

Attention Is All That Answer-Choices Need: Enhancing Visual Reasoning Using Pre-trained Language Models

Imene Ben Ammar, Daniel Richter, Anton Segeler and Omadbek Meliev

Abstract

This project introduces an innovative approach that combines Large Language Models (LLMs) and Visual Language Models (VLMs) to solve visual reasoning problems. By leveraging the strengths of both model types, our method focuses on improving the understanding and interpretation of complex visual data, intending to achieve superior performance in visual reasoning tasks without the need for any fine-tuning techniques. We employ a novel architecture that allows dynamic interaction between the LLMs and VLMs, facilitating detailed analysis and insight generation from visual input. Details of the development and evaluation of our approach are described in this paper, demonstrating remarkable improvements over existing methods. We present a comprehensive analysis, supported by experiments on the A-OKVQA dataset, showing that the architecture of our proposed approach achieves higher accuracy and a deeper understanding of visual reasoning compared to state-of-the-art methods such as ViCor, which is outperformed by 2%. The code is available at <https://git.tu-berlin.de/data-science-project-sose24/visual-reasoning>.

1. Introduction

In recent years, the appearance of Large Language Models (LLMs) and Visual Language Models (VLMs) has revolutionized both computer vision and natural language processing fields, as these are continuously making it easier for machines to process and interact with visual and textual data. Hence, various applications have benefited from this development including the visual reasoning task.

Visual reasoning is a crucial task involving using and enhancing the models' capabilities to comprehend complex visual scenes, extract critical visual information, and leverage commonsense reasoning to draw insightful conclusions beyond what is explicitly depicted in images [1]. This enables the development of systems that can interact with the world and make decisions in a similar manner that humans do. Consequently, unlike simple question-answering, visual reasoning demands multimodal perception and a deep understanding of the world through both visual and commonsense knowledge [2].

Although recent research has shown that VLMs and LLMs are capable of handling visual reasoning tasks independently [2], various integrated pipelines have emerged to achieve higher performance. These pipelines combine VLMs, adept at processing and generating text from visual inputs, with LLMs, which excel in generating coherent and contextually relevant textual outputs. Even though these pipelines were proven to be effective (more details in the following section), the gap remains in fully leveraging the combined strengths of these models to

achieve even higher performance.

In this work, this gap is addressed by introducing a novel approach that integrates LLMs and VLMs more effectively and without considering any fine-tuning techniques, aiming to achieve a more profound understanding of visual data and therefore lead to more precise and insightful reasoning outcomes. Hence, the research question guiding this investigation can be formulated as follows:

How can the integration of LLMs and VLMs, without fine-tuning, be further optimized to achieve superior performance in visual reasoning tasks?

2. Related Work

Several approaches that combine both pre-trained LLMs and VLMs to perform visual reasoning tasks have been introduced in the past few years and have been able to achieve remarkable performance. Three of these state-of-the-art methods are presented in the following subsections.

2.1. ViCor

Zhou et al. presents with the ViCor framework an innovative approach to leverage the strengths of both VLMs and LLMs. They distinguish between two types of visual commonsense reasoning (VCR) problems: Visual Commonsense Understanding (VCU), which requires understanding the literal visual content, and Visual Commonsense Inference, which requires drawing conclusions beyond the content of the image. The approach uses pre-trained LLMs to act as problem classifiers to find the appropriate problem category before eliciting a response directly from the VLM or to actively create instructions

✉ imene.ben.ammar@tu-berlin.de (I. B. Ammar);
daniel.richter@campus.tu-berlin.de (D. Richter);
segeler@campus.tu-berlin.de (A. Segeler);
meliev@campus.tu-berlin.de (O. Meliev)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

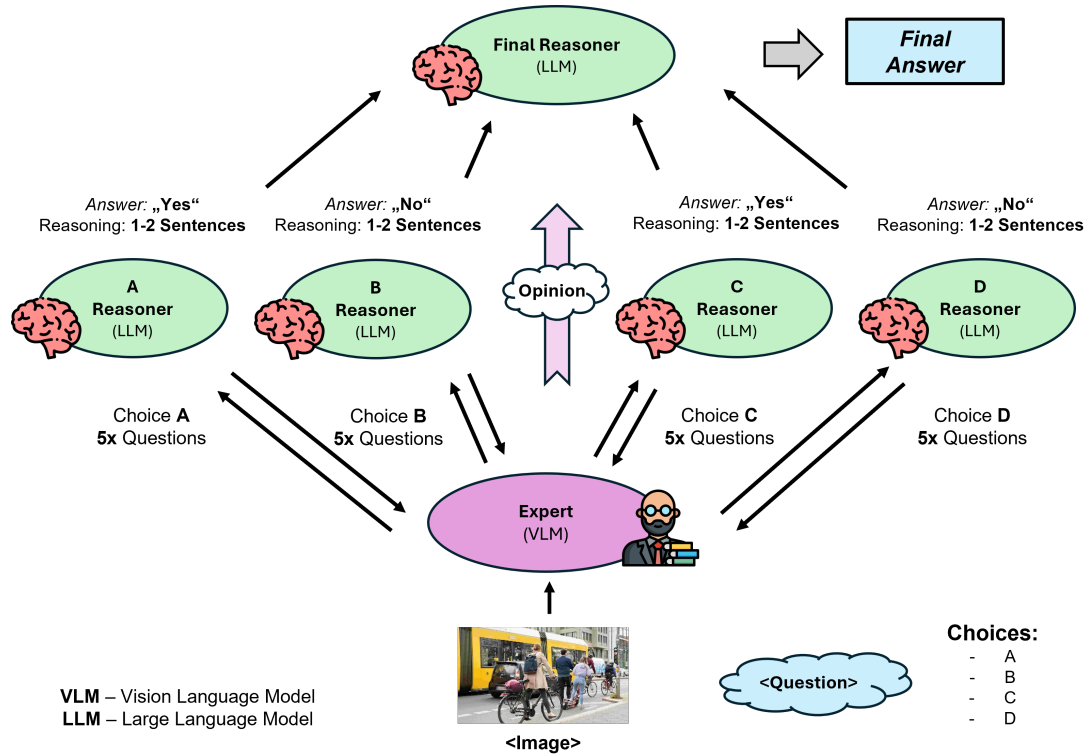


Figure 1: Overview of our approach

for the VLMs on what to focus on and collect relevant visual information to support the common sense inferences. The evaluation of this approach was done on the VCR [3] and A-OKVQA [4] data set. They surpassed all other approaches that don't need in-domain fine-tuning and achieved an accuracy of 75.6 percent on the A-OKVQA data set.

2.2. Cola

In another research, Chen et al. use the advantage of multiple VLMs with different preferred patterns in describing the visual context and predicting plausible answers in natural language. An LLM is used to coordinate and integrate their answers and efficiently and effectively leverage their respective strengths. Experiments show that this method exhibits competitive performance with finetuning, in zero-shoot settings and in few-shoot settings, compared to other state-of-the-art models in these respective settings. The model uses a simple approach consisting of a single prompt combining the captions and answers of two VLMs.

We investigated the idea of including a more complex

approach that uses multiple VLMs and multiple iterations of sub-questions, similar to the paper by You et al.. Having limitations that made it difficult to deploy multiple VLMs, let us come to the conclusion that this is out of the scope of this project. This still is a promising idea that might be worth looking into in a different context.

2.3. IdealGPT

In another significant contribution, You et al. introduce the IdealGPT framework [6], which employs a divide-and-conquer strategy. This framework consists of three main components working together: 1) The Questioner, an LLM, generates various sub-questions to gather additional context based on the given main question and image caption describing the image. 2) The Answerer, a VLM, then responds to these sub-questions using the provided image. 3) Finally, the Reasoner, another LLM, processes the sub-questions and answers, and reasons to arrive at the final answer to the main question. These components interact iteratively to provide sufficient information until the Reasoner is confident about the final answer. The conducted experiments showed the efficacy

of this approach in the zero-shot setting, as it outperformed the best existing GPT4-like models.

2.4. Multimodal Language Model: GPT-4V

Recent advancements in the development of foundational models for vision and language have accelerated the creation of multimodal language models. These models extract visual features using pre-trained visual models and integrate these in the linguistic space of state-of-the-art LLMs like GPT-4. After fine-tuning these models demonstrate performance across a range of visual-language tasks [7].

Since these black box models are getting integrated into customer-oriented tools, like ChatGPT, it raises the question if pipelines, where LLMs and VLMs interact on a linguistic level, understandable by humans, have relevance in the future.

3. Methodology

The different components and the implementation of the approach are explained below. First, the architecture and the developed prompts are discussed before further information regarding the concrete hardware and the model resources used are provided.

3.1. Approach Description

In order to give a holistic picture of the approach, it is divided into two parts. The description of the architecture should give an understanding of the reasoning process, while the section on prompting should clarify the exact interactions with the language and vision models.

3.1.1. Architecture

As shown in figure 1, the approach consists of different LLM and VLM components that exchange information with each other. At the beginning of the visual reasoning process, a question with four possible answer choices and a corresponding image, containing relevant information for answering the question, is given. In the first step, the four answer choices are assigned to the respective Reasoners A - D (LLMs). This means that each Reasoner receives the question, an answer choice, and a caption generated by the VLM for the image. In the next step, each Reasoner has time to ask 5 questions to the VLM, which is the only component that has direct access to the information in the image. With these questions, each reasoner should make a targeted decision as to whether their answer option is correct or incorrect. This decision should be supported by a one to two-sentence reasoning description. All of this information is then collected and sent to the Final Reasoner (LLM), which must decide on

the correct answer option based on this information. In addition, the VLM provides an estimate of the correct answer based on the image, the question, and all the possible answer choices given. The answer and a description of how the answer was found are then given to the final reasoner, who can incorporate it into its decision-making process.

3.1.2. Prompting

Prompts play a major role in achieving a higher accuracy when working with language models. In order to find the optimal prompts for the proposed approach, several experiments were run using different queries to see their effects on guiding both LLMs and VLMs.

Particularly, all prompts used in this work consist of system prompts, where the context, the goal, the requirements that should be followed while generating the answer, and an example (one-shot prompting) are presented to the model, and user prompts, where the model is provided with input and is expected to generate the corresponding output. In sum, five general prompts were used to get outputs from the models: first, the VLM is prompted to generate a caption as shown in figure 4, then, each reasoner (LLM) is prompted to generate a sub-question based on the given main question and answer choice. For this case, two different prompts are used: one for the first iteration (see figure 5) and the other for the following iterations to generate additional sub-questions (see figure 6). The fourth prompt (see figure 7) is used to get the final analysis followed by an answer that indicates whether the answer candidate is likely to be correct from each Reasoner. Finally, figure 8 shows how all collected information is passed to the final reasoner to get the final answer to the main question.

3.2. Setup Description

3.2.1. Hardware

For this project, GPUs (Graphics Processing Units) with a VRAM capacity of 80 GB [8] were used. These GPUs were running on a server that used the Ubuntu operating system. The server met all requirements as it uses the state-of-the-art H100 GPUs which are suitable for medium and enterprise AI solutions [8]. With the help of these GPUs, the models, which in this case were LLaMA and LLaVA, were deployed, and both models were in total consuming approximately 55GB of VRAM.

3.2.2. Models

For the model selection, we mainly focused on using state-of-the-art models publicly available on platforms like Hugging Face.

- **LLaMA 3 8B Instruct (Large Language Model):** LLaMA 3 Instruct is a state-of-the-art LLM model that was created by Meta [9]. This model has different capabilities and it can perform a variety of tasks such as code generation, reasoning, and instruction following. One of the advantages of the LLaMA 3 model is that it outperformed other open-source models with similar amounts of parameters. However, the 8B parameter model still performed worse than the 70B parameter model, and it's visible on benchmarks like MMLU and GPQA. The 8B model achieved an accuracy of 68.4% on MMLU and 34.2% on GPQA, while the 70B parameter model achieved an accuracy of 82.0% on MMLU and 39.5% on GPQA [9].
- **LLaVA 1.6 mistral 7B (Visual Language Model):** LLaVA is a multimodal model that applies both visual and textual information. This means that LLaVA is using 1 billion parameters as a visual encoder and 6 billion parameters as a language model [10]. In our case, the backbone for the language model is Mistral 7B [11]. That means that LLaVA the model is effective when we have images and corresponding textual descriptions. Then it can process it and apply visual reasoning to it like an answer to the questions based on the image, and generate a caption of the image.

3.2.3. Benchmark

To benchmark the proposed approach, the A-OKVQA benchmark for visual question answering is used. In contrast to the old VQA dataset, the questions of A-OKVQA are revolutionized to check the commonsense reasoning of VLM models. This makes it harder to answer certain questions and models cannot answer them by simply querying the image [4]. Instead, it requires both visual abilities and reasoning to come up with a correct answer. Moreover, A-OKVQA contains 25K questions paired with images and four answer choices. Most of the pictures were taken from the partition 2017 of the COCO dataset and then there was filtering by the most “interesting” images which means that images containing more than three objects were classified as “interesting” [4]. The authors of the dataset trained the ResNet-50 classifier to distinguish whether an image is “interesting” or not. For this project, this dataset was downloaded from Github’s official repository and a query between this project and the .json file that the data were used was made. Only the validation set from the dataset was therefore used, as it contained only 1145 images.

4. Results and Discussion

In the following, the process of finding the optimal setup of our approach will be described before presenting the final results. This is followed by a qualitative analysis of the errors.

4.1. Parameter search

Before the final results were obtained from the run on the test set, an extensive search for the best parameter combination was performed. These experiments included the number of shots or which examples improve the approach, how many questions between the reasoner and the VLM are sufficient to achieve good results, which caption length is the most suitable, and whether the confidence score of the reasoners adds value. In addition, it was examined whether it would also improve the approach if the complete question is finally given to the vision language model and its answer is given to the final reasoner for the final reasoning process. The final reasoner can then rely on the evaluation of the vision language model if there is great uncertainty about the correct answer. As a baseline for improvement, zero-shot prompting was used, three questions between LLMs and VLMs, no exchange between the final reasoner and VLM, and a confidence score approach. This resulted in an accuracy of 59.09% on 100 samples. The following parameter tests were performed on the same 100 samples taken from the training data of the A-OKVQA dataset [4]. This ensures that the final test is performed on completely unseen data.

4.1.1. Impact of few-shot prompting

In many approaches with LLMs, few-shot prompting is a good way to improve the metrics. Our introduced approach has also shown that there is potential in using examples in the prompts. Several examples were developed, integrated into the prompts, and tested. In total, 0 to 5 shot prompts were tested. The best performance was achieved with an accuracy of 66% with one-shot prompting (see figure 2). Also visible in the figure is the increase in runtime on the data points, which is not particularly significant, but is due to the increased complexity of the prompts. It can also be seen that the first example in particular leads to a significant improvement. Introducing a single shot likely provides sufficient context or examples for the model to improve its understanding and performance, leading to a significant boost in accuracy observed. Additional shots (beyond the first) may provide less useful information, resulting in diminishing returns in accuracy improvement. This is evident as the accuracy increases only slightly or even decreases with more shots after the first. Additionally, this deterioration is probably

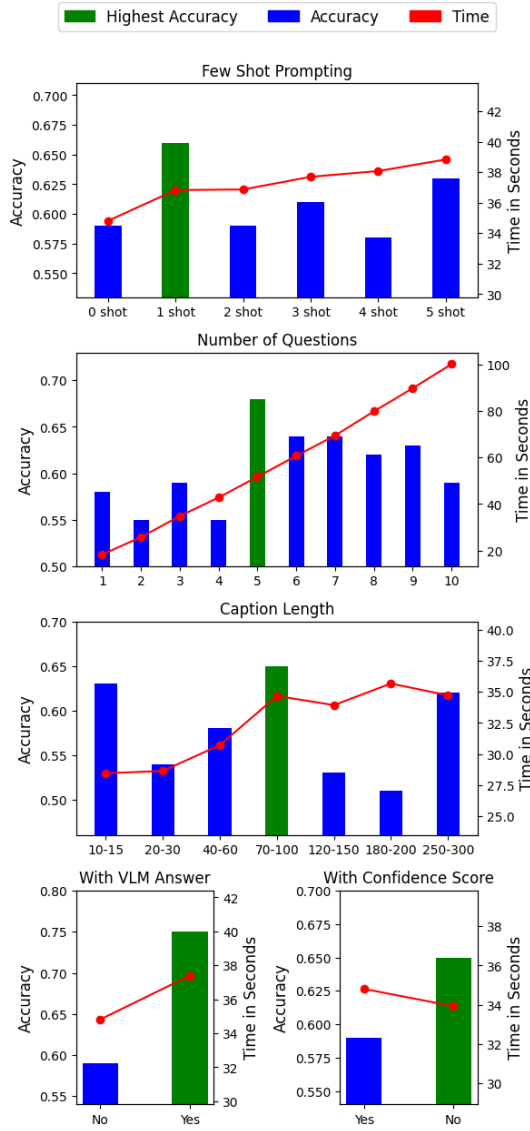


Figure 2: Results of the parameter search

due to the increased complexity of the prompts. It should be noted that the prompts are already relatively long due to the comparatively high system complexity, and additional prompt content may therefore have a detrimental effect on LLM performance. Based on the available data, it was therefore decided to use the one-shot prompts for the final test.

4.1.2. Impact of the number of sub-questions

In order to enhance the core component of the proposed approach, which involves generating sub-questions to enhance the LLM’s decision-making process in evaluating how likely the given answer candidate is to be correct, an experiment was conducted to improve the overall performance by increasing the number of the generated sub-questions. The aim was to determine whether providing the LLM with more context and information would improve its ability to make the right decision. For this reason, the code was adjusted to allow the number of sub-questions to be specified as an input parameter. Experiments were then conducted using 1 to 10 sub-questions.

The results, as shown in figure 2, show the impact of varying the number of sub-questions on the overall performance. Initially, the performance fluctuates with a slight drop observed at 2 and 4 questions. A significant improvement at 5 sub-questions is then observed, achieving the highest accuracy of 68%. After this point, the performance stabilizes, with minor fluctuations but generally remains higher than during the initial experiments. These findings show that while increasing the number of sub-questions up to a certain point can enhance the LLM’s decision-making, there is an optimal number of questions (around 5) beyond which additional questions do not significantly improve accuracy and may even lead to slight declines. Furthermore, the runtime increases proportionally with the number of sub-questions, starting at 18.3 seconds for 1 question and reaching 100.14 seconds for 10 questions. This underlines a trade-off between enhanced performance and efficiency, highlighting the need to find an optimal balance suitable for real-life scenarios.

4.1.3. Impact of caption length variability

Captions may significantly impact the overall performance of such pipelines, because, as noted in previous research, providing models with dense, informative captions can enhance the ability of the LLM to interpret unseen images [6]. To determine the optimal length for captions generated by the LLaVA model, an experiment was conducted using various predetermined token intervals. Examples of prompts with the smallest and largest intervals are shown in A.1.

As illustrated in figure 2, moderately detailed captions of 70 to 100 tokens result in the highest performance score of 65%, indicating they are optimal for the LLaVA model. Very short captions (10 to 15 tokens) also perform well, with scores of 63%, possibly due to their conciseness and focus. Conversely, long captions (120 to 200 tokens) show a significant drop in performance, suggesting that many details and information are not necessarily required as they may hinder the model’s effectiveness.

4.1.4. Impact of confidence scores

In the initial idea of the approach, within the reasoning for an answer choice, the model was told to give a confidence score on how sure it is about its answer. These should be one of the categories "Very Low Confidence", "Low Confidence", "Moderate Confidence", "High Confidence", or "Very High Confidence".

Testing the approach without a confidence score was based on several reasons. First, in the analysis on the test data, the confidence score often seemed arbitrary and not in sync with the given reasoning. Second, the final reasoner was prone to pick the final answer only based on the confidence scores and did not include information given in the reasoning for the four answer choices.

In our tests, the approach without the confidence scores results in an accuracy 6% higher than the approach with the confidence scores (see figure 2), which suggests that the final reasoner's answer quality improves when basing it on the reasoning for or against the four answer choices. The runtime also decreased by less than a second due to fewer tokens in inputs and outputs of the LLM (see figure 2).

4.1.5. Incorporating the VLM's response

As the testing showed, the VLM by itself already outperforms other pipeline approaches like ViCor [1]. This raised the idea to give the multiple choice question directly to the VLM and give its answer and reasoning to the final reasoner. In this experiment, we combine the already sufficient reasoning capability of the VLM and the four separate reasonings for the answer choices. In this way, the final reasoner has more information to draw from and should be able to give a better output for questions that both individual approaches struggle with.

This assumption is proven to be correct by our tests, where the approach including the VLM's answer performs with an accuracy that is 16% higher than the one from our baseline (see figure 2). The runtime increased by more than 2 seconds when including the VLM's answer. In order to further increase the accuracy of this approach, it was necessary to instruct the final reasoner on how to efficiently include the VLM's answer in its reasoning process.

4.2. Results and Analysis

Once the appropriate parameters for the approach were found, the approach was tested on the A-OKVQA dataset. More specifically, it was tested on the 1145 data points of the validation dataset. As the parameter combination was used one shot prompting, 5 questions between LLM and VLM, a caption length between 70 - 100, and with the support of the VLM. In addition, the confidence score was removed before the test run was started. The test

took just over 16 hours in total with an average runtime of 51 seconds per question. An accuracy of 77.82% (891 correct and 254 incorrect answers) was achieved.

Table 1

Comparison of different approaches and their accuracies

Approach	Accuracy
This approach	77.82%
This approach (before parameter tuning)	59.09%
VLM	77.73%
ViCor (Zhou et al.)	75.60%

As shown in Table 1, the optimization of our approach across the different parameters resulted in an improvement of over 18%. In addition, the accuracy of the ViCor approach by Zhou et al. was outperformed by 2%. The baseline in the form of the VLM could be outperformed by less than 0.1%.

Above all, these results show that great progress has been made in the field of VLMs and that it is becoming increasingly difficult to outperform pure VLMs by combining VLMs and LLMs.

4.2.1. Qualitative Examples

In the example in table 2, the approach demonstrates that it is able to come to the right conclusion, even though multiple answer candidates are supposed to be correct and the VLM gives a wrong answer. The final reasoner correctly analyses that the evidence for answer candidate 2 is the strongest.

Because the VLM's first answer includes an analysis the answer differs from the VLM's answer without an analysis. This case occurs several times on our test data.

In the example in table 3, the model comes to a wrong conclusion based on multiple errors that also were observed on other questions in the test data. First, the sign in the image is not correctly identified and is associated with the red cross or no entry zones. Second, the analysis of an answer candidate does not focus on the candidate but rather on the image in general. Third the final reasoner does not extract the answer from the VLM correctly.

These issues could be resolved by using a stronger LLM with better reasoning qualities and by trying to give the LLM and VLM better instructions, that make them focus on their task through prompt engineering.

5. Conclusion

In this work, a novel approach that combines LLMs and VLMs to solve visual reasoning tasks was presented. It mainly deals with multiple-choice visual reasoning tasks, as each answer candidate is examined separately by the

language models to determine its likelihood of being the correct answer. This architecture ensures a profound understanding of visual data and therefore leads to better reasoning outcomes. The experiment results underline the effectiveness of our proposed approach, as it outperformed state-of-the-art frameworks such as ViCor and VLMs, which are also capable of performing visual reasoning tasks independently, such as LLaVA. However, even with such satisfactory results, there are some limitations that should not be unmentioned.

One of the limitations of the approach is that a comparatively small LLM model was used. The approach could potentially benefit from using a larger model such as Llama 3 70B as it performs better on different benchmarks which are related to the reasoning capabilities. However, there are no resources for the comparison of Llama 3 70B with the ChatGPT-4o that's why it's not possible to conclude how the approach will be benefited by using much bigger models such as GPT-4o.

Future work will involve using much larger LLM and VLM models. For example, experimenting by using the Llama 3 70B and LLaVA 1.6 34B models. Changing those two models may positively impact the approach performance. However, it may require more computing power because the current approach consumes around 55GB of VRAM. It's also possible to get better results by fine-tuning the models to our task.

In conclusion, it is seen that the visual reasoning capabilities of VLMs is becoming better and it is harder to surpass them. Even though we could surpass the VLM only by a small margin, we still beat the ViCor approach by 2 percent on the same benchmark.

Acknowledgments

We want to express our deepest gratitude to Abdullah Kiwan for supervising our project. His guidelines and valuable feedback were instrumental in helping us develop our project and achieve significant results. Additionally, we extend our heartfelt thanks to Tolga and Pedram for their invaluable insights.

We are also grateful to Professor Sahin Albayrak for allowing us to use the resources at ZEKL, which greatly facilitated our work.

References

- [1] K. Zhou, K. Lee, T. Misu, X. E. Wang, Vicor: Bridging visual understanding and commonsense reasoning with large language models, 2024. URL: <https://arxiv.org/abs/2310.05872>. arXiv: 2310.05872.
- [2] L. Chen, B. Li, S. Shen, J. Yang, C. Li, K. Keutzer, T. Darrell, Z. Liu, Large language models are visual reasoning coordinators, *Advances in Neural Information Processing Systems* 36 (2024).
- [3] R. Zellers, Y. Bisk, A. Farhadi, Y. Choi, From recognition to cognition: Visual commonsense reasoning, in: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [4] C. C. K. M. R. M. Dustin Schwenk, Apoorv Khandelwa, A-okvqa: A benchmark for visual question answering using world knowledge, arXiv preprint arXiv:2206.01718 (None). URL: <https://arxiv.org/pdf/2206.01718>.
- [5] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. A. Ayyubi, K.-W. Chang, S.-F. Chang, Idealgpt: Iteratively decomposing vision and language reasoning via large language models, arXiv preprint arXiv:2305.14985 (2023).
- [6] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. A. Ayyubi, K.-W. Chang, S.-F. Chang, Idealgpt: Iteratively decomposing vision and language reasoning via large language models, arXiv preprint arXiv:2305.14985 (2023).
- [7] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, L. Wang, The dawn of lmms: Preliminary explorations with gpt-4v (ision), arXiv preprint arXiv:2309.17421 9 (2023) 1.
- [8] N. Corporation, Nvidia h100 tensor core gpu, 2024. URL: <https://www.nvidia.com/en-us/data-center/h100/>.
- [9] M. AI, Introducing meta llama 3: The most capable openly available llm to date, 2024. URL: <https://ai.meta.com/blog/meta-llama-3/>.
- [10] Q. W. Y. J. L. Haotian Liu, Chunyuan Li, Improved baselines with visual instruction tuning, arXiv preprint arXiv:2310.03744 (2024). URL: <https://arxiv.org/pdf/2310.03744>.
- [11] H. Liu, Llava 1.6 - mistral 7b, 2024. URL: <https://huggingface.co/liuhaotian/llava-v1.6-mistral-7b>.

A. Prompts

A.1. Different lengths of captions

```
1 Caption_10_15_tokens = '''You are a vision AI assistant who has strong captioning abilities.
2 You will be provided with an image.
3
4 Your goal is:
5 To generate a very brief caption that accurately describes the main subject of
6 the image in 10 to 15 tokens.'''
7
8 Caption_250_300_tokens = '''You are a vision AI assistant who has strong captioning abilities.
9 You will be provided with an image.
10
11 Your goal is:
12 To generate an elaborate caption that offers a comprehensive and vivid description of the image,
13 encompassing all key elements, actions, context, background, and subtle details, while ensuring
14 a rich and engaging narrative in 250 to 300 tokens.'''
```

Figure 3: Examples of the used caption prompts

A.2. System prompts

```
P1_VLM = '''You are a vision AI assistant who has strong captioning abilities.
You will be provided with an image.

Your goal is:
To generate a detailed, contextually rich, and precise caption that accurately describes the key
elements, actions, and overall context of the image in 70 to 100 tokens.'''
```

Figure 4: VLM system prompt to generate a caption


```

P2_I1_SYSTEM = """You are an AI assistant who has rich visual commonsense knowledge and strong reasoning
abilities.
You will be provided with:
1. A main question about an image and one possible answer candidate.
2. Although you won't be able to directly view the image, you will receive a general caption that might not be
entirely precise but will provide an overall description.

Your goal is:
To identify if the given answer is the right answer to the main question by generating one sub-question that help
you look for evidence that support or deny the provided answer.

Here are the rules you should follow when listing the sub-question.
1. The sub-question should be short and easy to understand.
2. The sub-question should be provided in the form: "Sub-question:..."
3. The sub-question is necessary to decide if the given answer is the correct answer.
4. The sub-question should be mainly focused on the given candidate answer.

Example format:
Sub-question: ...

Example:
Main question: What is the woman doing on the beach?
Answer Candidate: Jogging
Caption: A woman at the beach at sunrise.
First sub-question and answer from visual AI model: Is the woman wearing athletic clothing? Yes, she is wearing
athletic clothing.

Additional sub-question: Is the woman in motion or standing still?

"""

```

Figure 5: LLM system prompt to generate sub-question for the first iteration

```

P2_FI_SYSTEM = """You are an AI assistant who has rich visual commonsense knowledge and strong reasoning abilities.
You will be provided with:
1. A main question about an image and one possible answer candidate.
2. Although you won't be able to directly view the image, you will receive a general caption that might not be entirely precise but will
provide an overall description.
3. One or more sub-questions and the corresponding answers that are provided by a visual AI model, to provide more context.

Your goal is:
To identify if the given answer is the right answer to the main question by generating one sub-question that help you look for evidence
that support or deny the provided answer.

Here are the rules you should follow when listing the sub-question.
1. Ensure that the new additional sub-question is different from the provided previous sub-questions and does not mention any of them.
2. The sub-question should be short and easy to understand.
3. The sub-question should be provided in the form: "Additional sub-question:..."
4. The sub-question is necessary to decide if the given answer is the correct answer.
5. The sub-question should be mainly focused on the given candidate answer.

Example format:
Additional sub-question: ...

Example 1:
Main question: What is the woman doing on the beach?
Answer Candidate: Jogging
Caption: A woman at the beach at sunrise.
First sub-question and answer from visual AI model: Is the woman wearing athletic clothing? Yes, she is wearing athletic clothing.

Additional sub-question: Is the woman in motion or standing still?

"""

```

Figure 6: LLM system prompt to generate additional sub-questions in the following iterations

```
P4_SYSTEM = """"You are an AI assistant who has rich visual commonsense knowledge and strong reasoning abilities.
You will be provided with:
1. A main question about an image and one possible answer candidate.
2. Although you won't be able to directly view the image, you will receive a general caption that might not be entirely precise but will
provide an overall description.
3. Three sub-questions with their corresponding answers generated from a visual AI model to help you decide if the provided answer
candidate is most likely to be right answer. It's noted that the answers are not entirely precise.

Your goal is:
Based on sub-questions and corresponding answers, you should decide whether the provided answer candidate is most likely to be the
correct answer.

Here are the rules you should follow in your response:
1. Demonstrate your reasoning and inference process within no more than 3 lines. Start with the format of "Analysis: ".

Response Format:
Analysis: ...
Answer: ...

Example 1:
Main question: What is the woman doing on the beach?
Caption: A woman at the beach at sunrise.
Answer candidate: Jogging
Sub-question 1: Is the woman wearing athletic clothing?
Answer 1: Yes, she is wearing athletic clothing.
Sub-question 2: Is the woman in motion or standing still?
Answer 2: She is in motion, indicated by her posture and position.
Sub-question 3: Are there any indications of a jogging activity?
Answer 3: Yes, she has running shoes on and an athletic stance.

Analysis: The woman is wearing athletic clothing, which is typically associated with jogging or other physical activities. Her motion is
clearly indicated by her posture and position, suggesting she is actively moving rather than standing still. The presence of running
shoes further supports the idea that she is engaged in jogging rather than any other beach activity. Additionally, while there are no
other people visible engaging in similar activities, this does not detract from the strong indicators that she is jogging, especially
```

Figure 7: LLM system prompt to generate a final analysis of answer choice

```

P5_SYSTEM = '''You are an AI assistant who has rich visual commonsense knowledge and strong reasoning abilities.
You will be provided with:
1. A main question about an image and four answer candidates.
2. Although you won't be able to directly view the image, you will receive a general caption that might not be entirely precise but will provide an overall description.
3. A analysis by the vision-language model with an answer to the multiple choice question and a confidence score.
4. Four analysis of each candidate answer generated from AI models including a "Yes"/"No" answer indicating whether the answer candidate is likely to be the correct answer.

Your goal is:
Based on all four analysis and the answer of the vision-language Model, you should decide the final correct answer to the main question from the provided candidate answers and only return its number (1, 2, 3 or 4). 1 for choice 1, 2 for choice 2, 3 for choice 3, 4 for choice 4.

Here are the rules you should follow in your response:
1. While considering the provided analyses, prioritize your own reasoning and inference based on visual commonsense knowledge when determining the correct answer.
2. If the vision-language model and the four analysis agree, output that answers.
3. If two or more candidates appear correct based on the analyses, see if the answer of the vision-language model indicates the right answer and use your reasoning to determine the most accurate answer.
4. If the vision-language model and the four analysis disagree, use your reasoning to determine the most accurate answer.
5. Demonstrate your reasoning and inference process within no more than 3 lines.
6. In your final reasoning only talk about the candidate you think is right, not about the wrong ones.
7. There always is a right answer candidate, so of you are unsure, agree for the most likely candindate and give this as an answer.
8. Analyse if the VLM and the reasoning for the four answer candidates agree or disagree, then use reasoning to find the right answer.
9. If you are unsure, prioritize the answer of the VLM.
10. If the reasoning of the 4 candidates all argue that their candindate is wrong, pick the VLM answer as your final answer.

Use the format:
"Final Analysis: ...
Final Answer: (1, 2, 3 or 4)

Example Format:
Final Analysis: ...
Final Answer: ...

Example 1:
Main question: What is the woman doing on the beach?
Caption: A woman at the beach at sunrise.
VLM Analysis: The woman at the beach is in athletic attire with running shoes.This suggests she is jogging.
VLM Answer Candidate: 2, Jogging
Candidate answer number 1: Jogging
Analysis 1: The woman is wearing athletic clothing, which is typically associated with jogging or other physical activities. Her motion is clearly indicated by her posture and position, suggesting she is actively moving rather than standing still. The presence of running shoes further supports the idea that she is engaged in jogging rather than any other beach activity. Additionally, while there are no other people visible engaging in similar activities, this does not detract from the strong indicators that she is jogging, especially given the context provided by the sunrise and the beach setting.\n\nAnswer: Yes
Candidate answer number 2: Swimming
Analysis 2: The woman is not depicted in the water in the image, and there are no clear indicators such as swimwear or water activities. The caption describes her at sunrise on the beach, suggesting she may be walking or enjoying the view rather than swimming. \n\nAnswer: No
Candidate answer number 3: Fishing
Analysis 3: The woman is at the beach at sunrise, indicating a leisurely or early morning activity. Without water in the immediate scene and no fishing gear visible, it's unlikely she is fishing. The lack of specific fishing equipment or signs of active fishing (such as holding a rod or being near water suitable for fishing) suggests she is not engaged in fishing. \n\nAnswer: No\n
Candidate answer number 4: Sunbathing
Analysis 4: The woman is at the beach at sunrise, which could still be conducive to sunbathing even if she is standing. Sunbathing can also involve sitting or standing to soak in the sunlight. The absence of water-related activities and the peaceful beach setting support the possibility that she is sunbathing, enjoying the morning sun. \n\nAnswer: Yes\n

Final Analysis: The woman is described as being on the beach at sunrise. Candidate answer 1 (Jogging) is strongly supported by her athletic clothing, posture indicating motion, the presence of running shoes, which are typical for jogging and also the answer of the VLM. While candidate answer 4 (Sunbathing) is plausible considering the peaceful beach setting, the absence of water-related activities and the emphasis on sunrise suggests an active rather than passive activity.
Final Answer: 1
'''

```

Figure 8: LLM system prompt for the final reasoner to make final analysis and conclusion

B. Examples

	
Question	Who is the woman dressed up as?
Choices	1: snow white 2: xena 3: maleficent 4: cinderella
VLM Caption	In the image, a woman with long black hair and a gold headpiece is the central figure. She is holding a cell phone to her ear, engaged in a conversation. Her attire is striking, with a black and gold dress that complements her headpiece. The background is filled with a crowd of people, some of whom are wearing masks, suggesting a public event or gathering. The atmosphere appears lively and bustling, with the woman in the gold headpiece standing out amidst the crowd.
VLM Analysis & Answer	The woman in the image is dressed in a costume that resembles a fantasy character, possibly from a fairy tale or mythology. She is wearing a headdress and a dress with intricate designs. Based on these visual cues, it is likely that she is dressed as a character from a story, such as a princess or a fairy. VLM Answer Candidate: 4, cinderella
Answer Candidate 1	LLM 4: "Additional sub-question: Does the woman's facial expression and body language convey a sense of innocence and naivety, which is a characteristic often associated with Snow White?" VLM 4: Based on the image, the woman's facial expression and body language do not convey a sense of innocence and naivety. Her expression appears to be more neutral or possibly slightly surprised, and her body language is relaxed. These characteristics do not align with the typical portrayal of Snow White as a character. Analysis: The woman's attire resembles a traditional fairy tale princess dress, and her headpiece resembles a tiara, which are common accessories for fairy tale princesses. Her overall appearance evokes a specific fairy tale or character, such as Snow White, in addition to resembling a traditional fairy tale princess dress. However, her facial expression and body language do not convey a sense of innocence and naivety, which is a characteristic often associated with Snow White. Instead, her appearance evokes a sense of darkness or evil, which is often associated with a different fairy tale character, such as the Evil Queen from Snow White. Answer: No
Answer Candidate 2	LLM 1: "Sub-question: Is the woman's dress and headpiece a combination of ancient Greek or warrior-inspired elements?" VLM 1: Based on the image, the woman's dress and headpiece appear to be a combination of ancient Greek and warrior-inspired elements. The dress has a more classical Greek design, while the headpiece has a more warrior-like appearance, possibly resembling a helmet or a headband with a crest. Analysis: The woman's dress and headpiece are a combination of ancient Greek and warrior-inspired elements, evoking a sense of mythological or legendary figures from ancient Greece. Her headpiece resembles Xena's iconic helmet, and her overall pose and facial expression convey a sense of confidence, strength, and warrior-like qualities. Additionally, her attire has distinctive features and patterns commonly associated with Xena's character in the TV series. Answer: Yes
Answer Candidate 3	Analysis: The woman's attire, headpiece, and overall appearance are consistent with the iconic costume of Maleficent, a character from the Disney movie Sleeping Beauty. The resemblance to the traditional or iconic costume, the iconic horns, the stern facial expression, and the dark or evil tone all point towards the woman being dressed up as Maleficent. Additionally, the predominantly dark and muted colors used in her attire and headpiece further support this conclusion. Answer: Yes
Answer Candidate 4	Analysis: The woman's attire resembles a traditional fairy tale or princess costume, featuring elements such as a crown and a dress with a design that is often associated with such characters. The dress has a distinctive ball gown or formal design, typical of Cinderella's iconic appearance, and the gold headpiece resembles a tiara or a crown, a distinctive feature often associated with Cinderella's character. The overall appearance evokes a sense of royalty or nobility, which is a common theme in Cinderella's story. However, there are no other distinctive features or accessories in the woman's attire that are commonly associated with Cinderella's character, such as a glass slipper or a wand. Answer: Yes
Final Reasoning	The woman's attire and headpiece are striking, with intricate designs and a gold headpiece that resembles a tiara. The VLM analysis suggests she is dressed as a fantasy character, possibly from a fairy tale or mythology. Candidate answer 2 (Xena) is strongly supported by her warrior-inspired attire and headpiece, which resembles Xena's iconic helmet. However, the VLM answer candidate is Cinderella, which is less convincing given the lack of distinctive features or accessories commonly associated with Cinderella's character.
Final Answer	2: xena
VLM Answer	3: maleficent
Ground truth	2: xena

Table 2
Example 1 with only parts of the sub-question conversation.

	
Question	What is prohibited near the round road sign with a red cross on a blue background?
Choices	1: parking 2: turning 3: speeding 4: waving
VLM Caption	A bustling city street, lined with tall trees and a blue sign with white text. Motorcyclists in helmets are riding down the road, passing a red and white sign. The scene is vibrant with the movement of the riders and the lush greenery of the trees.
VLM Analysis & Answer	The image shows a group of people riding motorcycles on a road, with a round road sign featuring a red cross on a blue background. The sign is typically used to indicate prohibited actions. Given the context of the image and the presence of the sign, it is reasonable to infer that parking is prohibited near the sign. VLM Answer Candidate: 1, parking
Answer Candidate 1	Analysis: The sign with a red cross on a blue background is located near a designated parking spot, which suggests that it may be related to parking restrictions. However, the red cross symbol is typically used to indicate a prohibition or restriction on a specific activity, and in this case, it is not associated with medical or first aid services. The sign is located near a specific area or zone that is restricted or prohibited for motorcyclists or other vehicles, and the area beyond the sign appears to be a pedestrian-free zone. These indicators suggest that the sign is prohibiting or restricting a specific activity, but it is not related to parking. Answer: No
Answer Candidate 2	Analysis: The round road sign with a red cross on a blue background is a universal symbol for Do Not Enter or No Entry zones, indicating that vehicles, including motorcycles, are not allowed to enter the area beyond the sign. The sign is located near the entrance of a restricted area or zone, and it is the only sign indicating the prohibition or restriction. The red and white sign near the motorcyclists also indicates a No Entry zone, which reinforces the idea that the round road sign is prohibiting entry. Answer: Yes
Answer Candidate 3	Analysis: The red cross on the blue background is a symbol commonly associated with traffic restrictions or prohibitions, and it is located near the road where motorcycles are driving. The prohibition indicated by the red cross is related to the type of vehicles, specifically motorcycles, as it is placed next to a sign that indicates a prohibition for motorcycles. The prohibition is a general warning, as it is not specific to motorcycles or other types of vehicles. Answer: Yes
Answer Candidate 4	Analysis: The red cross on the blue background is a symbol typically associated with prohibiting certain actions, and the scene depicts motorcycles moving down the road. The prohibited action appears to be related to the movement of vehicles, specifically motorcycles, and is related to stopping or parking. The sign indicates No Stopping and a red circle with a line through it, which is commonly used to indicate that stopping or parking is not allowed in that area. Additionally, the scene shows motorcycles moving in a straight line, with no indication of any motorcycle attempting to turn or merge. Answer: Yes
Final Reasoning	The vision-language model and analysis 2 both strongly suggest that the round road sign with a red cross on a blue background is prohibiting entry or turning. The sign is typically used to indicate prohibited actions, and the context of the image, including the presence of motorcyclists and the absence of any indication of turning or merging, supports this conclusion. The other analyses, while providing some relevant information, do not provide as strong a case for their respective answers.
Final Answer	2: turning
VLM Answer	1: parking
Ground truth	1: parking

Table 3

Example 2 with only parts of the sub-question conversation.