

About Us





What to expect?

#multimodal-ml - Channel
@multimodal - Role

Challenges









Paper Reading Session

Guest Speaker Session

Coding Session

Deep Dive Session

Second and Fourth Friday 10 AM ET More Exciting Updates Soon

Code DeepDive
LLaVA-CoT: Let Vision
Language Models
Reason Step-by-Step

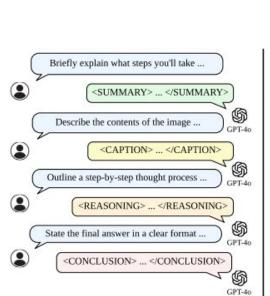
Recap

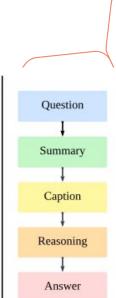
- LLaVA-CoT implements a four-stage vision language reasoning process: summary, caption, reasoning, and conclusion.
- Created LLaVA-CoT-100k dataset with detailed reasoning annotations for training.
- Introduced stage-level beam search for improved inference scaling.
- Achieves superior performance relative to larger and closed-source models, underscoring the effectiveness of LLaVA-CoT for multimodal reasoning.

Recap: Dataset Generation

Dataset	Type	Size
ShareGPT4V [9]	General VQA	31.3k
ChartQA [40]	General VQA	17.2k
A-OKVQA [48]	General VQA	16.1k
AI2D [25]	Science-Targeted VQA	11.4k
GeoQA+[7]	Science-Targeted VQA	11.4k
ScienceQA [36]	Science-Targeted VQA	5.6k
DocVQA [41]	General VQA	4.0k
PISC [30]	General VQA	1.0k
CLEVR [24]	General VQA	0.5k
CLEVR-Math [14]	Science-Targeted VQA	0.5k

Table 1. The number of samples selected from each benchmark.





Recap: Dataset Generation Prompt

I have an image and a question that I want you to answer. I need you to strictly follow the format with four specific sections: SUMMARY, CAPTION, REASONING, and CONCLUSION. It is crucial that you adhere to this structure exactly as outlined and that the final answer in the CONCLUSION matches the standard correct answer precisely.

To explain further: In SUMMARY, briefly explain what steps you'll take to solve the problem. In CAP-TION, describe the contents of the image, specifically focusing on details relevant to the question. In REASONING, outline a step-by-step thought process you would use to solve the problem based on the image. In CONCLUSION, give the final answer in a direct format, and it must match the correct answer exactly. If it's a multiple choice question, the conclusion should only include the option without repeating what the option is.

Here's how the format should look:

<SUMMARY>[Summarize how you will approach the problem and explain the steps you will take to reach the answer.] </SUMMARY>

<CAPTION>[Provide a detailed description of the image, particularly emphasizing the aspects related to the question.] </CAPTION>

<REASONING>[Provide a chain-of-thought, logical explanation of the problem. This should outline step-by-step reasoning.] </REASONING>

<CONCLUSION>[State the final answer in a clear and direct format. It must match the correct answer exactly.] </CONCLUSION>(Do not forget </CONCLUSION>!)

Please apply this format meticulously to analyze the given image and answer the related question, ensuring that the answer matches the standard one perfectly.

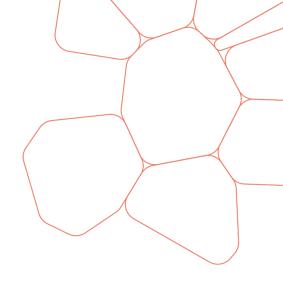
Recap: Data Verification Prompt

Evaluate whether the assistant's response is valid. Respond with 'valid' if the assistant's response is not a refusal and it aligns with the standard answer in meaning. Respond with 'invalid' if the response is a refusal or differs from the standard answer in a meaningful way.

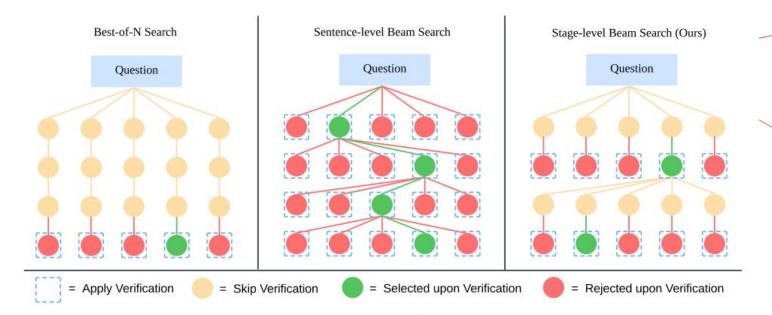
A refusal means the assistant states it cannot recognize a specific person/object or refuses to answer the question. Do not consider a response to be a refusal just because it includes the word 'no' or other negative terms.

Standard answer: {standard_answer}

Assistant's response: {assistant_response}



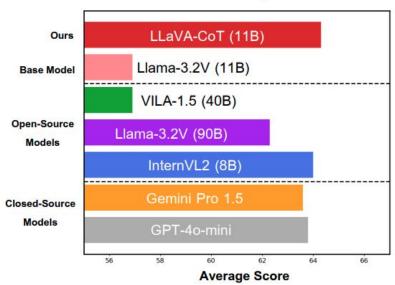
Recap: Inference time Scaling

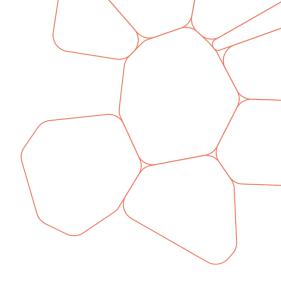


Now you act as a judge, helping me determine which of the two texts I provide better provides a summary/caption/reasoning/conclusion to solve the question.

Recap:Model Performance

6 Multimodal Reasoning Benchmarks





Recap: Model In Action

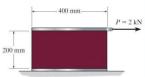
Example 1: Reasoning Problems



Ouestion: Subtract all tiny shiny balls. Subtract all purple objects. How many objects are left?

Options: A. 4, B. 8, C. 2, D. 6

Example 2: Science Problems



Question:

An acetal polymer block is fixed to the rigid plates at mm horizontally when it is subjected to a horizontal force P=2 kN, determine the shear modulus of the polymer. The width of the block is 100 mm. Assume that the polymer is linearly elastic and use small angle analysis.

Options:

A: 15 MPa, B: 5 MPa, C: 20 MPa, D: 2 MPa

Answer from Llama-3.2-11B-Vision-Instruct:

To determine the correct answer, we need to follow the instructions provided.

The first instruction is to subtract all tiny shiny balls. Since there are three tiny shiny balls in the image, we would be left with no shiny balls after performing this action. The second instruction is to subtract all purple objects. However, upon closer inspection, it becomes apparent that there are no shiny purple objects present in the image. Therefore, we cannot subtract any purple objects.

is C. 2, as initially, there were 2 purple objects, but they are not to be subtracted according to the instruction.

Answer from Llama-3.2-11B-Vision-Instruct:

To determine the shear modulus of the polymer, we can use the Summary stage: What's the problem? What should I do? formula for shear modulus in a small angle analysis: $\G = \frac{F}{A \gamma}$

where \(G \) is the shear modulus, \(F \) is the applied force, \(A \) is the area of the block, and \(\gamma \) is the shear

Given that the top plate of the block displaces horizontally by 2 mm when subjected to a horizontal force of 2 kN, we can calculate the shear strain \(\gamma \). Since the displacement is horizontal and the original length of the block is 400 mm the shear strain \(\\gamma\\) is:

 $\Gamma = \frac{\Delta x}{L} = \frac{2 \ \text{mm}}{400}$ $\text{text{mm}} = 0.005\]$

Next, we calculate the area \((A\)) of the block. The width of the block is given as 100 mm, and assuming the height of th its top and bottom surfaces. If the top plate displaces 2 block is also 200 mm (as given in the image analysis), the area \(A \) is:

 $A = 100 \ \text{text\{mm\} \times 200 \ \text{text\{mm\}} = 20000 \ }$ $\text{text{mm}}^2\$

Now, we can substitute these values into the formula for shear modulus:

 $G = \frac{F}{A \gamma} = \frac{2000 \cdot \text{text}}{20000}$ $\text{text{mm}^2 \times 0.005} = \frac{2000}{100} = 20$ \text{MPa}\] Therefore, the correct answer is C: 20 MPa.

Answer from LLaVA-CoT:

Summary stage: What's the problem? What should I do?

<SUMMARY> I will examine the image to identify and count all the objects, focusing on the tiny shiny balls and purple objects. Then, I will subtract these objects from the total count and determine the number of remaining objects, </SUMMARY>

Caption stage: What can I know from the image?

<CAPTION> The image shows various 3D geometric shapes on a flat surface, including spheres, cubes, and cylinders in different colors. Notably, there are small shiny green spheres and a purple cylinder. </CAPTION>

Reasoning stage: How to solve the problem step-by-step?

<REASONING> First, I will count all the objects in the image: there are 10 objects in total. Next, I will identify and subtract the tiny shiny balls. There is one small shiny green sphere. Given that we have already removed all the tiny shiny balls and Then, I will identify and subtract the purple objects. There is one purple cylinder. there were no purple objects to begin with, the correct answer By subtracting the two objects (the shiny ball and the purple cylinder) from the total, I find that 8 objects remain. </REASONING>

Conclusion stage: What is the final answer?

<CONCLUSION> B </CONCLUSION> Answer from LLaVA-CoT:

SUMMARY>To solve the problem, I will determine the shear modulus of the acetal polymer block using the relationship between the applied force, displacement, and dimensions of the block, I will use the formula for shear modulus and calculate the result. </SUMMARY> Caption stage: What can I know from the image?

<CAPTION>The image shows a rectangular block of acetal polymer fixed between two rigid plates. The top plate moves horizontally by 2 mm when a force $(P = 2 \setminus kN)$ is applied. The block's dimensions are given as 400 mm in length, 200 mm in height, and 100 mm in width. </CAPTION>

Reasoning stage: How to solve the problem step-by-step?

<REASONING>The shear modulus \(G \) of the polymer can be calculated using the formula: \| G = \frac{F}{A \cdot \frac{\delta}{I}} \|

\(F \) is the applied force (2 kN),

- (A) is the area of the block's top surface (100 mm x 400 mm = 40,000 mm²),

\(\delta\\) is the horizontal displacement (2 mm).

\(1 \) is the height of the block (200 mm).

First, convert the force to Newtons: $\langle F = 2 \rangle$, $\text{text}\{kN\} = 2000 \rangle$, $\text{text}\{N\} \rangle$. Next, calculate the shear modulus:

\[G = \frac{2000}{40,000 \cdot \frac{2}{200}} \]

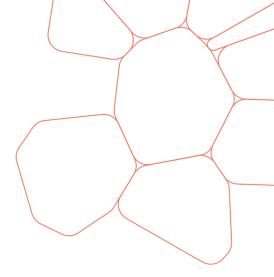
\[G = \frac{2000}{40,000 \cdot 0.01} \]

\[G = \frac{2000}{400} \]

 $\sqrt{G} = 5 \cdot \sqrt{\text{MPa}} \cdot \sqrt{\text{REASONING}}$

Conclusion stage: What is the final answer?

<CONCLUSION> B, 5 MPa </CONCLUSION>



Dataset

Model Demo



LLaVA-COT Code Repository VLMEvalKit LLaMa Finetuning

Challenge

- 1. Take a Small Domain VQA Dataset (~1k)
- 2. Generate the CoT Dataset using an open source VLM.
- 3. Fine-tune a VLM using created CoT Dataset
 - 4. Compare the results before and after fine tuning.
 - 5. Share Your Results in the Discord.

