The slide features a white background with decorative geometric elements in the corners. These include dark gray and light gray triangles and chevrons, some with thin black outlines, pointing towards the center. The main title is centered in a large, bold, dark gray sans-serif font.

VISUAL SPATIAL REASONING OF LARGE LANGUAGE MODELS

Presentation: Jinlong Liu



LIST OF CONTENT

01.

PROJECT INTRODUCTION

A brief description of this project

02.

AIMS AND OBJECTIVES

Research aims and objectives

03.

OUTPUT

A brief introduction to the experiment result

04.

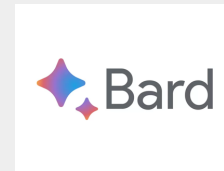
EVALUATION

Evaluate the experiment models in detailed



01. PROJECT INTRODUCTION

In recent years, Large Language Models (LLMs) have transformed the field of Natural Language Processing (NLP) with their impressive achievements. Such as ChatGPT, Bert and Bard. As they updated in a raptly speed, there are a lot of test work also starting.




Source of all images:

<https://chat.openai.com>

<https://bard.google.com/>

<https://ivet360.com/google-bert-for-vets/>

- 
- The testing works are focused on the professional domain, like the ability of logic, answer test questions in various expert domain.
 - This project is based on the work of testing ability of spatial reasoning, which is a subset of testing ability of commonsense reasoning.
 - But it is different with current research that adding visual input into prompt not just text input. Which can be considered as multi-model problem.




02.

AIMS AND OBJECTIVES



AIMS:

1. Investigate performance of Visual Question Answering (VQA) tasks.
 2. Make comparison between different versions of LLMs.
 3. Find challenges of LLM's on Visual Spatial Reasoning.
- 



MAIN OBJECTIVES

1. Design a diverse dataset with corresponding questions.
2. Conduct systematic experiments by using prepared dataset.
3. Analyse the experimental results to identify the challenges of LLMs
4. Provide the insight of strong points and weakness of current LLMs.



03. **OUTPUT**

Experiment Setting

In experiment, there are three models have been tested, and their test settings listed as below:

- Bard is tested on the official website[1]
- Visual ChatGPT is deployed on Colab Pro server with a Nvidia V100 graph card.
- MiniGPT-4 is test on the demon website provided by the development team, with default settings.[2]

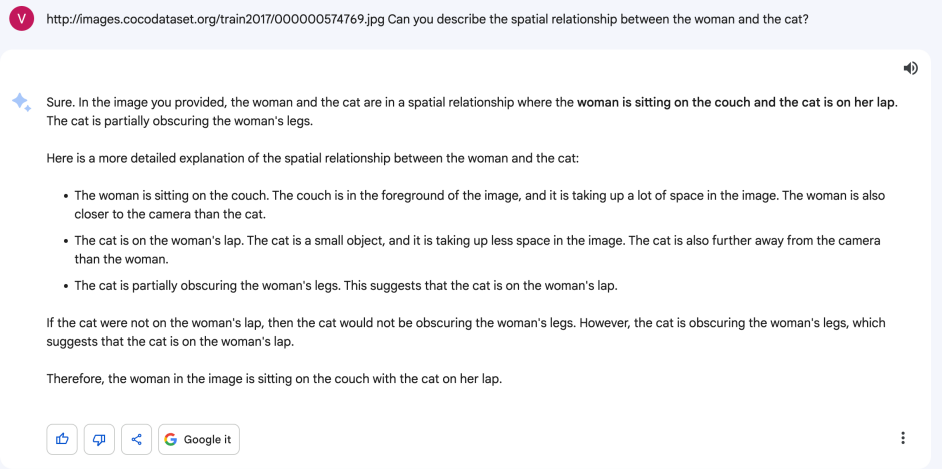


Fig.1 The user surface of Bard

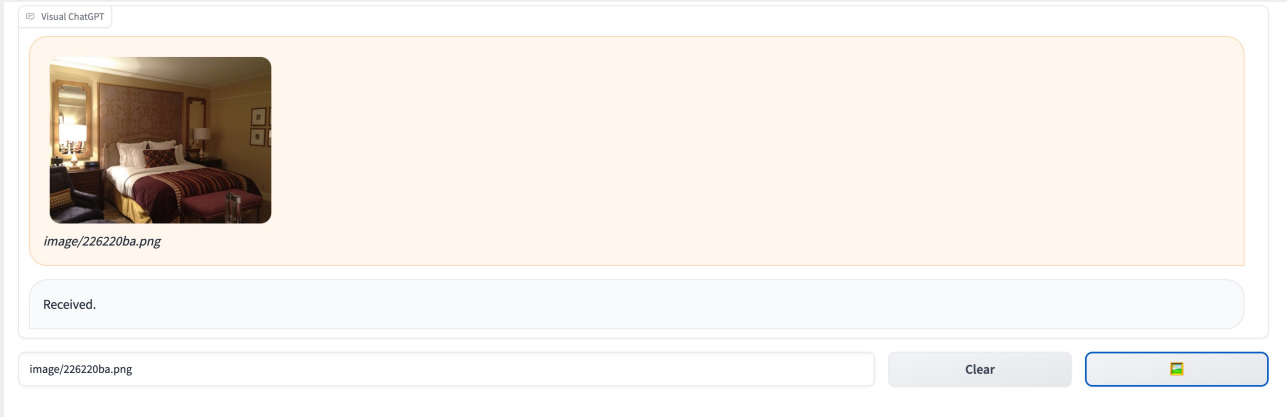



Fig.2 The user surface of Visual ChatGPT

Demo of MiniGPT-4

[NOTE] It is possible that you are waiting in a lengthy queue.
You can duplicate and use it with a paid private GPU.

 Duplicate this Space

Alternatively, you can also use the demo on our [project page](#).

This is the demo of MiniGPT-4. Upload your images and start chatting!

Project Page Github Code Paper PDF

Image

Drop Image Here
- or -
Click to Upload

Upload & Start Chat

Restart

beam search numbers1

Temperature1

MiniGPT-4

User

Please upload your image first

Fig.3 The user surface of MiniGPT-4

Question Template


Input Image	Questions
	<ol style="list-style-type: none">1) Can you describe the spatial relationship between these two laptops?2) Is this black laptop on the left of the white laptop?3) Show me your specific reasoning steps that lead you to the answer, better in detailed explanation.

Fig 4. Question Template

The question set is shown on the left side. It will have three questions in each set:

1. First one is to test the understanding abilities without leading.
2. Second one is to test the abilities with leading.
3. Third one is to test the reasoning process of all models.

There are 80 question sets in total and divided into two categories: Easy and hard.

Overall Accuracy

Model	Easy	Hard	Total
Bard	45.0%	42.5%	43.75%
Visual ChatGPT	62.5%	57.5%	60.0%
MiniGPT-4	65.0%	57.5%	61.25%

Fig 5. The accuracy of Question 2

Based on the metrics, the MiniGPT-4 model manifests superior accuracy, succeeded by Visual ChatGPT and subsequently Bard. Notably, Visual ChatGPT's performance closely parallels that of MiniGPT-4. A more granular analysis will be presented in the ensuing section. The reason of giving up Question 1's answer is that Bard gives all the possible answer to the questions, which is not fair for other models to compare with.



04.

EVALUATION

Examination of Accuracy

In the analysis of all answers of question 2, there are 10 types of error occurred:

1. Left-Right Confusion
2. Recognition Error
3. Judgement Fault in Front and Back
4. Positional Fault
5. Item Misorientation
6. Confused Expression
7. No Exact Answer
8. No Response
9. Logic Fault
10. Vertical Fault

The detailed explanation of the types of error and the types of error of each question you can find it in my report document.

Examination of Accuracy

In this slides, the detailed discussion will focus on “Left-Right Confusion”, “Judgement Fault in Front and Back” and “Positional Fault”. For more detailed answer of all models, you can check it in my report document and appendix part, there are 80 pages.

Left-Right Confusion:

Example Question 9-2: Is the cat on the right of the laptop?

Bard: No, the cat is not on the right of the laptop. The cat is sitting on the left side of the laptop, ...

Visual ChatGPT: Yes, the cat is on the right of the laptop.

MiniGPT-4: No, the cat is on the left of the laptop.

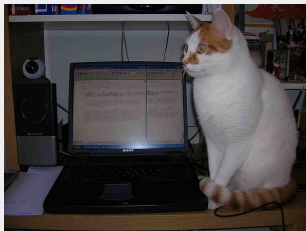


Fig 6. Image of Question 9

From the answer we can find that Bard and MiniGPT-4 have the opposite spatial relationship compare with the correct one. The cause of this is not sure yet, because in the former question9-1: MiniGPT-4 have the correct answer, then it change to wrong.

Judgement Fault in Front and Back:

Example Question 14-2: Is the horse on the back of the vehicle?

Bard: No, the horse is not on the back of the vehicle. The horse is in front of the vehicle, pulling it ...

Visual ChatGPT: The image shows a red truck parked next to a horse. The horse is not on the back of the vehicle.

MiniGPT-4: No, the horse is not on the back of the vehicle. The horse is standing in front of the vehicle, next to it...



Fig 7. Image of Question 14

From the answer, it is clear that models have difficulties in judging spatial relationships in this angle of view. Because they didn't see the back of all objects, they can understand which side the object facing, but just use for front side, not for other relationships.

Positional Fault:

Example Question 64-2 : Is this microwave behind the toaster?

Bard: No, the microwave in the image you provided is not behind the toaster. The microwave is on top of the toaster ...

Visual ChatGPT: The microwave is not behind the toaster, it is located to the left of the toaster.

MiniGPT-4: No, the microwave is in front of the toaster.



Fig 8. Image of Question 64

Compare with the image and all answers, all models seems have issues in this kind of images, that objects are not in a prefect view, part of them covered by other objects, letting models hard to make judgements. That may lead to this type of error.

Another type of error mostly appears on Bard: Vertical Fault

Example Question 6-2 : Is the blue bag above the yellow package?

Bard: No, the blue bag is not above the yellow package. The blue bag and the yellow package are both sitting on the bench, and there is no space between the two objects. Therefore, the blue bag cannot be above the yellow package.



Fig 9. Image of Question 6

For this type of error, this may be caused by the model's judgement methods, that it only considers up and down when two objects' center line is matching perfectly. This type of error is not only on this type of relationships, but also in left and right.

Scrutiny of the Reasoning Process

In this subsection, we will discuss about the Reasoning Steps shows by the models, by analyzing these answers, I found that every model has its style of expression ways and the logic chain is quite the same in each reasoning answer for each of them.

The example question is Question 17, and all models have the correct in question 2.

Question 17-3: Show me your specific reasoning steps that lead you to the answer, better in detailed explanation. Image of Question 17 shows by fig 10.



Fig 10. Image of Question 17

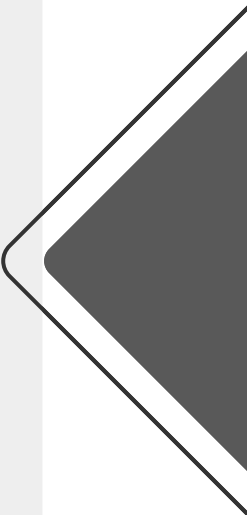


Bard:

Sure, here are my specific reasoning steps that lead me to the answer: I first looked at the image and saw that the boy was facing a pizza on a table. I then considered the question "Is the pizza in front of the boy?". I realized that the answer to this question depends on the definition of "in front of".

I decided to use the following definition of "in front of": "located in front of or facing someone or something". Based on this definition, I concluded that the pizza was in front of the boy, because the boy was facing the pizza and his hands were on the table, near the pizza...

Bard's reasoning step likes to understand the meaning of "in front of", only just a few part of content is analyzing the content of the spatial relationship of objects, And for the key point support it's answer is that the boy is facing the pizza, which is not strong enough to convince human.



Visual ChatGPT:

The image provided is of a boy holding a pizza. From the description, it is clear that the pizza is in front of the boy. To confirm this, I used the Get Photo Description tool to observe the image and confirm that the boy is indeed holding the pizza in his hands.

Visual ChatGPT's reasoning step is lack of performance, it totally has no proof steps, just wants to tell you the image just like this.

MiniGPT-4:

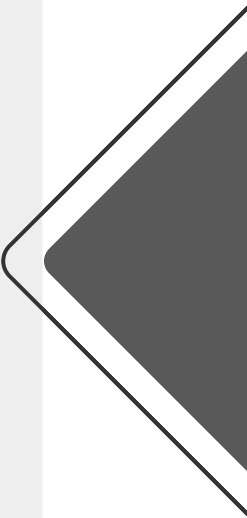
Sure, Sure, here are the specific reasoning steps that led me to the answer:

1. The image shows a young boy sitting at a table with a large pizza in front of him.
2. The pizza is on a large round plate, which suggests that it is a personal-sized pizza.
3. The boy is using a fork to cut into the pizza, which suggests that he is eating the pizza.
4. The boy is wearing a red and white plaid shirt, which suggests that he is in a casual setting.
5. The tablecloth is white and the walls of the room are painted a light blue, which suggests that the setting is a restaurant or cafe.
6. There are a few small white flowers on the table in front of the boy, which suggests that the setting is a nice, upscale restaurant or cafe.
7. The pizza appears to be a pepperoni pizza, which is a common type of pizza.
8. Based on these observations, it is likely that the pizza is in front of the boy.



MiniGPT-4:

MiniGPT-4's reasoning steps is just describing the content of the image, still no proof steps about how it gets this correct answer.



Scrutiny of the Reasoning Process

Based on the analysis above, we can easily found that Bard have the best performance in reasoning steps, but it still not good as human. It still greater than Visual ChatGPT and MiniGPT-4. MiniGPT-4 in some case could have a clear reasoning steps, but for Visual ChatGPT is not showing any abilities of reasoning.

The reasonable explanation might be this: Bard can keep its model up to date so that lead to a better answer, and MiniGPT-4 build with the LLMs that still new in this year, but for Visual ChatGPT is using GPT-3 for text generation part and it is not build in the model, that leads the bad performance in reasoning.

Deliberative Discussions

Form the evaluation steps above, we can found that in the Q&A stage, the models can have a good performance, especially in Easy. With the complexity of image improved, the performance is getting down.

In reasoning stage, the models start struggling, that shows them lack of abilities in logic, or they have issues in image reading so that lead to the performance in reasoning.

Deliberative Discussions

So there are 5 challenges the models are facing:

1. The models' struggles in identifying spatial interrelations among image objects.
2. Difficulties in reasoning about these spatial interconnections.
3. A constrained ability to generate comprehensive justifications for their determinations.
4. A limited understanding of visual intricacies within images.
5. The language generation frameworks not being refined enough to weave detailed explanations.



REFERENCES

REFERENCES MENTIONED IN THIS SLIDES

- [1]. <https://bard.google.com/>
- [2]. <https://huggingface.co/spaces/Vision-CAIR/minigpt4>



**THANK
YOU**

