OPTiCAL: An Abstract Positional Reasoning Benchmark for Vision Language Models

Anonymous Author(s)
Affiliation
Address
email

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 visual reasoning skills—specifically, positional understanding—using a novel benchmarking dataset, Shapes30k, generated by our tool, ShapeMaker. Our primary 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 is available here: `https://anonymous.4open.science/r/optical-benchmark-DAE9/`

1 Introduction

Multimodal LLMs 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 therefore detect a fundamental 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].

We provide the following contributions:

- A Python script for generating a scalable image benchmark of shapes in front of a white or transparent background, dubbed ShapeMaker.
- A benchmarking experiment wherein six VLMs perform an abstract visual reasoning task on data generated by ShapeMaker, dubbed Shapes30k.
- The benchmark Shapes30k generated by the ShapeMaker for the benchmarking experiment, available with our code.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

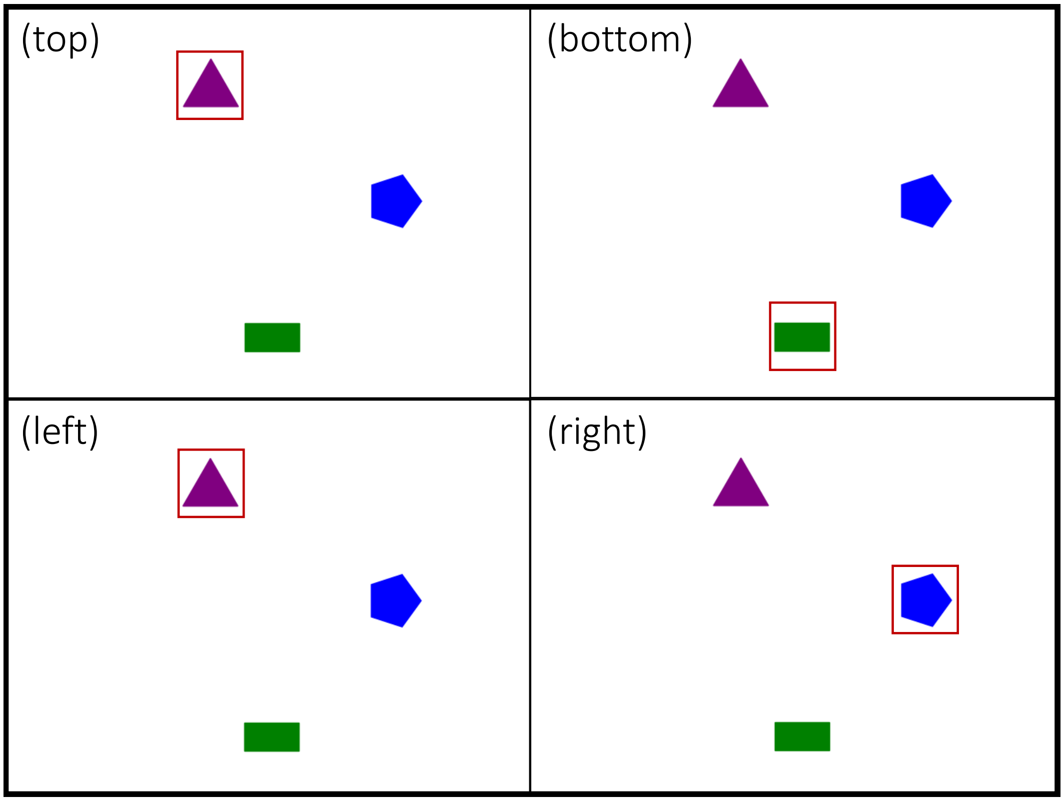

Figure 1: An image from Shapes30k repeated four times to demonstrate the task directions in our benchmark. VLMs are prompted with the question **"Which shape is on __"**; and the direction indicators ***top*, *the bottom*, *the left***, or ***the right*** fill the blank. Answers are marked here with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

2 Related Work

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].

Although CLIP training datasets are compositionally rich, compositional (object relationship) understanding is unnecessary for CLIP models following the present contrastive training strategy [7]. We study VLM performance in a VQA task but use an unconventional dataset to uncover the faults in VLM reasoning that could explain object hallucinations. 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 model perform inadequately without compositional understanding [7].

In this work, we use many VLMs with pretrained CLIP vision encoders, and we expect to observe the bag-of-words phenomenon. 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, we expect the bag-of-word phenomenon to dominate.

Future work must evaluate object hallucination and mitigation techniques in a benchmark that isolates visual reasoning and object category understanding, such as the one in this work. Because of the ubiquity of pre-trained CLIP models in modern VLMs, it is likely that deficiencies in downstream tasks are an offshoot of their difficulties with image composition and the bag-of-words. In the following sections, we investigate this hypothesis using a dataset that isolates positional understanding of VLM models because it lacks visual grounding similar to CLIP's pre-training data, and we demonstrate that VLM models' performance in this task is consistent with previous findings.

3 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*. 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.

We dub the dataset described above 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.

4 Results

Table 1 displays overall accuracy and accuracy per task direction for all six VLMs. Most models score 40% to 60% accuracy. There is significant ($p < 0.01$) directional bias in the accuracy of each model. Table 2 displays overall accuracy again, and accuracy when specific shapes were the answer.

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

HF model/task	***all***	***left***	***right***	***top***	***bottom***
blip2-flan-t5-xl [8]	0.117	0.112	0.144	0.0539	0.160
cogvlm-chat-hf [9]	0.634	0.508	0.623	0.618	0.786
cogvlm2-llama3-chat-19B [10]	0.592	0.505	0.535	0.606	0.722
instructblip-vicuna-7b [11]	0.315	0.278	0.289	0.343	0.350
llava-v1.6-mistral-7b-hf [12]	0.566	0.500	0.571	0.602	0.593
paligemma2-10b-pt-224 [13]	0.410	0.403	0.348	0.405	0.483

Table 2: Accuracy of HF models. Columns with shape names report a calculation of accuracy only for responses where that shape was the answer.

HF model/task	***all***	***circle***	***pentagon***	***rectangle***	***square***	***triangle***
blip2-flan-t5-xl [8]	0.117	0.000	0.000	0.000	0.0924	0.490
cogvlm-chat-hf [9]	0.634	0.926	0.00691	0.653	0.811	0.781
cogvlm2-llama3-chat-19B [10]	0.592	0.825	0.0961	0.563	0.530	0.949
instructblip-vicuna-7b [11]	0.315	0.170	0.000	0.000	0.399	0.999
llava-v1.6-mistral-7b-hf [12]	0.566	0.910	0.000	0.0534	0.885	0.982
paligemma2-10b-pt-224 [13]	0.410	0.639	0.126	0.148	0.168	0.961

5 Discussion

CogVLM [9] scores the greatest overall accuracy at 63.4%. Paradoxically, the newer, related model CogVLM2 [10] lags behind. LLaVA-1.6 [12] performs third best and is the last model whose overall accuracy in the task is greater than 50%. Despite its simplicity, models appeared to 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. Alarmingly, Flan-T5 [8] scores below chance (20%) in for each direction.

In Table 2, 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.

6 Conclusions

The VLMs frequently suffer from object hallucination and fail at spatial reasoning. The common misidentification of pentagons as hexagons underscores a significant limitation in current VLMs. The models surveyed do not perceive spatial relationships between objects accurately and cannot even 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.

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. We hope these findings inspire emphasis on embedding-aware design and evaluation of abstract spatial reasoning performance prior to deployment for grounded tasks.

7 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, etc. The additional grounding in images for those downstream tasks could unexpectedly confound mitigation techniques used to improve upstream performance.

Though we observe the bag-of-words phenomenon, where models are unable to reason about object relationships, we cannot establish a cause for object hallucination in our VLMs. Further, our significance testing for directional bias indicates different directional biases exist for each individual model that should be considered further. Although our work is consistent with previous reports about CLIP vision encoders and encoder-decoder achitectures, 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] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

A Technical Appendices and Supplementary Material

Table 3: p-values from two-way Fisher's exact test on paired task types performed by Salesforce/blip2-flan-t5-xl [8]. 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 4: p-values from two-way Fisher's exact test on paired task types performed by THUDM/cogvlm-chat-hf [9]. 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 5: p-values from two-way Fisher's exact test on paired task types performed by THUDM/cogvlm2-llama3-chat-19B [10]. 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 6: p-values from two-way Fisher's exact test on paired task types performed by Salesforce/instructblip-vicuna-7b [11]. 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 7: p-values from two-way Fisher's exact test on paired task types performed by llava-hf/llava-v1.6-mistral-7b-hf [12]. 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 8: p-values from two-way Fisher's exact test on paired task types performed by google/paligemma2-10b-pt-224 [13]. 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	–

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract summarizes our method and contribution accurately.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our approach in Section 7. Principally, we cannot guarantee that performance improvements on our abstract task will translate necessarily to improvements in downstream tasks, and though our finding are consistent with previous reports on VLM reasoning, we cannot establish a causation with our data.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No theoretical results are provided.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The models used are open source, the code used to benchmark them is provided, and the experimental settings are communicated in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A link to an anonymized repository is provided in the abstract because the code is central to our submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We are evaluating rather than training models in this case, but the dataset specifications are communicated in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We perform meticulous significance testing to demonstrate directional bias in the results. p-values for our tests are recorded in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources including the number of CPUs and GPUs and the model of GPU are listed in Section 3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The only harm encountered was potential copyright abuse. This is mitigated through the use of open source models. No societal impacts are anticipated.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work in this paper focuses on abstract reasoning for VLMs rather than potential deployments. Other than giving an example to motivate our work, societal impacts, positive or negative, do not emerge.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The data poses no risk if released, and we encourage the generation of new datasets using the same or similar parameters using our code. We do not use models that are not already publicly available, so the risks of releasing them have already been taken mitigated by their authors.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The data used in our benchmark is our original data. The models we cite in our benchmarking experiment are open source, and the licenses are available on HF, where our readers are directed to find them.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only asset introduced is our Shapes30k dataset, and we provide documentation, including the number of samples generated and the attirbutes of each sample.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing was and no human subjects were involved in this research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects were involved in this research.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: As a benchmark of encoder-decoder VLM performance, LLMs are central to the methodology because the VLMs use LLMs as their decoder modules. We describe the models used and the method of accessing them.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.