DISSECTING ZERO-SHOT VISUAL REASONING CAPABILITIES IN VISION AND LANGUAGE MODELS

Aishik Nagar
ASUS Intelligent Cloud Services (AICS)
ASUS Global Pte. Ltd.
aishiknagar@gmail.com

Shantanu Jaiswal & Cheston Tan
Centre for Frontier AI Research (CFAR)
Agency for Science, Technology and Research (A*STAR)
jaiswals_shantanu@ihpc.a-star.edu.sg

ABSTRACT

Vision-language models (VLMs) have shown impressive zero- and few-shot performance on real-world visual question answering (VQA) benchmarks, alluding to their capabilities as visual reasoning engines. However, existing works (typically) use benchmarks that conflate "pure" visual reasoning with world knowledge, and also have questions that involve a limited number of reasoning steps. Thus, it remains unclear whether a VLM's apparent visual reasoning performance is due to its world knowledge, or due to actual *visual* reasoning capabilities. To clarify this ambiguity, we systematically benchmark and dissect the zero-shot visual reasoning capabilities of VLMs through synthetic datasets that require minimal world knowledge, and allow for analysis over a broad range of reasoning steps. We specifically focus on evaluating the impact of conveying scene information as either visual embeddings or purely textual scene descriptions to the underlying large language model (LLM) of the VLM. We notably find that the underlying LLMs, when provided textual scene descriptions, consistently perform significantly better compared to being provided visual embeddings. Our work comprehensively identifies limitations of VLMs for compositional visual reasoning, and highlights the important role that LLMs can play in scene understanding and visual reasoning.

1 INTRODUCTION

We perform systematic analyses of zero-shot visual reasoning capabilities in Vision-and-Language Models (VLMs) and their backbone Large-Language Models (LLMs) using synthetic datasets CLEVR (Johnson et al., 2017) and PTR (Hong et al., 2021). These datasets require minimal world knowledge and provide complete scene metadata, which overcomes a major limitation of traditional VQA benchmarks that conflate visual reasoning with factual or world knowledge (Goyal et al., 2017; Marino et al., 2019; Hudson & Manning, 2019). This approach thus enables us to benchmark the pure visual reasoning and scene understanding capabilities uninfluenced by a VLM's world knowledge.

We comprehensively evaluated state-of-the-art VLMs and LLMs over various factors such as the scale of the models, question complexity due to reasoning step lengths, the type of reasoning required by different question categories, and the impact of conveying scene information through visual embeddings versus purely textual descriptions. Our work: **i)** Demonstrates, for the first time, the ability of LLMs to perform visual reasoning and scene understanding when prompted with text-based scene metadata; **ii)** Reveals that LLMs, when provided with ground-truth textual descriptions of scenes, exhibit superior performance in compositional reasoning and scene understanding tasks, compared to Vision Language Models that rely on visual embeddings; **iii)** Underscores the inherent limitations in current VLMs' visual reasoning abilities; **iv)** Emphasizes the substantial, yet under-explored, potential of LLMs in the realm of visual understanding and interpretation.

2 EXPERIMENTS AND FINDINGS

We used three instruction-tuned VLMs: BLIP2-Flan-T5-XL (3B), BLIP2-Flan-T5-XXL (11B) (Li et al., 2023b) and GPT-4V. These were compared to their LLM counterparts: Flan-T5-XL (3B), Flan-T5-XXL (11B) (Chung et al., 2022) and GPT-4. We used 2 synthetic datasets (CLEVR and

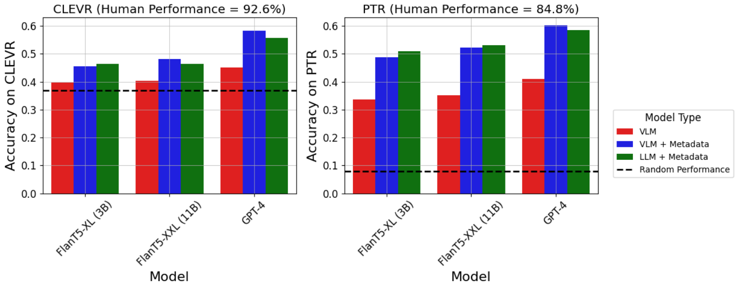


Figure 1: VLM versus VLM+Metadata versus LLM performance on CLEVR and PTR.

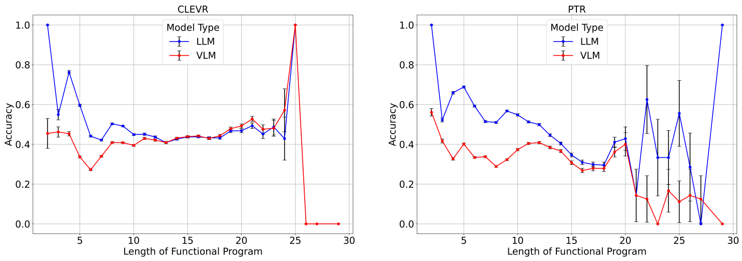


Figure 2: LLM versus VLM performance of Flan-T5-XXL, analyzed by length of functional program.

PTR). As shown in the experimental setup in Figure 3, we compared the "traditional VLM" (i.e. an LLM receiving scene information as visual embeddings from a base vision model) against an LLM simply receiving a completely textual representation of the scene. We found that **LLMs consistently outperform VLMs that utilize the same base LLMs**, as shown in Figure 1. Specifically, BLIP2-Flan-T5 using only its base LLM (Flan-T5) without the visual front-end achieved ∼18% higher accuracy on PTR, while GPT-4 was ∼17% more accurate than GPT-4V on CLEVR. In general, models with scene metadata (LLM and VLM+metadata) obtained significantly better accuracies than VLM-only models, suggesting scene metadata embeddings are potentially more informative and amenable for reasoning than passing only visual embeddings to the VLM decoder. **This is not an a-priori obvious finding, as one might expect that visual embeddings are better representations than text for spatial reasoning, for instance.**

Additionally, to verify that lower VLM performance is not merely due to poor transfer to synthetic images, we also evaluated on the real-world compositional VQA dataset, GQA (Hudson & Manning, 2019). As shown in Table 2 (in the Appendix), we similarly found LLMs with scene metadata performed significantly better than VLMs on GQA. Another key finding is that for questions which can be solved in 2 to 5 "reasoning steps", LLMs showed performance levels which are significantly above chance, suggesting that **LLMs may in fact possess reasonable capabilities as zero-shot visual reasoning engines**. This is an important finding as the LLMs were able to answer questions which require multi-step reasoning and are not necessarily observed during model pretraining. Both LLMs and VLMs generally showed declining performance as the number of "reasoning steps" increases.

The LLM performed better than the VLM in most question categories. Most notably, both LLMs and VLMs have their worst performance on the "analogy" question family of the PTR dataset. This indicates that model reasoning was not complex enough to create analogies, a process which involved multiple stages of reasoning, including identifying the relevant relationships, applying it to a new context, and generating or selecting the correct answer.Another observation is that the VLM when provided with the scene metadata in addition to the image performed ∼2% better than the base LLM only in the case of GPT-4, but not BLIP2. This indicates that the visual frontend for GPT-4 provided additional benefits to the LLM for visual reasoning. The "question family" analysis revealed limitations in current LLMs regarding their ability to create visual representations from textual descriptions. Unlike humans who can easily visualize and understand scenes from text, **LLMs still struggle to generate abstract representations** for complex reasoning and analogical tasks.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

Acknowledgement

This work was supported by A*STAR through a CRF award to Cheston Tan.

References

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-GPT: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. PTR: A benchmark for part-based conceptual, relational, and physical reasoning. *Advances in Neural Information Processing Systems*, 34:17427–17440, 2021.

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5648–5656, 2018.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023b.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204, 2019.

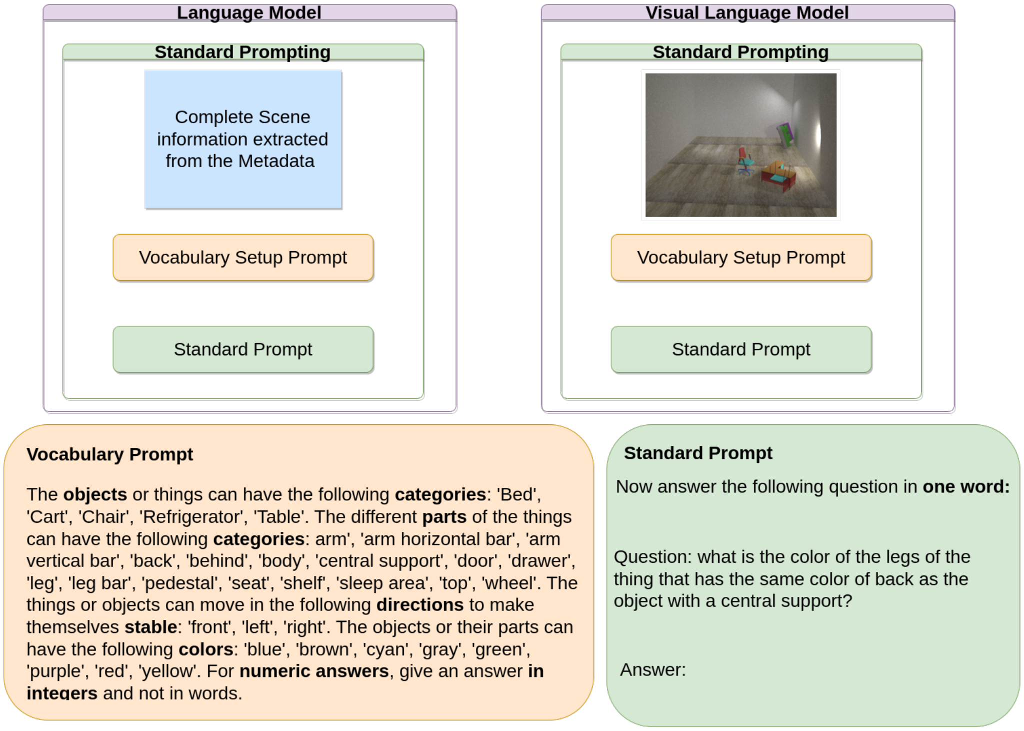
BigScience Workshop: Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, and et al. BLOOM: A 176b-parameter open-access multilingual language model, 2023.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223, 2017.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Appendix

A.1 Results and Analyses



Vocabulary Prompt

The **objects** or things can have the following **categories**: 'Bed', 'Cart', 'Chair', 'Refrigerator', 'Table'. The different **parts** of the things can have the following **categories**: arm', 'arm horizontal bar', 'arm vertical bar', 'back', 'behind', 'body', 'central support', 'door', 'drawer', 'leg', 'leg bar', 'pedestal', 'seat', 'shelf', 'sleep area', 'top', 'wheel'. The things or objects can move in the following **directions** to make themselves **stable**: 'front', 'left', 'right'. The objects or their parts can have the following **colors**: 'blue', 'brown', 'cyan', 'gray', 'green', 'purple', 'red', 'yellow'. For **numeric answers**, give an answer **in integers** and not in words.

Standard Prompt

Now answer the following question in **one word:**

Question: what is the color of the legs of the thing that has the same color of back as the object with a central support?

Answer:

Figure 3: Experimental setup, using the same prompts for both VLMs and their pure-LLM variants.

A.1.1 Comparing LLMs with scene descriptions versus VLMs

LLMs with scene descriptions outperform VLMs: Figure 1 shows the impact of visual grounding using BLIP-2 on the reasoning effectiveness of the models. Pure LLMs generally outperform or have similar performance to their counterpart VLM models across both scales and datasets. A t-test was performed to test if the pure LLMs performed better than VLMs. A p-value of 0.0088 indicates that the difference is statistically significant. This might seem counter-intuitive, as one might expect the VLM to be able to effectively utilize the "visual frontend" provided by the image encoder used in the BLIP-2 setup for querying the relevant aspects of the image. There are 2 possible explanations: 1) There are underlying issues in the VLM architecture which prevent the visual front-end from

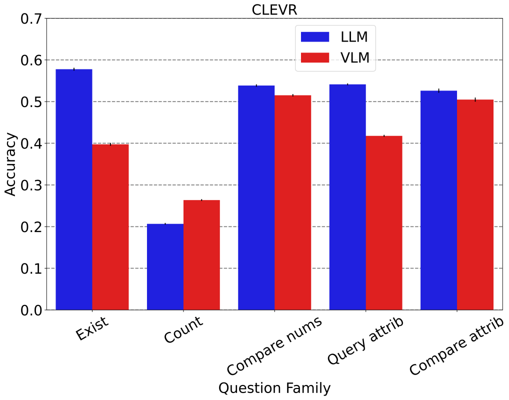


Figure 4: LLM versus VLM model performance of Flan-T5-XXL on CLEVR and PTR, organized by question family.

providing relevant information to the model. 2) The complexity of the tasks is not enough that a visual front-end which queries only the relevant information from the scene can be better than providing the complete, unfiltered information to the reasoning engine: which in this case is the LLM. To guard against data contamination (i.e. LLMs trained on CLEVR or PTR), we ran image-free baselines (Appendix A.8), which performed at chance, indicating no contamination.

LLM advantage for CLEVR versus PTR: The difference in performance between the LLM and the VLM is more pronounced in PTR than CLEVR. For CLEVR, the LLM outperforms the VLM by roughly 6-7%, while for PTR the gap is roughly 17-18%. One possible explanation is that the objects in PTR are more complex, with multiple parts, hence the task for the VLM's visual frontend is more challenging, and more errors and uncertainty are introduced. Providing the ground-truth scene description to the LLM eliminates this challenging visual frontend task. Conversely, the objects in CLEVR are simple geometric objects, hence access to the ground-truth scene description provides less of an advantage to the LLM.

Analysis by number of "reasoning steps": Both CLEVR and PTR provide functional programs which programmatically describe the solution for the reasoning tasks. We used the length of these functional programs **as a proxy** for the number of "reasoning steps" needed. We analyzed the results by number of "reasoning steps" (Fig. 2). For questions requiring relatively fewer "reasoning steps" (up to around 12-17), LLMs generally outperform VLMs. As seen in Fig. 2 (right), for PTR, both LLMs and VLMs generally show declining performance as the number of "reasoning steps" increases, unsurprisingly. However, when it comes to CLEVR (Fig. 2, left), the performance of VLMs seems to be somewhat independent of the number of "reasoning steps". This could be due to the nature of the CLEVR dataset. CLEVR questions are usually abstract and require deep reasoning, regardless of the number of steps. As such, even tasks with fewer steps might be inherently complex in nature, demanding similar levels of abstraction and reasoning as tasks with more steps.

Moreover, because CLEVR consists of geometric shapes rather than recognizable object parts, the VLMs may not gain as much valuable information from the visual encoder for each additional reasoning step. It is important to note that while the program length provides a heuristic for reasoning complexity, it might not always perfectly capture the cognitive complexity for humans. However, it is still worthwhile to study the impact of length of functional programs on performance.

Analysis by question family (CLEVR): The LLM performs better than the VLM in most categories (Fig. 4, left). The "exist" and "query attribute" categories show the most significant difference in performance, with the LLM noticeably better. Interestingly, the multimodal model performs better in the "count" category for Flan-T5 while it is comparable to the LLM in the case of GPT-4. The observed results could potentially be explained by a few factors. For the LLMs, the "exist" and "query attribute" questions are the most straightforward tasks since this information requires a direct lookup from the scene metadata which already contains this information. The VLMs, on the other hand, require identification of the correct object(s) and their attributes even for "exist" and "query

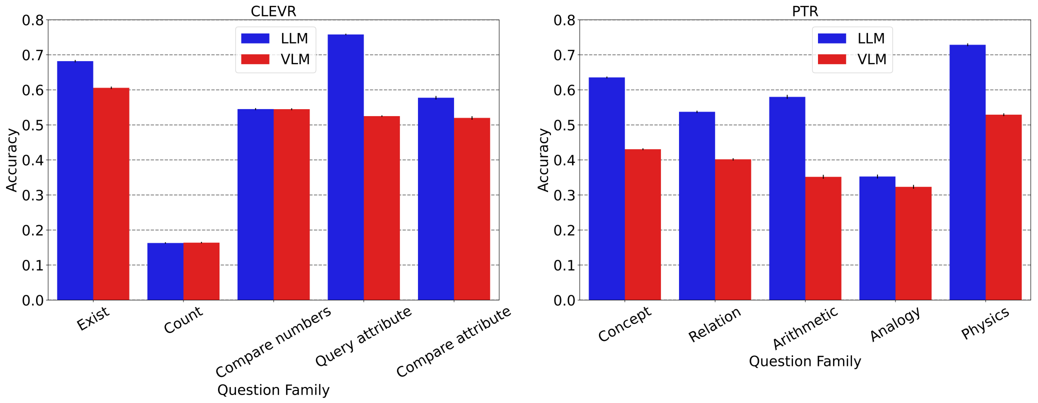


Figure 5: LLM versus VLM model performance of GPT-4 on CLEVR and PTR, organized by question family.

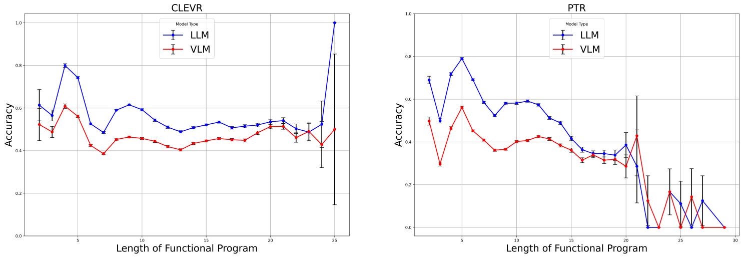


Figure 6: LLM versus VLM performance of GPT-4, analyzed by length of functional program.

attribute" questions. For "counting" questions, on the other hand, it's possible that VLMs, with their ability to process visual data, are more efficient in tasks like counting where visual cues can be valuable. One interesting observation is that GPT-4V is significantly better than Flan-T5 in "exist" category of CLEVR while both GPT-4 LLM and VLM are worse than Flan-T5 in "count" category. This could point to potential advantages in reasoning abilities of the two models for those specific question categories.

Analysis by question family (PTR): The LLM outperforms the VLM across all question families on PTR (Fig. 4, right). The largest performance gap is observed in the "concept" and "relation" categories. "Concept" questions in PTR evaluate a model's capability to understand and reason about basic part-whole relations. Similar to the findings in CLEVR, the question families which require simple "lookups" from the metadata for the LLM have the largest gap in performance. Interestingly, the performance of LLMs on "arithmetic" questions is better than VLMs for this dataset (unlike the "count" questions in CLEVR). This can be attributed to the fact that the level of reasoning required for arithmetic questions is much higher. While such questions in CLEVR were limited to counting objects or comparing numbers, PTR questions require making complex selections of object parts based before performing arithmetic operations.

Visual analogy questions in the PTR dataset require complex reasoning that pose significant challenges for both LLMs and VLMs. This is evident from both the models having their worst performance on the "analogy" question family. This process involves multiple stages of reasoning, including identifying the relevant relationship, applying it to a new context, and generating or selecting the correct answer. The models must not only identify the relationship between A and B, but also accurately project it onto C and D. This complexity could make these tasks particularly challenging for both types of models. Additionally, the geometric and spatial properties involved in analogical reasoning may be difficult for both models.

Note that "count" category from CLEVR and "Analogy" category from PTR remain as the most difficult reasoning categories. We also observe that there are common trends in performance on question categories across the the Blip2-Flan-T5 as well as the GPT-4 models on both the CLEVR and the PTR datasets. This highlights the fact that the question family analysis indeed show a significant bottleneck in the general reasoning abilities of VLMs.

This question family can also provide insights into the abilities of LLMs to make visual representations of textual descriptions. When provided such a text description of a scene, most humans will try to create a visualization to easily identify the parts or objects which are relevant to the problem at hand. This ability to generate abstract representations from descriptions, or use visual inputs to perform complex projections and analogies still seems to be lacking in existing systems.

Drawbacks of current VLM Architecture: VLMs, even those leveraging LLMs, have inherent architectural bottlenecks that may hinder their performance. During inference, they function in two separate phases: 1) visual information querying, where the model's visual frontend extracts scene details based on an initial text query, and 2) text generation, where the LLM uses this extracted information for reasoning and response. This process lacks a feedback loop, preventing the LLM from requesting additional visual information if needed during the generation phase. In contrast, when LLMs receive full scene descriptions in text form, they can access the entire description while generating responses, thereby better retrieving relevant information to answer the question. These drawbacks of VLM architecture are further evidenced by the fact that even when given access to scene metadata, VLMs consistently perform similar to LLMs. This indicates that they are unable to take significant advantage of the additional visual information.

VLM performance on synthetic vs real images. One concern of using VLMs on synthetic datasets is that the vision models are not trained on synthetic data, which could lead to lower performance compared to LLMs. We conducted experiments on the GQA (Hudson & Manning, 2019) dataset using a similar LLM vs VLM comparison, and confirmed that the LLMs also performed better than VLMs on natural images. Full analysis and results are in Appendix A.7.

A.2 Limitations and Future Work

More varied tasks. We used datasets for physical reasoning, due to the availability of comprehensive scene metadata and minimal dependency on world knowledge. Future work can extend to a broader range of visual reasoning tasks, such as abstract data interpretation (Kafle et al., 2018), image-based statement classification (Suhr et al., 2017), etc.

Future work. We plan to extend our study by benchmarking some of the latest instructed-generation capable VLMs such as Otter (Li et al., 2023a), MultiModal-GPT (Gong et al., 2023) and InstructBLIP (Dai et al., 2023) besides recent LLMs such as Chat-GLM (Du et al., 2022), Vicuna (Chiang et al., 2023), OPT (Zhang et al., 2022) and Bloom (Scao et al., 2023) in order to capture trends, bottlenecks and emergent properties for visual reasoning.

A.3 Experiment Code and Reproducibility

All the relevant code and scripts to process the dataset, run all experiments and evaluate the results is available with the supplemental submission . The code uses 2 major libraries for the experiments:

1. The huggingface transformers library for LLM experiments.
2. The Salesforce-LAVIS library for VLM experiments.

Setup instructions have been included in markdown where required.

The 3 major datasets used (CLEVR, PTR and GQA), can be downloaded from these links:

1. CLEVR
2. PTR
3. GQA

The experiment code can be found in the *code* folder provided along with the supplemental submission. The folder structure is provided in the *README.md* in the root folder and separate files are provided to process the dataset as well as run each experiment for the different model families on different datasets.

A.4 Examples of Reasoning step complexities

More details about the the reasoning steps and question families can be found within the papers of the respective datasets:

- CLEVR Dataset [Figure 2 of the paper Johnson et al. (2017)]

- PTR Dataset [Figure 1 of the paper Hong et al. (2021)

A.5 Example of VLM vs LLM

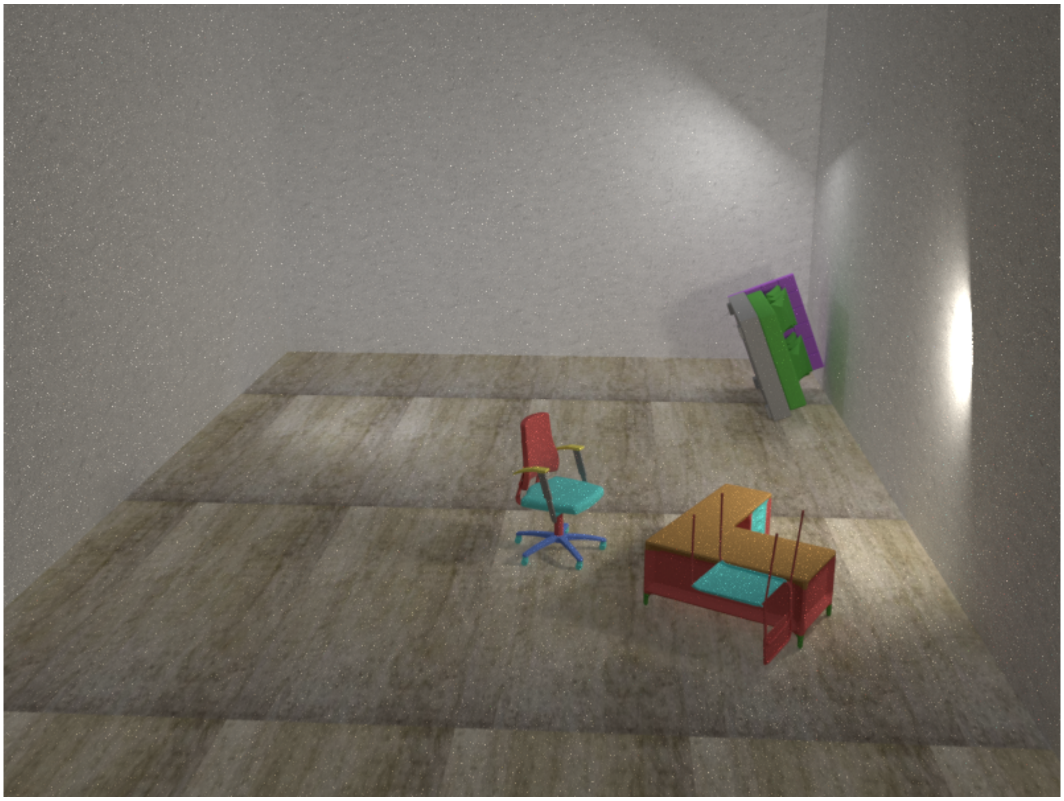


Figure 7: Example of a PTR scene.

Figure 7 shows a scene used in the evaluation of the GPT-4 and GPT-4 Vision models from the PTR dataset. The following was one of the questions asked to the models:"what is the color of the legs of the thing that has the same color of back as the object with a central support?"

Figure 8 shows the reasoning steps required to reach the answer starting from the scene input.The correct answer for this question was "red". The following were the answers provided by the LLM and the VLM:

- GPT-4 (LLM + scene metadata): "red"

- GPT-4 Vision (VLM): "cyan"

As we can see, the LLM arrives at the answer correctly, while the VLM fails to do so. This example concretely shows the complex reasoning required to arrive at the answer, as well as a case where the LLM performs better at reasoning than the VLM.

A.6 Full Experimental Results

The results for all experiments performed are given in the Table 1

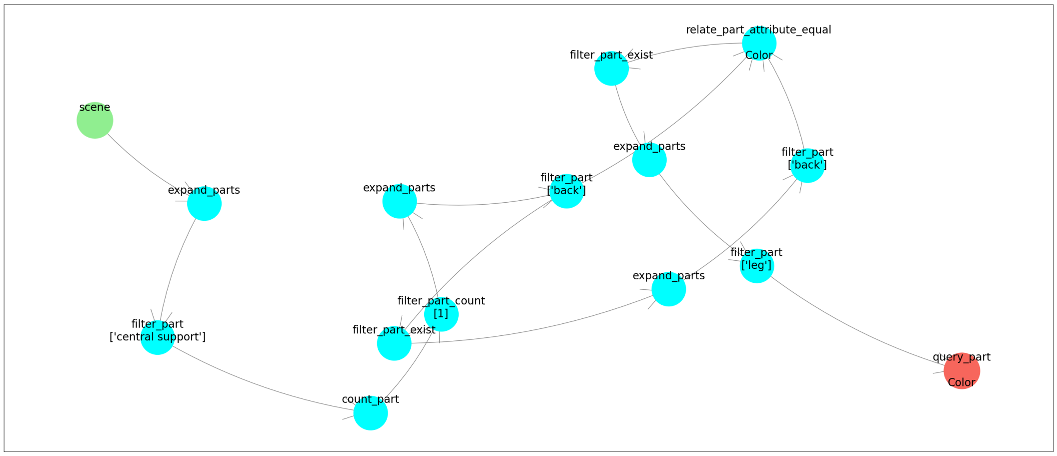


Figure 8: Reasoning steps required at the answer for the question “what is the color of the legs of the thing that has the same color of back as the object with a central support?”. The Green Node signifies the input step while the Red node signifies the output step. The arrows indicate the flow of reasoning.

Table 1: Experiment Results

Model	Scale (Billions of parameters)	Dataset	Type	Accuracy
FLAN T5	3.00	CLEVR	Llm	0.463932
FLAN T5	3.00	CLEVR	Vlm	0.396497
FLAN T5	3.00	CLEVR	Vlm_metadata	0.455474
FLAN T5	11.00	CLEVR	Llm	0.463932
FLAN T5	11.00	CLEVR	Vlm	0.402938
FLAN T5	11.00	CLEVR	Vlm_metadata	0.481456
GPT	0.35	CLEVR	Llm	0.095729
GPT	1.30	CLEVR	Llm	0.175713
GPT	6.70	CLEVR	Llm	0.296915
GPT	175.00	CLEVR	Llm	0.409037
GPT	1800.00	CLEVR	Llm	0.556250
GPT	1800.00	CLEVR	Vlm	0.450970
GPT	1800.00	CLEVR	Vlm_metadata	0.584013
OPT	2.70	CLEVR	Vlm	0.138835
FLAN T5	3.00	PTR	Llm	0.508657
FLAN T5	3.00	PTR	Vlm	0.336524
FLAN T5	3.00	PTR	Vlm_metadata	0.488672
FLAN T5	11.00	PTR	Llm	0.531447
FLAN T5	11.00	PTR	Vlm	0.352028
FLAN T5	11.00	PTR	Vlm_metadata	0.522143
GPT	0.35	PTR	Llm	0.038419
GPT	1.30	PTR	Llm	0.149849
GPT	6.70	PTR	Llm	0.242263
GPT	175.00	PTR	Llm	0.461586
GPT	1800.00	PTR	Llm	0.586389
GPT	1800.00	PTR	Vlm	0.409767
GPT	1800.00	PTR	Vlm_metadata	0.601770
OPT	2.70	PTR	Vlm	0.276101

A.7 GQA Experiments

The GQA dataset was used to test the experimental setup on a dataset which uses natural images instead of synthetically generated images. This was done in order to check the fairness of the VLM vs LLM comparision on the original datasets. The rationale behind this was that the Visual encoders in the VLMs were not trained on synthetic images, which affect the performance on the datasets selected in the original paper. The GQA dataset was as it provided access to comprehensive scene metadata as well as functional programs to arrive at the answer, similar to the (Johnson et al., 2017) and (Hong et al., 2021) datasets used in the main experiments.

Analysis of the results. We can see that the LLM outperforms the VLM on the dataset, as well as over the length of functional programs and question families. This result is consistent with the findings of the main paper. It is important to note that there are not many questions with a large length of the functional programs in the dataset, the scene metadata covers all the important relationships and informations in a more verbose manner and the answers seem to be generally simpler to answer than the synthetic datasets, which could explain a relatively larger gap in the LLM vs VLM performance.

Table 2: Experiment Results on Sampled GQA Dataset

Model	Dataset	Accuracy
Flan-T5 XXL	Sampled GQA Dataset	78.72
Blip-2 Flan-T5 XXL	Sampled GQA Dataset	56.81

A.8 Image-Free Baseline and Random Chance

The following tables provide the probabilities of getting an answer correct by randomly picking an option from the question vocabulary for CLEVR 3 as well as for the PTR 4 . The image free baselines for GPT-4 on CLEVR and PTR were **36.85%** and **10.16%** respectively. Image free baseline results indicate that the model performance in the absence of scene metadata is basically random chance.

Table 3: Random Chance and Total Questions for CLEVR

Category	Random Chance (%)	Total Questions
exist	50.00	20196
colors	12.50	13404
material	50.00	30545
compare attribute	50.00	35422
shape	33.33	13544
size	50.00	10094
count	10.00	13273
compare numbers	50.00	13513
Overall random chance		36.86

Table 4: Random Chance and Total Questions for PTR

Category	Random Chance (%)	Total Questions
concept	2.63	38972
relation	4.35	22905
physics	50.00	7413
analogy	5.26	7472
arithmetic	8.33	14958
Overall random chance		8.03

The code required to run the experiments as well as the analysis has been provided with the supplemental submission under the “image_free” folder.