

CIS 7000 Final Project Report

Yankai Li

University of Pennsylvania
liyankai@seas.upenn.edu

Hong Lyu

University of Pennsylvania
hlyu@seas.upenn.edu

Renyi Qu

University of Pennsylvania
requ@seas.upenn.edu

Abstract

This project attempts to tackle the issue of object relation inference in VQA (Visual Question-Answering) using VLMs (Vision-Language Models) and LLMs (Large Language Models). Inspired by [Hu et al. \(2022\)](#), we utilize the chain-of-thought reasoning capability of contemporary LLMs to interact with the VLM with clarifying questions. The VLM is responsible for providing the LLM with sufficient text information about the image and answering whatever clarifying question the LLM asks. The LLM is responsible for understanding the text information provided by the VLM, acquiring more useful information through clarifying questions, and ultimately answering the input question. We experimented on the GQA dataset, and our results show that adding the LLM did not necessarily improve the performance of the state-of-the-art VLMs on object relation inference. While GPT-3.5 was able to understand and utilize concepts from text information properly, the whole system was very limited by the capabilities of the VQA module and many other limitations. We suggest further research to tackle these issues.

This is particularly important for the topic of concepts in deep learning as it shows that current VQA systems do not possess the ability to understand the concept of physical relations whereas humans are able to do this with ease ([Yi et al., 2018](#)).

Recently, there has been remarkable progress in the development of LLMs, namely GPT-3 ([Brown et al., 2020](#)) and its dialogue version ChatGPT (or GPT-3.5). They have demonstrated excellent performance in various natural language processing tasks due to their exceptional abilities in extracting, comprehending, and utilizing various concepts from the input context to generate incredibly useful texts. Therefore, many researchers and practitioners have attempted to incorporate LLMs into the VQA systems ([Tsimpoukelli et al., 2021](#); [Gao et al., 2022](#); [Yang et al., 2022b](#); [Hu et al., 2022](#); [Zhang et al., 2023](#)). They managed to achieve great performance in some question types but not necessarily all of them. In particular, they still suffered from relational questions due to the fatal issue that current image captioning or description systems do not encode positional information.

1 Introduction

VQA (Visual Question-Answering) is an essential multimodal task where a VLM (Vision-Language Model) attempts to output an answer given an input image and an input question. Specifically, it is an essential task for the topic of concepts in deep learning because the VLM needs to accurately understand the concepts of The contemporary state-of-the-art VQA (Visual Question-Answering) models are able to achieve reasonable results for many question types such as verification and some basic query questions. However, they do not perform well on object relation inference, where the model is asked to infer the relation between two or multiple objects in the image ([Sharma and Jalal, 2021](#)).

Inspired by this fatal issue and the theme of this class, we investigate the possibility of using LLM as the concept management system to tackle relational VQA. The VLM functions as an oracle that provides text information from the image to the LLM whenever queried, and the LLM functions as a student who is trying to solve a given question with the text information provided by the VLM. Section 3 will describe our method in detail. We experimented our system together with its variations against the baseline VLMs on the GQA dataset, and our results suggested that the addition of LLM does not necessarily improve the VLM's performance on object relation inference. We list our limitations in Section 5 and suggest further research to tackle these issues.

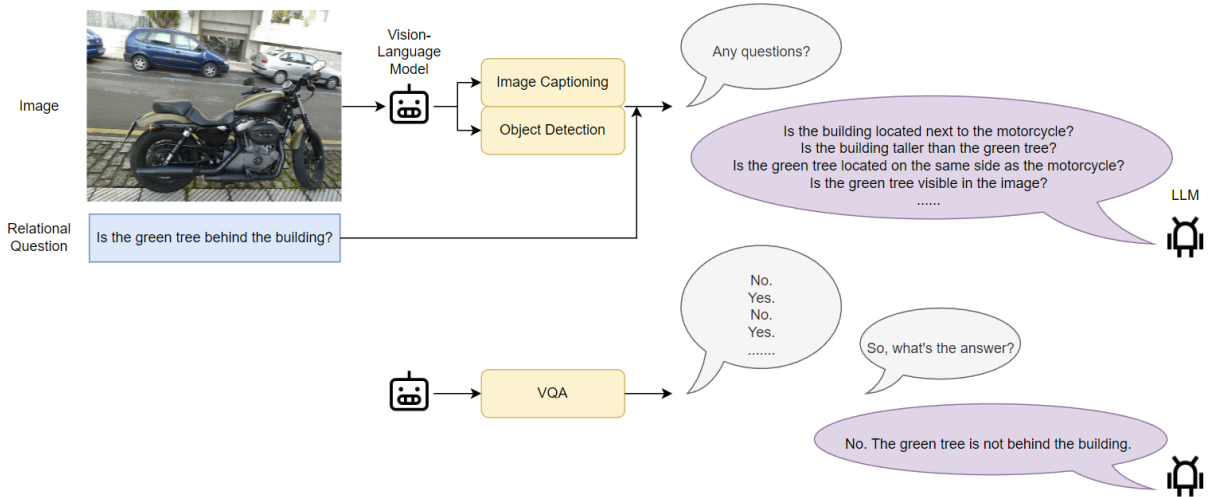


Figure 1: Our method. There are 2 major interactions: the initial stage and the question-answer exchange. At the initial stage, the VLM extracts text information from the image and gives them to the LLM. At the exchange stage, the LLM asks clarifying questions and the VLM answers them. Eventually, the LLM attempts to answer the input relational question.

2 Related Work

In previous research, [Kim et al. \(2021\)](#) first attempted to incorporate transformers into a vision-language setting. [Tsimpoukelli et al. \(2021\)](#) first attempted to incorporate LLMs into the multimodal setting by transforming image embeddings into the same shape as text token embeddings and placing them into the input prompt as zero-shot or few-shot prompting for LLMs. For VQA, their frozen method can convert the image and question into the same token embedding shape, concatenate them sequentially, and prompt the LLM to provide a text answer. However, their method did not work too well because the transition from image to token embedding failed to capture useful concepts from the image and resulted in a huge loss of information. [Gao et al. \(2022\)](#) demonstrated that it is possible for a model to answer VQA problems using only language information, such as captions, label lists, and OCR of the input image, without access to the image itself. This suggests that high accuracy can be achieved without utilizing image embeddings, but they did not investigate the explicit use of LLMs to address this problem either.

Subsequent research by [Yang et al. \(2022b\)](#) explored the explicit use of LLMs for VQA by extracting text information from the image and incorporating such text into the prompt for GPT-3 to generate the final answer directly. Their findings provided evidence that LLMs can effectively answer VQA problems based on language information alone. A

more recent work by [Zhang et al. \(2023\)](#) further suggested the Chain-of-Thought reasoning capabilities of LLMs in the multimodal setting. By simply providing the image, question, and answer choices, LLMs are able to generate their own rationale for logical reasoning and then choose an answer from the given choices. Furthermore, [Yang et al. \(2022a\)](#) provided evidence that LLMs can extract specific concepts from images given proper prompts. Our method is closely related to these recent works and focuses specifically on object relation inference, which none of the previous works specialized in.

3 Method

Figure 1 is a diagram of our method, named "Chain-of-Clarifying-Questions". Our method consists of a total of 4 interactions between the selected VLM and LLM. First, given an input image and an input relational question, the VLM extracts core information from the image through image captioning and object detection and provides them to the LLM in the form of texts. This step is inspired by [Yang et al. \(2022b\)](#) and [Gao et al. \(2022\)](#) as image captioning and object detection are the best ways to convey information about an image in the form of language at the moment. Second, the LLM asks clarifying questions to the VLM based on the current information it knows about the image to further help it answer the input question. Third, the VLM provides answers to the clarifying questions. Lastly, the LLM attempts to answer the input question.

Our approach is designed to address relational questions, which require the LLM to identify and understand concepts in the image in order to answer the question accurately. Motivated by the results from Zhang et al. (2023) and Yang et al. (2022a), we hope that the LLM understands what objects are of the highest importance and what relation is being asked, and that the LLM can ask clarifying questions that are beneficial to understanding the core concepts involved in the input relational question.

4 Experiment

We use the GQA (Hudson and Manning, 2019) dataset, consisting of symbolically generated questions about real-world images. These questions are intrinsically logical and thus good candidates for investigating the capabilities of extracting, understanding, and utilizing concepts of the LLM in the multimodal setting. In particular, we select 100 samples from the "relational_testdev.json" file as our test set. We did not use the training set because our VLMs have likely seen the data in the training set in the past, which might result in biases in the final result.

We use 2 state-of-the-art VLMs: BLIP-2 (Li et al., 2023) and MDETR (Kamath et al., 2021). BLIP-2 is the state-of-the-art model in the overall vision-language setting. It is capable of any task that requires generating some text related to the image given any prompt in a zero-shot or few-shot manner. It has been tested on GQA with relatively the best performance, but objectively speaking, the largest version of BLIP-2 only has an accuracy of 44.7% on the test-dev set which our subset belongs to. On the other hand, MDETR is specifically tailored for highly accurate object detection and relational VQA tasks. It was trained on GQA and achieved competitive performance on the test-dev set. However, it is unknown whether it could generalize to other VQA dataset or real-world VQA in general.

We use GPT-3.5 as our LLM, and we use the LangChain package (Chase, 2022) to assist our construction of clarifying question chain. Please refer to Appendix for the technical detail.

4.1 Extensions

In addition to our primary method, we explored several approaches that are built upon the foundation of our primary method.

4.1.1 A "Good" clarifying question

One issue that emerged during the testing process was the randomness in the generated texts of GPT-3.5. Running the same script on the same sample leads to different formats and contents of generated clarifying questions from GPT-3.5. Though we were able to control the format through rigorous prompt engineering, we were not able to control the content unless tuning the temperature parameter which we forgot during the experiment.

Nevertheless, it is hard to define what a "good" clarifying question is since the definition of "good" is very situational. Generally speaking, a good question should guide the LLM to gain more information from the VQA system and the image. There are 3 aspects of a "good" question that we looked into, but they are not feasible given the time and resource constraints of the project.

First, a good question should always result in a high and positive information gain. This is straightforward as the more information the LLM has, the more likely it will answer the input question correctly. However, there is a lack of clear measurement of such information gain. We may vaguely determine whether or not a clarifying question leads to a high information gain by looking at it, but we are not going to stop the automatic generation chain at every single step for every single clarifying question given the time scope of this project. It is reasonable to outsource human evaluations for this part, but we could not afford them. It is also possible to build some BERT-based model to measure the distance between the embeddings of the current information that the LLM has and the new information that is gained from the clarifying question, but a far distance does not necessarily contribute to a positive information gain as the clarifying question could be completely irrelevant to the input question at all and still got evaluated as a positive gain based on the distance measure.

Similar to the first aspect, the second aspect is that a good clarifying question should be somewhat relevant to the original question. However, it shall not be too relevant or there will be no information gain at all. In particular, it should focus on the objects involved in the queried relation. We can use some existing semantic similarity metrics to measure the similarity between the clarifying and input questions, but we do not know the particular threshold for the similarity, which would again require human annotations that we could not afford.

Nevertheless, we can explicitly ask the LLM to extract the core entities from the input question and to ask clarifying questions centered around these entities. Meanwhile, the LLM might already be equipped with implicit capabilities to ask informative clarifying questions related to the core entities. For this reason, we extend our major method with an additional step of explicit concept extraction as described previously.

Lastly, a good clarifying question should be answerable by the current VQA systems. If the current VQA systems provide a wrong answer to the clarifying question, this directly leads to a negative information gain. This is the trickiest part because the contemporary VQA systems always provide you with an answer regardless of whether they can answer the question correctly or not. As shown by the result section of Li et al. (2023), the best VQA model BLIP-2 provides wrong answers in the majority of the test-dev cases in GQA, and we simply have no way of knowing them because we do not have the ground-truth answers for the LLM-generated clarifying questions. Human evaluation is again necessary for this step.

4.1.2 Incorporating VLM Original Answer

This approach involves utilizing the original answer generated by VLMs as the basis for generating a corrective answer. BLIP2 is used to generate captions for the input image. We then feed the generated caption and input question to ChatGPT, which generates five clarifying questions to aid in answering the input question. The answers to these clarifying questions are obtained using VLMs and serve as additional information.

This method involves prompting ChatGPT to refrain from altering the original answer generated by VLMs unless it is certain about the answer with the additional information provided. This approach effectively allows ChatGPT to serve as a corrector to VLMs, correcting any obvious mistakes while preserving the accurate original answers from VLMs that contribute to the final accuracy.

Overall, this approach is designed to improve the accuracy of VQA by leveraging the strengths of VLMs and ChatGPT while correcting any errors and preserving the original answers that contribute to the overall accuracy of the system.

4.2 Results

Table 1 shows the accuracy comparison of different methods for BLIP2. The baseline accuracy is

Method	Accuracy
BLIP-2 (Baseline 1)	0.54
BLIP-2 + GPT-3.5	
w/o caption & w/o concept extraction	0.59
with caption & w/o concept extraction	0.55
w/o caption & with concept extraction	0.21
with caption & with concept extraction	0.51
w/o caption & with BLIP2 original answer	0.64
with caption & with BLIP2 original answer	0.55
...	...

Table 1: Accuracy comparison of different methods for BLIP2

Method	Accuracy
MDETR (Baseline 2)	0.77
MDETR + GPT-3.5	
w/o caption & with MDETR Original Answer	0.67
with caption & with MDETR Original Answer	0.73

Table 2: Accuracy comparison of different methods for MDETR

0.54. When incorporating GPT-3.5, the accuracy improves to 0.59 without captioning and concept extraction. With captioning but without concept extraction, the accuracy is 0.55. However, when concept extraction is included, the accuracy drops to 0.21. When incorporating the original answer generated by BLIP2 into LLMs, the accuracy improves to 0.64 without caption and 0.55 with caption. The findings from our experiments indicate that utilizing the original answers generated by VQA systems can enhance the accuracy of LLMs. However, providing explicit instructions for concept extraction may not always result in improved performance. In fact, LLMs, such as ChatGPT, are already quite adept at comprehending the underlying concepts present in natural language questions. They are proficient in identifying entities and relationships that require further clarification, even without explicit guidance. This suggests that LLMs’ inherent abilities to grasp complex concepts and discern the necessary information might be sufficient for many tasks, and additional guidance may not always lead to better outcomes.

For MDETR, which has been pre-trained on the GQA dataset, the baseline performance is already outstanding. As demonstrated in Table 2, our experiments indicate that incorporating ChatGPT does

not lead to any improvement in MDETR’s performance. This outcome primarily arises from the fact that MDETR is trained on a fixed answering dictionary. Consequently, when MDETR is presented with clarifying questions generated by LLMs, its performance suffers significantly if the expected answers fall outside the dictionary. This issue results in MDETR providing a substantial amount of incorrect information to ChatGPT, ultimately causing a decline in overall performance.

5 Limitations

5.1 Limitations by LLMs

The limitations of the use of chatGPT are mainly 2-folds. The first significant limitation is the difficulty in prompting chatGPT to concentrate on specific concepts without degrading the pipeline’s performance. When attempting to guide the model toward focusing on entities or relations, the model might, contrary to our expectations, deviate from the main questions and start to ask tedious clarifying questions like "What kind of color is blue in this context?".

Another concern with chatGPT is the potential for problematic answers that might influence our metric calculation. For instance, in the early stage of our experiments, chatGPT might reply "With the information given, I am not able to give an answer to the question", which might be counted as correct if the ground truth answer is "No". Another kind of problematic answer is that chatGPT sometimes provides answers that are close to but not an exact match of ground truth. For instance, chatGPT might answer "shuttle" or "van" when the expected ground truth is "bus". This discrepancy often sources from the answers to clarifying questions by VLMs.

5.2 Limitations by VQA models

As for the challenges posed by the Visual Question Answering models used in our approach. A significant limitation is the poor performance of VQA models when answering clarifying questions from ChatGPT. For instance, MDETR, which was trained with a fixed answering dictionary, experiences a substantial performance degradation if the expected answers to the clarifying questions fall outside the dictionary. Therefore the introduction of clarifying questions by LLM to MDETR worsens the model’s performance on its own.

On the other hand, BLIP2 is designed and expected to answer open questions, but frequently provides incorrect or uninformative answers to questions. A notable example is an image of a woman holding an umbrella for a reading man. When asked, "Is there any visible clothing on the man?", BLIP2 incorrectly responded with "No," even though the man was properly dressed. Furthermore, BLIP2 falsely claimed that the man was holding the umbrella instead of the woman. Examples of uninformative answers include "The kitchen table is a table" and "The sphere is made of sphere". These limitations are the main performance bottleneck for our pipeline, and highlight the need for further improvements in VQA models to better handle the complexity of real-world scenarios and provide accurate responses to LLM-generated clarifying questions. Specific examples would be provided in the Appendix section.

5.3 Limitations in evaluating the quality of clarifying questions

Another limitation in our approach is the lack of resources to quantitatively measure the quality of the clarifying questions generated by the LLMs. As mentioned in Section 4.1.1, an effective clarifying question should be closely related to the original question, informative enough to acquire relevant knowledge for answering the GQA question, and, crucially, answerable by the VQA models under the current settings.

However, there is no automatic metric available for assessing these qualities when it comes to the generated questions. Due to resource constraints, we were unable to manually evaluate each clarifying QA pair to measure the VQA models’ performance and help LLMs understand their shortcomings. This limitation presents a challenge in optimizing the performance of our pipeline and accurately evaluating the benefits of using LLM-generated clarifying questions for VQA tasks. Further research is needed to develop metrics or evaluation methods that can accurately assess the quality of the clarifying questions generated by LLMs in this context.

5.4 Limitations by resources

In this section, we discuss the constraints posed by limited computational resources and the OpenAI API quota. Due to these restrictions, we were only able to report the performance of our pipeline

over the first 100 relational questions in the GQA development set. This sample size might be considered small and potentially not representative of the broader dataset.

Nonetheless, in a few additional tests that expanded the scope to include up to 200 questions, we observed that the accuracy did not deviate significantly from the initial 100-question evaluation. While these findings suggest that our approach might generalize to larger samples, it is important to acknowledge the limitations imposed by resource constraints and the possibility of biased results due to such a small sample size. Further research with more extensive testing is necessary to draw more definitive conclusions about the efficacy of our approach in enhancing VQA tasks.

6 Conclusion

This project examines the potential benefits of incorporating LLMs into VLMs for object relation inference in VQA. Our findings suggest that the addition of LLMs may not necessarily improve the performance of VLMs, which remain limited by their own performance. Furthermore, our study indicates that LLMs are effective in dealing with concepts and that explicit extraction and extra focus on concepts does not necessarily improve their ability to understand entity relations and solve VQA tasks. Additionally, our project proposes a method for incorporating the original answer generated by VLMs into LLMs, which can increase the accuracy of VLMs and can be easily adapted to any VLMs to further improve their accuracy.

Acknowledgements

We would like to express our sincere gratitude to all those who have supported us in the completion of this research project. In particular, we would like to thank Professor Lyle Ungar and Professor Konrad Kording for their valuable suggestions and insightful feedback, which greatly contributed to the quality of this project. We would like to also acknowledge the paper choices from all members in this class on the concept understanding aspect of contemporary multimodal methods, which ultimately led to the formation of this project.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harrison Chase. 2022. [LangChain](#).

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. A thousand words are worth more than a picture: Natural language-centric outside-knowledge visual question answering. *arXiv preprint arXiv:2201.05299*.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Drew A Hudson and Christopher D Manning. 2019. 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*, pages 6700–6709.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

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

Himanshu Sharma and Anand Singh Jalal. 2021. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*, 116:104327.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2022a. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. *arXiv preprint arXiv:2211.11158*.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022b. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

A Appendix 1: Prompt examples

In this appendix, we present a selection of prompts used in our pipeline to illustrate the process of utilizing Large Language Models (LLMs) to improve relational Visual Question Answering (VQA) tasks. The prompts were designed to encourage LLMs, specifically ChatGPT, to ask clarifying questions to VQA models given GQA questions, and then use the VQA responses to generate a final answer.

A.1 Example 1

Context:

Given an image. You are allowed to ask an Image qa model (BLIP2) at most 5 clarifying questions to answer the final question (Are there any vans in front of the flag?).

Ask informative questions to maximize your knowledge about relations in the image so that you can answer the final question. Response should be a list of questions and nothing else:

The questions you raised and the corresponding answers are:

'Question: Are there any vehicles of any kind in front of the flag? Answer: Yes, there are a few vans and a few cars', 'Question: Is the flag in the foreground or background of the image? Answer: Background', 'Question: Is there any prominent object that could obstruct the view of the flag in the foreground? Answer: No', 'Question: Are there any groups of people standing in front of the flag? Answer: Yes, there are people standing in front of the flag', 'Question: Is the area in front of the flag clear or cluttered with objects? Answer: Cluttered' Ignore the qa pairs that you think do not make sense. Now explicitly answer the original question (Are there any vans in front of the flag?) in one word or phrase if you can answer it with the information provided, answer NA otherwise.

Response:

Yes, there are vans in front of the flag.

Ground truth:

yes

Result:

True

A.2 Example 2

We also provide the following screenshots as an example of comparing the approaches without or with the original answer by BLIP2.

B Appendix 2: VLM failure examples

In this appendix, we present a selection of screenshot examples where the Visual Question Answering (VQA) model BLIP2 fails to answer clarifying questions posed by ChatGPT or provides uninformative responses. These examples highlight the limitations of the VQA models used in our approach and the need for further improvements.

You have the following information:

1. What is the room made of?: Office
2. Is there any other furniture in the room?: No
3. How are the shelves attached to the wall?: With screws
4. What is the size of the shelves?: Small
5. What is the room's interior design style?: Modern

Your task is to answer question: What are the shelves mounted on?

You are given the answer to the question: They are shelves.

But you are not sure if the answer is correct and you may want to adjust the answer based on additional information above.

The additional information may also be wrong, please evaluate them all together and give a final answer.

If you are not sure about the final answer, just give the given answer as the final answer.

Now explicitly answer the original question (What are the shelves mounted on?) in one word or phrase.

predicted: Wall.

ground truth: wall

True

Final Answer Given by ChatGPT

Original Question

Original Answer From BLIP2

(a) Enhanced Accuracy: ChatGPT's Capability to Update BLIP2 Answers with Supplementary Information

You have the following information:

1. What is the van's position in relation to the flag?: The van is positioned in relation to the flag
2. Is the flag behind or in front of the van?: In front of the van
3. Can the flag be seen through the van's windows?: Yes, it can
4. Is the flag located at a higher or lower elevation than the van?: The flag is located at a higher elevation than the van
5. What is the distance between the van and the flag?: The distance between the van and the flag is the distance between the van and the flag

Your task is to answer question: Is the van in front of a flag?

You are given the answer to the question: Yes, it is a flag van.

But you are not sure if the answer is correct and you may want to adjust the answer based on additional information above.

The additional information may also be wrong, please evaluate them all together and give a final answer.

If you are not sure about the final answer, just give the given answer as the final answer.

Now explicitly answer the original question (Is the van in front of a flag?) in one word or phrase.

predicted: In front.

ground truth: yes

False

Final Answer Given by ChatGPT

Original Question

Original Answer From BLIP2

(b) Unrecognized Accuracy: ChatGPT's Correct Answers Marked Incorrect due to Evaluation Criteria

Figure 2: Example for the approach: w/o caption & with BLIP2 original answer

