OPTiCAL: An Abstract Positional Reasoning Benchmark for Vision Language Models

Christopher Driggers-Ellis, Gabriel Ayoubi, Christan Grant
Department of Computer & Information Sciences & Engineering
University of Florida
Gainesville, FL, USA
{driggersellis.cw, gabriel.ayoubi, christan}@ufl.edu

***Abstract*—Vision question answering (VQA) tasks increasingly employ Visual Language Models (VLMs), but the performance of these models degrades substantially when applied to out-of-distribution or compositional reasoning tasks. This is especially concerning with wide access to pretrained VLMs, which could lead to misuse and overdependence on the reasoning capabilities of these models. In this work, we analyze the root causes of poor VLM performance by isolating and testing basic positional reasoning skills using a novel benchmarking dataset, *Shapes30k*, generated by our tool, *ShapeMaker*. Our main metric is VLM accuracy in the positional reasoning task, and we perform significance testing to detect directional bias in the results. Pretrained VLMs sometimes score below chance (20%) in our benchmark, and we detect varied and significant ($p < 0.01$) directional biases in each model. Our code and data are available here: https://github.com/ufdatastudio/optical**

***Index Terms*—Benchmark testing, Deep learning, Vision transformers, Large language models**

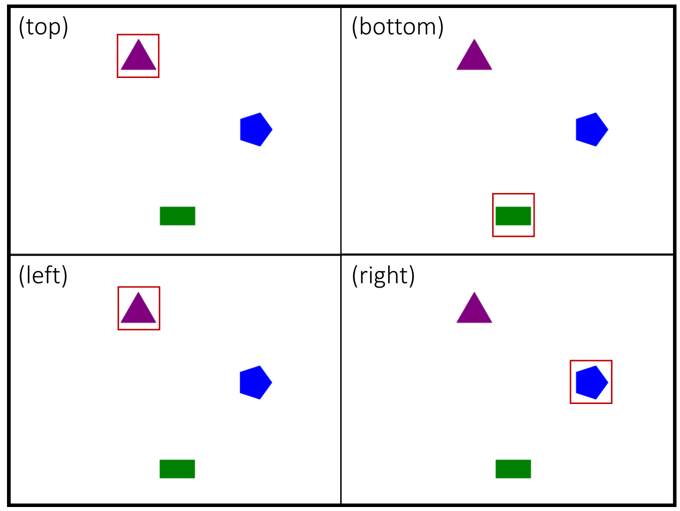

Fig. 1. An image from Shapes30k repeated four times to demonstrate the task directions in our benchmark. VLMs are prompted with the question **"Which shape is on __"**; and the direction indicators ***top***, ***the bottom***, ***the left***, or ***the right*** fill the blank. Answers are marked here with red boxes: "triangle" (top, left), "rectangle" (bottom), and "pentagon" (right).

I. Introduction

Multimodal Large Language Models and other Vision Language Models (VLMs) are applied to a variety of tasks, including visual question and answering (VQA). Low performance plagues this VQA task in numerous questioning contexts and models [1], [2]. This low performance is especially concerning in light of the increasing availability of pretrained, open-source VLMs and LLMs through free APIs, for this easy access is a vector for application of models to highly specialized, reasoning-intensive tasks.

We detect a need to understand VLM and LLM reasoning beyond performance in downstream tasks. Instead, because pre-trained models may be deployed and fail unpredictably, we must understand VLM and LLM reasoning in the abstract and how abstract reasoning correlates with performance on grounded inference tasks. Thus, we benchmark VLM reasoning capabilities with basic, abstract composition with samples like Figure 1. We also emphasize an urgent need for abstract understanding in light of harms that have already occurred. For example, CVE records a critical vulnerability in the row-level database security policies of websites generated by the vibe coding platform Lovable wherein websites permit arbitrary read-write access to database tables [3]. When such vulnerabilities come to light, it doubtlessly impacts user trust and poses a severe liability for developers.

Because pretrained VLM and LLM models can be deployed in unpredictable ways, current benchmarking methods do not provide adequate coverage for these deployments. Users do not deploy models in the controlled way VLMs are typically benchmarked, with consistent levels or types of visual grounding, so we must understand performance with abstract data and investigate whether the controls present in other benchmarks.

This investigation provides a methodology to shed light on compositional reasoning, an area where VLMs must improve, and provides baseline results for future investigations into abstract reasoning tasks for VLMs. In the following sections, we investigate VLM performance in a VQA positional reasoning task using a dataset that isolates positional understanding of VLM models because it lacks visual grounding similar to typical benchmarking or pre-training datasets like MSCOCO, and we demonstrate that VLM models' performance in this task is consistent with previous findings about VLM positional and compositional reasoning.

As first steps, we provide the following contributions:

- Shapemaker, script for generating Shapes30k, a customizable image benchmark of shapes in front of a white or transparent background.

- A benchmarking experiment wherein six VLMs perform an abstract visual reasoning task on Shapes30k.
- Shapes30k, an abstract dataset consisting of 30000 images of shapes on transparent backgrounds.

II. Background

Hallucination in VLMs is a common problem observed while performing a variety of tasks. Object hallucinations occur when models incorrectly classify an object's category, attributes, or its relationship with other objects and are usually studied in image captioning and VQA tasks [4], [5]. Some research investigates the causes of object hallucination in depth. Experiments on the pretrained CLIP models ubiquitous in VLMs suggest that CLIP's objective during contrastive training does not require the model to differentiate between fine details in images and that this can lead to object hallucination [6].

CLIP models often act like a ***bag-of-words***, meaning that they do not manage well with reasoning about the attributes or relationships of objects that form an image's composition, and research into this problem blames the contrastive training CLIP models receive [7]. In this work, we examine many VLMs with pretrained CLIP vision encoders, and successful completion of the task in our benchmark requires the model to correctly ascertain the composition of abstract shapes on a blank background. Because we lack visual grounding to confound the results by providing the models with additional visual grounding information, we expect the bag-of-words phenomenon to dominate.

CLIP training datasets are compositionally rich, but compositional (object relationship) understanding is unnecessary for CLIP models following the present contrastive training strategy [7]. CLIP models are trained and evaluated with conventional datasets, consisting of real-world scenes and objects, where visual grounding enables shortcuts in compositional reasoning such as falling back on associations from the pretraining data.

Previous efforts to benchmark VLM visual reasoning, such as ARO [7]; SPHERE [8]; and Winoground [9], use real-world datasets similar to the CLIP models' training data. Image sources for these benchmarks include Flick30k, COCO, and Getty Images [7]–[9]. The visual grounding available in such images is similar to the CLIP training data and therefore enables many of the same shortcuts. CLEVR [10] approaches the abstractness of the OPTiCAL benchmark because its samples are 3D geometric shapes positioned on blank surfaces. While superficially similar to our work, the images are 3D renders of geometric shapes in an environment with realistic lighting. Some of the cues leaned on during training with real world images, such as shadows; relative size of objects; and reflections, remain available in CLEVR [10]. The flat, 2D benchmark we propose contains no such cues and better isolates the raw positional reasoning ability of VLMs.

Two aspects of complex tasks and data stand as confounding forces against unraveling relational reasoning deficiencies in VLMs. Complex tasks can fail because the VLM fails a subtask other than understanding object relationships, and empirical evidence suggests CLIP models perform inadequately without compositional understanding [7]. We study VLM performance in a VQA task but use an unconventional, abstract dataset to uncover the faults in VLM reasoning that could explain object hallucinations and low compositional reasoning.

III. Experiment

The experiment begins by constructing a dataset generator that constructs images of s randomly positioned shapes on a $n \times n$ grid of plots and saves them in PNG format. For the procedure in this research, the generator is utilized to construct a dataset consisting of $30,000$ such images with 3 shapes on a white background of size 5×5. To disambiguate the task in our experiment, the generator does not place multiple shapes on the same horizontal or vertical coordinate, and only one instance of each shape may appear in a given image. Five shape types are included in the dataset. These are *triangle, square, rectangle, pentagon*, and *circle*.

The dataset design follows rationales inspired by the related work. We have designed the dataset in this benchmark to lack grounding information in order to isolate reasoning ability from inferences the models. Models can use grounding information in real-life images to fallback on patterns in their original training data rather than to reason about the positions and relationships between objects in the inference image. As a result, a key contribution of this benchmark are therefore insights into encoder-decoder VLMs' ability to reason about of the positions and relationships of objects in an image without the confounding effect of image grounding.

We study a set of similar positional reasoning tasks. For each of the $30,000$ generated images, the VLM is asked to determine which of the shapes is to the *left, right, top*, or *bottom*. Thus, there are 5 possible answers for each task and 4 task types. An ensemble of VLMs perform each of the tasks on the dataset; the following VLMs are used: Blip-FlanT5 [11], CogVLM [12], CogVLM2 [13], InstructBlip-Vicunna7B [14], LLaVA1.6 [15], and PaliGemma2 [16]. We name the dataset Shapes30k and the script used to generate it the ShapeMaker. Utilizing Shapes30k as a benchmark, we perform the following procedure on 6 open source VLMs accessed through Hugging Face (HF) APIs (license terms available on HF).

All experiments are performed with 2 A100 GPUs and 16 CPUs. We load the VLMs, and present each of the models with the same prompt-image pairs. We record responses and compare them to image labels. For the given tasks, the answer is a single word, the name of the shape in a given direction relative to the others, and the VLM is prompted to answer with just the name of that shape, although it is not told what the possible answers are. We measure accuracy by counting exact matches of the casefold of the response and label and dividing the number of exact matches by the total number of images. Finally, we use the two-way Fisher's exact test to detect directional bias in VLM performance.

TABLE I

ACCURACY OF HF MODELS BY TASK. COLUMNS WITH TASK NAMES REPORT A CALCULATION OF ACCURACY ONLY FOR RESPONSES RESPONDING TO THAT TASK'S PROMPT. BEST PERFORMANCE FOR EACH COLUMN IN **BOLD**, AND RED INDICATES PERFORMANCE UNDER THE CHANCE VALUE (20%).

HF model/task	*all*	*left*	*right*	*top*	*bottom*
blip2-flan-t5-xl [11]	0.117	0.112	0.144	0.0539	0.160
cogvlm-chat-hf [12]	**0.634**	**0.508**	**0.623**	**0.618**	**0.786**
cogvlm2-llama3-chat-19B [13]	0.592	0.505	0.535	0.606	0.722
instructblip-vicuna-7b [14]	0.315	0.278	0.289	0.343	0.350
llava-v1.6-mistral-7b-hf [15]	0.566	0.500	0.571	0.602	0.593
paligemma2-10b-pt-224 [16]	0.410	0.403	0.348	0.405	0.483

TABLE II

ACCURACY OF HF MODELS. COLUMNS WITH SHAPE NAMES REPORT A CALCULATION OF ACCURACY ONLY FOR RESPONSES WHERE THAT SHAPE WAS THE ANSWER. BEST PERFORMANCE FOR EACH COLUMN IN **BOLD**, AND RED INDICATES PERFORMANCE UNDER THE CHANCE VALUE (20%).

HF model/task	*all*	*circle*	*pentagon*	*rectangle*	*square*	*triangle*
blip2-flan-t5-xl [11]	0.117	0.000	0.000	0.000	0.0924	0.490
cogvlm-chat-hf [12]	**0.634**	**0.926**	0.00691	**0.653**	0.811	0.781
cogvlm2-llama3-chat-19B [13]	0.592	0.825	0.0961	0.563	0.530	0.949
instructblip-vicuna-7b [14]	0.315	0.170	0.000	0.000	0.399	**0.999**
llava-v1.6-mistral-7b-hf [15]	0.566	0.910	0.000	0.0534	**0.885**	0.982
paligemma2-10b-pt-224 [16]	0.410	0.639	**0.126**	0.148	0.168	0.961

We benchmark a collection of models with pre-trained CLIP or other contrastive vision encoders, so the procedure we outline should provide data on their performance within a variety of encoder-decoder architectures. The number of samples available to us for use in the benchmark is unrestricted, so we choose a large number to provide the greatest chance that any patterns in the performance data pose significant results.

IV. RESULTS

Table I displays overall accuracy and accuracy per task direction for all six VLMs. Most models score 40% to 60% accuracy. CogVLM [12] scores the greatest overall accuracy at 63.4%. Paradoxically, the newer, related model CogVLM2 [13] has lower overall performance. LLaVA-1.6 [15] performs third best and is the last model whose overall accuracy in the task is greater than 50%. Flan-T5 [11] scores below chance (20%) for all task directions. There is also significant ($p < 0.01$) directional bias in the accuracy of each model. Table II displays overall accuracy again, and accuracy when specific shapes were the answer. Columns display accuracy for cases where the labeled shape was the answer, and the overall column is the same as in Table I. Models perform below chance much more frequently whenever specific shapes are the answer than in the overall case.

V. DISCUSSION

Despite its simplicity, models struggle with the positional reasoning task put before them in our experiment. The task is only a matter of recognizing the sample image's composition, and the "noise" present in real-world images is absent in the data we use for our experiment. We must question why VLMs *incorrectly* identify the shape about one out of three times in the best case. Alarmingly, Flan-T5 [11] scores below chance (20%) in for each direction, and every model in the experiment performs below chance for some shape whenever that shape is the answer.

CogVLM's [12] improved performance over the updated CogVLM2 [13] is interesting. While it would be outside the scope of this investigation to determine the cause of performance discrepancies, it is possible that differences in model architectures could explain this and other performance differences.

In Table II, we observe that, when *pentagon* was the answer, three out of six models studied achieved a 0% accuracy in our positional reasoning task, meaning that, in 30000 trials, the model did not once correctly identify a pentagon when it was the answer. The models often stated *hexagon* as their answer instead, whereas there were no hexagons present in the dataset used for this experiment. It appears that models are not able to see the pentagons in our dataset and frequently hallucinate hexagons that were not present in the original data.

We also identify a significant ($p < 0.01$) directional bias in task accuracy for each model studied in at least four out of six directional pairs and provide these tests in Appendix A. The results are concerning because consistent bias explains the differences in model performance across task directions, and lack of a consistent pattern in the biases suggests that explanations of the biases differ by model.

VI. CONCLUSIONS

The VLMs frequently suffer from object hallucination and fail at positional reasoning. The common misidentification of pentagons as hexagons underscores a significant limitation in current VLMs. The models we benchmark do not perceive spatial relationships between objects accurately and cannot correctly identify certain shapes. The models we test perform poorly, sometimes worse than chance (20%), on the basic positional reasoning task. The consistency of the results combined with the noiselessness of the data indicate that the hallucinations observed are not outliers or symptoms of distraction caused by extraneous input features but rather symptoms of a fundamental weakness in decoder-encoder VLMs and is consistent with the hypothesis that VLM utilize cues in real-world image data in a positive way.

Our findings show high rates of category object hallucination and suggest that models cannot determine the relative positions of abstract shapes in the abstract images. These findings are likewise consistent with previous work that suggests that CLIP vision encoders, which are central to most encoder-decoder VLMs, struggle with spatial understanding due to the limitations of their contrastive training objectives rather than confusion of visual grounding. Additionally, our work reveals a high directional bias in the outputs of the different models evaluated. Bias varies greatly between different models, and the source of biases and disparities in bias cannot be traced with the current data, although architectural differences appear to play a role.

Improving the performance of VLMs in spatial reasoning tasks will require hallucination mitigation techniques that improve preservation of objection relationships from the original image in the text embedding space. These findings inspire emphasis on embedding-aware design and evaluation of abstract spatial reasoning performance prior to deployment for grounded tasks, especially prior to their deployment in tasks like grounded language inference and VQA.

VII. Future Work

Future work must implement and test relevant VLMs and hallucination mitigation techniques related works offer on a variety of VLMs using an abstract benchmark similar to the one in our experiment. Our investigation is largely limited encoder-decoder models with CLIP vision encoders, but the visual reasoning ability non-contrastive architectures should be investigated in abstract settings. Although our results reflect what is already reported about CLIP encoders in the VLM literature, non-CLIP models may also struggle with positional reasoning in the same contexts, albeit for different reasons.

Because we are unaware of any mechanism which secures the link between increased performance in OPTiCAL with other visual reasoning benchmarks, we also propose correlational studies of OPTiCAL performance with performance on other benchmarks to determine if a performance correlation exists and whether changes to model architectures that improve performance in other visual reasoning benchmarks promote corresponding improvements in OPTiCAL.

Investigating each of the important architectures, models, and methods in the abstract setting poses a daunting endeavor for any one team, and in light of the seriousness of the VLM reasoning pitfalls already established, we prioritize spreading awareness of the issues and open the door for new perspectives by leaving the wider exploration of methods and models in abstract positional reasoning benchmarks as future work.

VIII. Limitations

We lack evidence that increases in performance on fundamental reasoning tasks will translate to increased performance in downstream tasks. Concretely, we cannot show that improvements on our abstract reasoning benchmark will translate to increases in performance on benchmarks for VQA, visual inference, and similar tasks. The additional grounding in images for those downstream tasks could unexpectedly confound mitigation techniques used to improve upstream performance.

We observe the bag-of-words phenomenon, for the models are unable to reason about object relationships, but we cannot establish a cause for object hallucination in our VLMs. Further, significance testing for directional bias indicates different directional biases exist for each individual model that should be considered further. Although our benchmarking results are consistent with previous reports about CLIP vision encoders and encoder-decoder architectures, something else is at work in each of the models. Hallucination mitigation strategies will likely need tuning to specific models as a result.

References

[1] M. Mitchell, A. B. Palmarini, and A. Moskvichev, "Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks," 2023.

[2] P. Verma, M.-H. Van, and X. Wu, "Beyond human vision: The role of large vision language models in microscope image analysis," 2024.

[3] N. I. of Standards and Technology, "Cve-2025-48757 detail," 2025.

[4] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," 2019.

[5] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou, "Hallucination of multimodal large language models: A survey," 2025.

[6] Y. Liu, T. Ji, C. Sun, Y. Wu, and A. Zhou, "Investigating and mitigating object hallucinations in pretrained vision-language (clip) models," 2024.

[7] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?," 2023.

[8] W. Zhang, W. E. Ng, L. Ma, Y. Wang, J. Zhao, A. Koenecke, B. Li, and W. Wanglu, "SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, eds.), (Vienna, Austria), pp. 11591–11609, Association for Computational Linguistics, July 2025.

[9] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, " Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality ," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 5228–5238, IEEE Computer Society, June 2022.

[10] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, " CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning ," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Los Alamitos, CA, USA), pp. 1988–1997, IEEE Computer Society, July 2017.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023.

[12] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, and J. Tang, "Cogvlm: Visual expert for pretrained language models," 2023.

[13] W. Hong, W. Wang, M. Ding, W. Yu, Q. Lv, Y. Wang, Y. Cheng, S. Huang, J. Ji, Z. Xue, L. Zhao, Z. Yang, X. Gu, X. Zhang, G. Feng, D. Yin, Z. Wang, J. Qi, X. Song, P. Zhang, D. Liu, B. Xu, J. Li, Y. Dong, and J. Tang, "Cogvlm2: Visual language models for image and video understanding," 2024.

[14] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[15] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[16] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long, S. Qin, R. Ingle, E. Bugliarello, S. Kazemzadeh, T. Mesnard, I. Alabdulmohsin, L. Beyer, and X. Zhai, "Paligemma 2: A family of versatile vlms for transfer," 2024.

APPENDIX

TABLE III

p-VALUES FROM TWO-WAY FISHER'S EXACT TEST ON PAIRED TASK TYPES PERFORMED BY SALESFORCE/BLIP2-FLAN-T5-XL [11]. INSIGNIFICANT RESULTS ($p \geq 0.01$) IN RED.

key pairs	left	right	top	bottom
left	–	–	–	–
right	5.58e-10	–	–	–
top	4.43e-37	3.72e-79	–	–
bottom	8.42e-19	0.00836	7.36e-102	–

TABLE IV

p-VALUES FROM TWO-WAY FISHER'S EXACT TEST ON PAIRED TASK TYPES PERFORMED BY THUDM/COGVLM-CHAT-HF [12]. INSIGNIFICANT RESULTS ($p \geq 0.01$) IN RED.

key pairs	left	right	top	bottom
left	–	–	–	–
right	1.17e-45	–	–	–
top	2.68e-42	0.579	–	–
bottom	5.24e-283	5.46e-107	2.54e-112	–

TABLE V

p-VALUES FROM TWO-WAY FISHER'S EXACT TEST ON PAIRED TASK TYPES PERFORMED BY THUDM/COGVLM2-LLAMA3-CHAT-19B [13]. INSIGNIFICANT RESULTS ($p \geq 0.01$) IN RED.

key pairs	left	right	top	bottom
left	–	–	–	–
right	0.000236	–	–	–
top	7.56e-36	1.23e-18	–	–
bottom	4.19e-166	2.18e-125	3.44e-51	–

TABLE VI

p-VALUES FROM TWO-WAY FISHER'S EXACT TEST ON PAIRED TASK TYPES PERFORMED BY SALESFORCE/INSTRUCTBLIP-VICUNA-7B [14]. INSIGNIFICANT RESULTS ($p \geq 0.01$) IN RED.

key pairs	left	right	top	bottom
left	–	–	–	–
right	0.147	–	–	–
top	1.29e-17	1.44e-12	–	–
bottom	6.35e-21	2.38e-15	0.400	–

TABLE VII

p-VALUES FROM TWO-WAY FISHER'S EXACT TEST ON PAIRED TASK TYPES PERFORMED BY LLAVA-HF/LLAVA-V1.6-MISTRAL-7B-HF [15]. INSIGNIFICANT RESULTS ($p \geq 0.01$) IN RED.

key pairs	left	right	top	bottom
left	–	–	–	–
right	3.96e-18	–	–	–
top	3.87e-36	0.000112	–	–
bottom	1.51e-30	0.00515	0.287	–

TABLE VIII

p-VALUES FROM TWO-WAY FISHER'S EXACT TEST ON PAIRED TASK TYPES PERFORMED BY GOOGLE/PALIGEMMA2-10B-PT-224 [16]. INSIGNIFICANT RESULTS ($p \geq 0.01$) IN RED.

key pairs	left	right	top	bottom
left	–	–	–	–
right	1.79e-12	–	–	–
top	0.880	5.34e-13	–	–
bottom	1.31e-22	1.46e-63	5.72e-22	–