. Discussion

We ran an evaluation study to test the imaginative capacities of three state-of-the-art vision language models (VLMs). These systems provide integrated frameworks capable of multimodal analysis by tokenizing text and images and subjecting them to transformer architectures in parallel ways. The approach has produced new capabilities, such as the robust ability to highlight and label objects in images based on natural language queries. In theory, these systems should permit simple forms of imaginative processing as well, since text embeddings can yield updated representations of embeddings of images and vice versa [32, 38, 41]. Our investigations were designed to test a rudimentary form of imagination: they tasked VLMs with imagining a new "virtual" object in a scene of two objects, and then queried for a list of all the objects in the scene. A system capable of updating its representations appropriately should list all three objects. Our analyses show, however, that VLMs systematically lost track of the virtual objects and were perturbed by factors that should not affect processing, such as whether prompts were in the present or past tense.

The inability to track virtual objects suggests complementary limitations on more complex tasks. VLMs could be used for many simple forms of hypothetical and imaginative reasoning by querying for the system to consider: when one object is replaced with another; when it is moved relative to another; when its size or some other property is changed, and so on. If VLMs cannot perform these tasks, they cannot be said to possess reliable visuospatial reasoning capabilities, and so their usage must be circumscribed around those tasks for which they're suited.

How could AI architectures learn to track virtual objects? Unlike contemporary AI systems, humans integrate verbal and perceptual information by building sparse, discrete, abstract "mental models". They construct multiple models to imagine alternative spatial configurations [17, 21, 24, 36]. Mental models discard irrelevant perceptual details [4, 7, 25] to yield abstract, mutable structures, which permit rapid and flexible spatial reasoning of both visualizable and non-visualizable concepts [11, 26, 37]. But they demand piecemeal and serial manipulation of representations [18, 23], which makes human reasoners slower than AI systems at many visuospatial tasks. VLMs, in contrast, leverage parallel pipelines for processing text and image embeddings holistically, but they may have difficulty integrating the resulting distributed representations in coherent ways that permit rapid updating and analysis. Systems capable of human-like imaginative processing may have to create a synthesis of these approaches, e.g., by reasoning over both distributed and discretized structures.

References

[1] Anna Abraham. The imaginative mind. *Human brain mapping*, 37(11):4197–4211, 2016. 1

[2] Jessica R Andrews-Hanna and Matthew D Grilli. Mapping the imaginative mind: Charting new paths forward. *Current Directions in Psychological Science*, 30(1):82–89, 2021. 1

[3] Christopher J Bates, Iker Yildirim, Joshua B Tenenbaum, and Peter W Battaglia. Humans predict liquid dynamics using probabilistic simulation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2015. 1

[4] Eric J Bigelow, John P McCoy, and Tomer D Ullman. Non-commitment in mental imagery. *Cognition*, 238:105498, 2023. 4

[5] Derek H Brown. Infusing perception with imagination. *Perceptual imagination and perceptual memory*, pages 133–160, 2018. 1

[6] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024. 1

[7] Patrick Cavanagh. The artist as neuroscientist. *Nature*, 434 (7031):301–307, 2005. 4

[8] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1

[9] Liangyu Chen, Bo Li, Sheng Shen, Jingkang Yang, Chunyuan Li, Kurt Keutzer, Trevor Darrell, and Ziwei Liu. Large language models are visual reasoning coordinators. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. *arXiv preprint arXiv:2406.01584*, 2024. 1

[11] Robert A Cortes, Adam B Weinberger, Griffin A Colaizzi, Grace F Porter, Emily L Dyke, Holly O Keaton, Dakota L Walker, and Adam E Green. What makes mental modeling difficult? normative data for the multidimensional relational reasoning task. *Frontiers in psychology*, 12:668256, 2021. 4

[12] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. 1

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2

[14] Ronald A Finke. *Creative imagery: Discoveries and inventions in visualization*. Psychology press, 2014. 1

[15] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 1

[16] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024. 1

[17] Mary Hegarty. Mental animation: Inferring motion from static displays of mechanical systems. *Journal of experimental psychology: learning, memory, and cognition*, 18(5): 1084, 1992. 1, 4

[18] Mary Hegarty. Components of spatial intelligence. In *Psychology of learning and motivation*, pages 265–297. Elsevier, 2010. 4

[19] Hartwig H Hochmair, Levente Juhász, and Takoda Kemp. Correctness comparison of chatgpt-4, gemini, claude-3, and copilot for spatial tasks. *Transactions in GIS*, 2024. 1

[20] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002, 2024. 1

[21] Philip N Johnson-Laird. Imagery, visualization, and thinking. *Perception and cognition at century's end*, pages 441–467, 1998. 1, 4

[22] Sangeet Khemlani, Tyler Tran, Nathaniel Gyory, Anthony M. Harrison, Wallace E. Lawson, Ravenna Thielstrom, Hunter Thompson, Taaren Singh, and J. Gregory Trafton. Vision language models are unreliable at trivial spatial cognition. In *Proceedings of the International Conference on Computer Vision 2025*, 2025. 2

[23] Sangeet Suresh Khemlani, Robert Mackiewicz, Monica Bucciarelli, and Philip N Johnson-Laird. Kinematic mental simulations in abduction and deduction. *proceedings of the national academy of sciences*, 110(42):16766–16771, 2013. 4

[24] Markus Knauff. *Space to reason: A spatial theory of human thought*. Mit Press, 2013. 4

[25] Markus Knauff and Phil N Johnson-Laird. Visual imagery can impede reasoning. *Memory & cognition*, 30:363–371, 2002. 4

[26] Maria Kon and Sangeet Khemlani. How spatial simulations distinguish "tracking" verbs. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2024. 4

[27] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 2

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[29] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024. 1

[30] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

[31] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1

[32] Jingming Liu, Yumeng Li, Boyuan Xiao, Yichang Jian, Ziang Qin, Tianjia Shao, Yao-Xiang Ding, and Kun Zhou. Enhancing visual reasoning with autonomous imagination in multimodal large language models. *arXiv preprint arXiv:2411.18142*, 2024. 1, 4

[33] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature reviews neuroscience*, 20(10):624–634, 2019. 1

[34] Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. 1

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[36] Barbara Tversky. Visuospatial reasoning. In *Handbook of reasoning*, pages 209–249. Cambridge University Press, 2005. 1, 4

[37] Tomer D Ullman, Andreas Stuhlmüller, Noah D Goodman, and Joshua B Tenenbaum. Learning physical parameters from dynamic scenes. *Cognitive psychology*, 104:57–82, 2018. 4

[38] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023. 1, 4

[39] Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024. 1

[40] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[41] Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2410.09874*, 2024. 1, 4