OPTiCAL: An Abstract Positional Reasoning Benchmark for Vision Language Models

Christopher Driggers-Ellis, Gabriel Ayoubi, Christan Grant

University of Florida Data Studio | ufdatastudio.com

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.

Background

Investigations have observed a **bag-of-words** phenomenon in vision language models (VLMs) [1]. The models understand what objects are in images but fail in complex compositional and positional reasoning tasks.

The contrastive CLIP vision transformers (ViTs) that serve as image encoders for many VLMs are not trained in a way that requires them to discriminate between fine details in an image [1], [2]. This could be the cause of shortcomings in VLM performance in positional and compositional reasoning tasks.

Additionally, high levels of visual grounding in real-world image datasets may allow models to draw inferences based on image grounding rather than actual reasoning abilities, confounding positional or compositional reasoning benchmarks that utilize real-world image data.

Figure 1: The shapes used in our method. From left to right, *circle*, *pentagon*, *rectangle*, *square*, and *triangle*.

Experiment

To test whether the bag-of-words persists or worsens without visual grounding, we devise the ShapeMaker to generate abstract benchmarks containing image-task pairs. The OPTiCAL benchmark we propose consists of the Shapes30k dataset and an equal number of positional reasoning tasks. The images are generated by the ShapeMaker and consist of images like the one in Figure 1 paired with questions asking models to identify which shape is most to one of the cardinal directions: *top*, *bottom*, *left*, and *right*.

ShapeMaker can be parameterized to generate other benchmarks with different shapes and distributions of shapes, but in the benchmark generated for this paper (Shapes30k), there are 30,000 images that each contain 3 different shapes arranged in an unmarked 5x5 grid on a transparent background. The shapes are placed such that no two out of the three shapes can be in the same column or row of the grid to ensure that there is always an unambiguous answer. Figure 1 shows each of the shapes in our benchmark as they appear in our data.

Models are shown image-task pairs and asked to answer with only the name of the shape that is most to the given direction. The question asked is shown in Figure 2's caption.

Use the QR Code
to view OPTiCAL!

Herbert Wertheim College of Engineering
Department of Computer & Information Science & Engineering
UNIVERSITY *of* FLORIDA

DATA STUDIO

Our Data

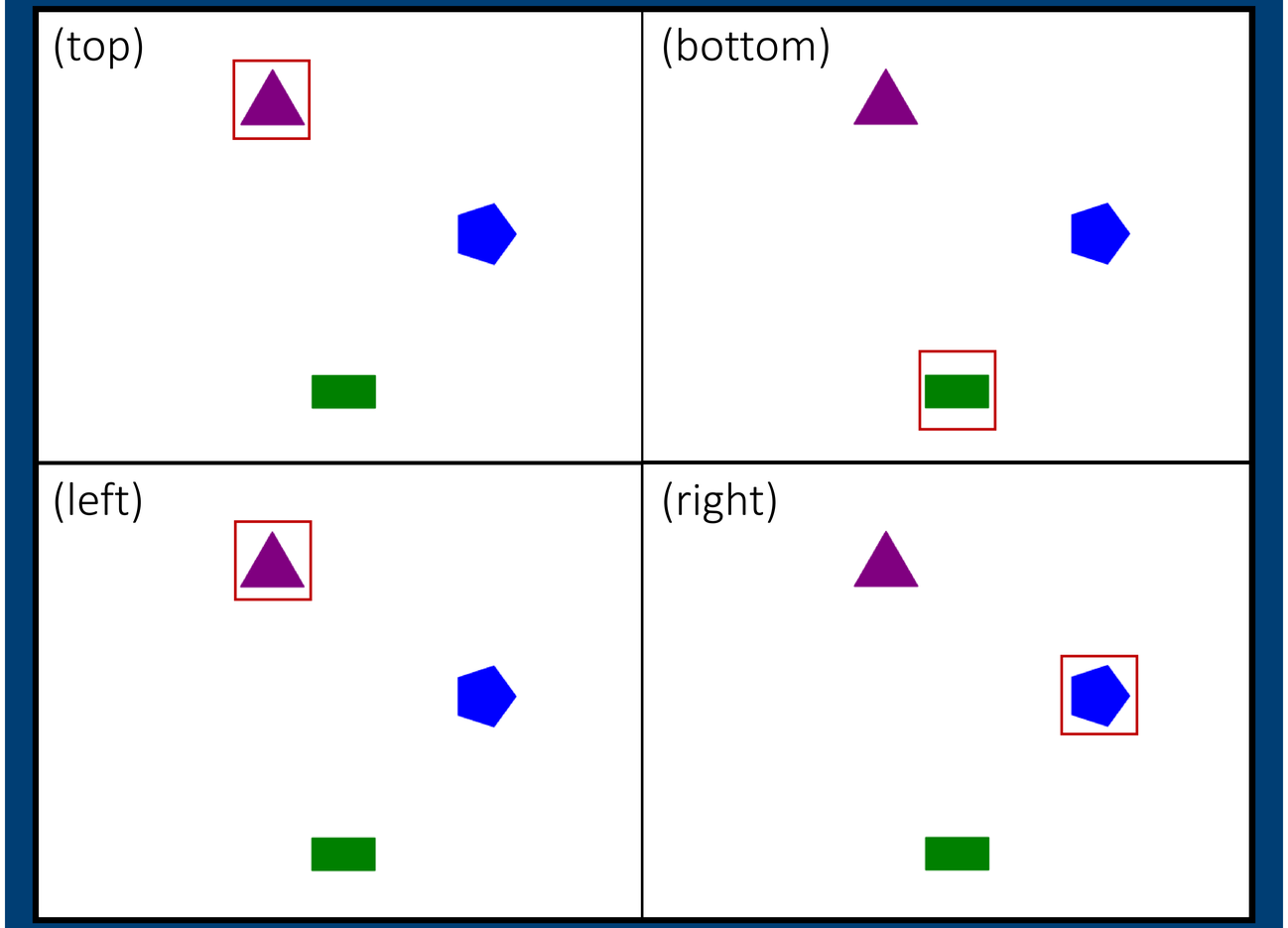

Figure 2: An image from Shapes30k repeated four times to demonstrate the task directions in our benchmark. VLMs are prompted with the question **"Which shape is on __"**; and the direction indicators ***top***, ***the bottom***, ***the left***, or ***the right*** fill the blank. Answers are marked here with red boxes: "triangle" (top, left), "rectangle" (bottom), and "pentagon" (right).

Results & Discussion

Tables 1 and 2 show the results of our benchmark for various freely-available vision language modesls (VLMs) accessed via hugging face (HF).

We identify significant ($p < 0.01$) directional performance bias for at least four out of six direction pairs in each of the models studied. The directional bias is different for different models and indicates that each model is better at identifying shapes in some directional extremes better than in others. The cause of directional bias is an open question.

HF model/task	*all*	*left*	*right*	*top*	*bottom*
blip2-flan-t5-xl [3]	0.117	0.112	0.144	0.0539	0.160
cogvlm-chat-hf [4]	**0.634**	**0.508**	**0.623**	**0.618**	**0.786**
cogvlm2-llama3-chat-19B [5]	0.592	0.505	0.535	0.606	0.722
instructblip-vicuna-7b [6]	0.315	0.278	0.289	0.343	0.350
llava-v1.6-mistral-7b-hf [7]	0.566	0.500	0.571	0.602	0.593
paligemma2-10b-pt-224 [8]	0.410	0.403	0.348	0.405	0.483

Table 1: 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*	*circle*	*pentagon*	*rectangle*	*square*	*triangle*
blip2-flan-t5-xl [3]	0.117	0.000	0.000	0.000	0.0924	0.490
cogvlm-chat-hf [4]	**0.634**	**0.926**	0.00691	**0.653**	0.811	0.781
cogvlm2-llama3-chat-19B [5]	0.592	0.825	0.0961	0.563	0.530	0.949
instructblip-vicuna-7b [6]	0.315	0.170	0.000	0.000	0.399	**0.999**
llava-v1.6-mistral-7b-hf [7]	0.566	0.910	0.000	0.0534	**0.885**	0.982
paligemma2-10b-pt-224 [8]	0.410	0.639	**0.126**	0.148	0.168	0.961

Table 2: 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%).

Limitations

We lack evidence that increased performance in our abstract benchmark will correlate with increased performance on concrete tasks. Additional grounding in normal VQA tasks could confound any measures taken to improve performance on the abstract benchmark.

While we observe the bag-of-words phenomenon, we cannot establish the causes of high object hallucination or of poor performance on the abstract benchmark in the VLMs studied. Significant ($p < 0.01$) directional biases along different axes exist in each of the VLMs, also. Though our analysis provides further evidence of bag-of-words, something else causes this directional bias; and different mitigation techniques may need to be applied to different VLMs.

References

[1] 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. arXiv: 2210.01936 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2210.01936.

[2] Y. Liu, T. Ji, C. Sun, Y. Wu, and A. Zhou, *Investigating and mitigating object hallucinations in pretrained vision-language (clip) models*, 2024. arXiv: 2410.03176 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2410.03176.

[3] 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. arXiv: 2301.12597 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2301.12597.

[4] W. Wang et al., "Cogvlm: Visual expert for pretrained language models," 2023. arXiv: 2311.03079 [cs.CV].

[5] W. Hong et al., *Cogvlm2: Visual language models for image and video understanding*, 2024. arXiv: 2408.16500 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2408.16500.

[6] W. Dai et al., *Instructblip: Towards general-purpose vision-language models with instruction tuning*, 2023. arXiv: 2305.06500 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2305.06500.

[7] H. Liu, C. Li, Y. Li, and Y. J. Lee, *Improved baselines with visual instruction tuning*, 2023. arXiv: 2310.03744 [cs.CV].

[8] A. Steiner et al., *Paligemma 2: A family of versatile vlms for transfer*, 2024. arXiv: 2412.03555 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2412.03555.

Paper ID: S19205