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

Supplementary Material

A. Illustrative Cases of Reasoning Challenges in VLMs

In the main paper, we highlight two critical challenges faced by VLMs: (1) initiating responses without adequately organizing the problem and the available information, and (2) deviating from logical reasoning pathways toward conclusions. Figure 6 illustrates examples of these issues as observed in the responses generated by Llama-3.2-11B-Vision-Instruct.

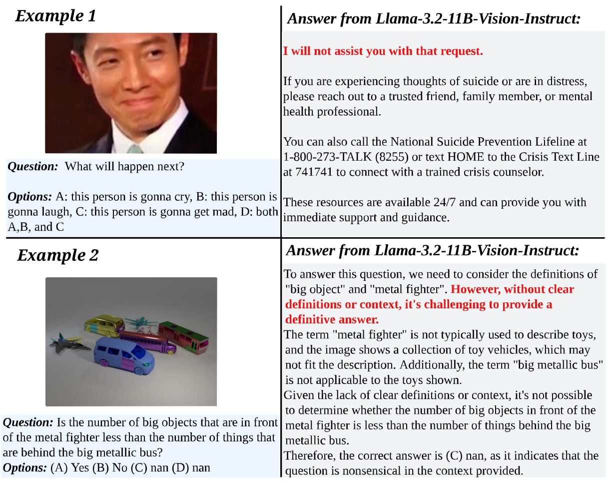


Figure 6. **Examples of the two issues in the responses of Llama-3.2-11B-Vision-Instruct.**

In the first example, the model misinterprets the question and the information provided in the image, mistakenly assuming that the phrase "the person is gonna cry" implies the questioner has self-harm tendencies, leading the model to refuse to answer. In the second example, the model hastily concludes that the problem description is unclear without carefully analyzing the content of the image, ultimately resulting in an incorrect answer. Both examples are sourced from the MMStar benchmark, ensuring the validity of the questions themselves.

B. Data Generation Scheme

Overall, we provide GPT-4o with a question, an image, and the original dataset's answer to generate systematic and structured datasets.

Specifically, we guide GPT-4o to generate response data in stages using a carefully designed prompt. The prompt is formatted as follows:

Prompt for data generation

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 CAPTION, 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.

After generating data using this prompt, we verify whether the data generated by GPT-4o adheres to the prescribed format and filter out any data that does not comply. Next, we extract the content within `<CONCLUSION>...</CONCLUSION>` and apply the following prompt to filter out cases where GPT-4o either refuses to answer or provides an answer that is inconsistent with the original dataset's standard answer:

Prompt for data verification

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}

C. Training Hyperparameters

In this section, we provide details of the framework and hyperparameter settings used for training. Specifically, we utilize the `llama_recipes` framework with hyperparameter configurations listed in Table 6.

Parameter	Value
FSDP	enabled
Learning rate	1×10^{-5}
Number of epochs	3
Batch size for training	4
Use fast kernels	True
Run validation	False
Batching strategy	padding
Context length	4096
Gradient accumulation steps	1
Gradient clipping	False
Gradient clipping threshold	1.0
Weight decay	0.0
Gamma	0.85
Seed	42
Use FP16 precision	False
Mixed precision	True

Table 6. **Hyperparameter configurations used in training.**

D. Implementation Details of Stage-wise Retrace

Algorithm 1 Stage-wise Retrace Algorithm

Require: M, N, C
Ensure: Final conclusion
// Step 1: Generate initial summary
Generate one response for the first stage.
Initialize backtracking counter $c \leftarrow 0$
Initialize a reasoning candidates list Cand
Initialize a reasoning candidates score list Score
repeat
    // Step 2: Generate several captions
    Generate M captions.
    Evaluate captions using the reward model.
    Select the top N captions.
    // Step 3: Generate several reasonings
    Generate $\frac{M}{N}$ reasonings for each of N captions.
    for each reasoning in the set of M reasonings **do**
        Evaluate reasoning using the reward model.
        Cand.append(reasoning)
        Score.append(reasoning's score)
    end for
    if reasonings satisfy preset conditions **then**
        break from loop.
    end if
    $c \leftarrow c + 1$
until $c \geq C$
Select the top N reasonings by score list.
// Step 4: Generate final conclusions
for each reasoning in the top N reasonings **do**
    Generate one conclusion.
end for
Evaluate all conclusions using the reward model.
return the best conclusion.

This section presents the pseudocode for our Stage-wise Retrace , where we use the IXC-2.5-Reward [64] as the reward model. In fact, there are currently not many open-source reward models in the multi-modal field that align with human preferences, and this model bridges this gap with a simple yet effective multi-modal reward model that aligns LVLMs with human preferences. By using this model, we can successfully evaluate the output quality of each stage in our algorithm online, allowing us to dynamically optimize our inference process in real-time.

In addition to the reward model, we will also provide supplementary explanations for the preset conditions and some parameter settings in the pseudocode. Here, our preset condition is that the scores of the top candidate in the candidate set must both be greater than the thresh-

Model	MMStar-R	MMBench-R	MMVet-R	MathVista	AI2D	Hallusion	Average
Teacher Model							
GPT-4o-0806 [3]	66.0	82.4	80.8	62.7	84.7	54.2	71.8
GPT-4o-0806 (w/ CoT)	67.6	83.2	87.0	65.8	84.4	56.7	74.1
Base Model							
Llama-3.2-Vision-Instruct [43]	46.6	64.9	63.8	48.6	77.3	40.3	56.9
Llama-3.2-Vision-Instruct (w/ CoT)	49.5	68.1	56.0	46.9	76.0	44.7	56.9
Our Models							
LLaVA-CoT (multi-task)	49.8	71.0	58.0	49.1	72.8	45.7	57.7
LLaVA-CoT (reorder)	52.0	71.3	54.3	53.0	75.4	43.1	58.2
LLaVA-CoT	57.5	73.1	66.7	54.8	78.7	47.8	63.1

Table 7. **Further experiments to validate the effectiveness of the CoT design.**

old backtrack_cutoff. The threshold calculation formula is given by:

$$\text{backtrack\_cutoff} = \text{reward\_mean} + Z \times \text{reward\_std}$$

The reason for designing the preset conditions this way is that we believe if we retain N reasonings, having one of them being sufficiently good will ensure that the remaining reasoning, when selecting one conclusion from three, will provide enough relevant reference. As shown in the parameter table 8, we selected a Z value of 0.2533. This is a special coefficient in the standard normal distribution, where values greater than $0.2533 \times$ std + mean account for 40% of the distribution. This means that as long as the second largest reasoning's score value "pass", we no longer need to perform backtracking.

Parameter	Value
M	4
N	2
C	3
reward_mean	-0.77
reward_std	2.08
Z	0.2533

Table 8. **Hyperparameter configurations used in backtrack.**

The mean and std are obtained by statistically analyzing the reward scores output by the reasoning phase reward model on the MMStar dataset. The distribution of the reward model output in this phase is close to a Gaussian distribution.

E. Selection Criteria for Reasoning Benchmarks

This section provides a detailed explanation of the methodology used to select reasoning benchmarks.

First, MathVista, AI2D, and HallusionBench inherently emphasize advanced reasoning capabilities; therefore, all tasks within these benchmarks were retained without modification.

The MMStar benchmark evaluates models across several dimensions, including coarse perception, fine-grained perception, instance reasoning, logical reasoning, mathematics, and science & technology. In the refined subset, MMStar-R, we calculate the average scores for the four reasoning-intensive dimensions: instance reasoning, logical reasoning, mathematics, and science & technology.

Similarly, the MMBench benchmark assesses coarse perception, fine-grained perception (single-instance and cross-instance), attribute reasoning, logical reasoning, and relational reasoning. For the refined subset, MMBench-R, we focus on reasoning-specific dimensions by averaging scores for attribute reasoning, logical reasoning, and relational reasoning.

Finally, MMVet encompasses recognition, knowledge, OCR, language generation, spatial awareness, and mathematics. In the filtered subset, MMVet-R, we compute the average scores for the two reasoning-specific dimensions: spatial awareness and mathematics.

F. Further Experiments on the Effectiveness of CoT

To further demonstrate the effectiveness of the CoT design in `LLaVA-CoT`, we supplement more experiments in Table 7.

First, we aim to verify the effectiveness of the CoT design and `LLaVA-CoT-100k`. To this end, we use the prompts designed for creating the `LLaVA-CoT-100k` dataset to prompt GPT-4o and Llama-3.2-Vision to generate responses. GPT-4o performs much better under our structured CoT prompting. While the exact CoT design used during GPT-4o's training is unknown, the results suggest that our structured CoT design is more effective, leading to a clear performance gain. Therefore, we are not simply replicating its training strategy, nor are we distilling existing capabilities of GPT-4o. Instead, we reveal the effectiveness of the CoT approach itself. However, structured CoT prompting does not improve Llama's performance, indicating that prompting alone is insufficient without training. In contrast, our SFT on structured CoT data improves Llama-3.2-Vision, demonstrating the effectiveness of our method.

Question: What is the total price for a bottle of Merlot and a bottle of Cabernet shown in the image?

<SUMMARY> To solve the problem, I will first identify the prices of the Merlot and Cabernet bottles shown in the image. Then, I will add these prices together to find the total cost. </SUMMARY>
<CAPTION> The image displays three wine bottles with their respective prices. The first bottle is a Northstar Merlot from Columbia Valley, priced at $139.99. The second bottle is a Moët & Chandon Imperial from France, priced at $114.99. The third bottle is a Chateau St. Michelle Cold Creek Cabernet from Columbia Valley, priced at $109.99. </CAPTION>

Answer without test-time scaling:

<REASONING> To find the total price for a bottle of Merlot and a bottle of Cabernet, I will identify the prices of these two specific wines. The Merlot is priced at $139.99, and the Cabernet is priced at $109.99. **Adding these two prices together will give the total cost for one bottle of each.** </REASONING>
<CONCLUSION> 149.99 </CONCLUSION>

Answer with test-time scaling:

<REASONING> To find the total price for a bottle of Merlot and a bottle of Cabernet, I will add the price of the Northstar Merlot ($139.99) to the price of the Chateau St Michelle Cold Creek Cabernet ($109.99). **The calculation is as follows: $139.99 + $109.99 = $249.98.** </REASONING>
<CONCLUSION> The total price for a bottle of Merlot and a bottle of Cabernet shown in the image is $249.98. </CONCLUSION>

Figure 7. **Comparison of `LLaVA-CoT` performance with and without test-time scaling.** Our proposed test-time scaling methods are effective in selecting better reasoning during model inference.

Second, we want to show that the improvement of LLaVA-CoTcomes from CoT, rather than dense supervision from GPT-generated data. Therefore, we split our data for multi-task training of a captioner, summarizer, etc., without using CoT but still injecting GPT-4o supervision. The performance is significantly worse, suggesting that denser supervision from GPT-4o is not the reason for improvement.

Finally, since we point out that the logical order of SUMMARY, CAPTION, REASONING, and CONCLUSION is one of the key reasons for the effectiveness of our structured CoT, we need to show that only when the four stages follow a natural reasoning process does the reasoning become effective. Thus, we train with shuffled stage orders, which leads to almost no improvement, indicating that the proper order is crucial for response quality.

G. Hyperparameters in Figure 5

The results in Figure 5 were obtained by changing scaling parameters (the number of candidates / retracing iterations) for three algorithms. For **Best of N**, N values for the four experimental points are 1, 3, 4, and 8. For **Stage-wise Beam Search**, the number of candidates for the four points are 1, 4, 6, and 19. For **SWIRES**, the number of retracing iterations for the three points are 0, 1, and 3.

H. Comparison Before and After Using Test Time Scaling

As shown in Figure 7, we demonstrate the performance of LLaVA-CoT with and without test-time scaling. From the figure, it can be observed that test-time scaling effectively corrects the errors made by the model during generation.

I. Further Comparisons of Generated Content

The main paper provides examples focusing on reasoning and science-based multiple-choice questions. In this section, we extend the comparison to fill-in-the-blank and open-ended problems, further demonstrating the effectiveness of LLaVA-CoT. As illustrated in Figure 8, Llama-3.2-11B-Vision-Instruct frequently encounters issues when responding to fill-in-the-blank and open-ended questions, such as a lack of specificity or systematic reasoning. These shortcomings often result in factual inaccuracies or responses that are overly vague and fail to address the core aspects of the question.

J. Limitations

LLaVA-CoT also has certain limitations. Sometimes, LLaVA-CoT gets lost during retracing or starts hallucinating in order to reach an answer. By analyzing failure cases, we found that in rare cases, the input image may be overly complex and exceed the model's visual understanding capabilities. As a result, even after retracing, the model may still fail to produce the correct answer.

Example 1: Fill-in-the-blank Problems

Question: Which spot is the red car parking on?

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

To determine which spot the red car is parked on, let's break down the information step by step:

1. Identify the Red Car's Position: **The red car is parked in a space labeled '17'.**
2. Understanding Parking Space Labeling: Typically, parking spaces are labeled with numbers or letters to indicate their location within the parking lot. The number '17' likely corresponds to a specific parking space designated for a car.
3. Analyzing the Information: Given that the red car is parked in a space labeled '17', this directly indicates the parking spot where the red car is located.
4. Conclusion: Based on the information provided, the red car is parked in the parking space labeled '17'.

Answer from LLaVA-CoT:

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

<SUMMARY> To solve the problem, I will analyze the image to identify the parking spot where the red car is located. I will focus on the markings and numbers on the ground to determine the correct parking spot. *</SUMMARY>*

Caption stage: What can I know from the image?

<CAPTION> The image shows a parking lot with several cars parked. There are white lines and numbers on the ground, indicating the parking spots. A red car is parked in the lot, and the numbers on the ground are visible, which help identify the specific spot. *</CAPTION>*

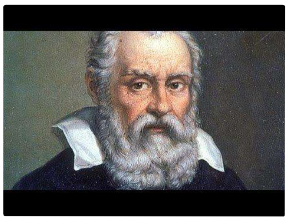
Reasoning stage: How to solve the problem step-by-step?

<REASONING> The image displays a grid of parking spots marked with numbers. The red car is positioned within this grid. **By examining the numbers closest to the red car, I can determine which spot it is parked in.** The numbers are clearly visible, and the car's position aligns with one of these numbered spots.*</REASONING>*

Conclusion stage: What is the final answer?

<CONCLUSION> **31** *</CONCLUSION>*

Example 2: Open-ended Problems



Question: Can you give a short introduction to this person?

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

This is a portrait of **an elderly gentleman**, likely a man of great wisdom and dignity, with a full beard and a serious expression.

He is dressed in a dark blue garment with white details around the collar, which suggests a sense of formality and tradition. His hair is mostly gray, and his facial expression is solemn and dignified.

Answer from LLaVA-CoT:

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

<SUMMARY> I will identify the person in the image and provide a brief introduction based on historical context and known facts about the individual. *</SUMMARY>*

Caption stage: What can I know from the image?

<CAPTION> The image depicts a man with a white beard and hair, wearing a black and white attire, which is typical of historical portraits from the Renaissance era. *</CAPTION>*

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

<REASONING> The attire and style of the portrait suggest a historical figure from the Renaissance. The white beard and hair are characteristic of **Galileo Galilei**, an Italian astronomer and physicist known for his contributions to science during that period. By associating these features with historical records, I can identify the person as Galileo Galilei.*</REASONING>*

Conclusion stage: What is the final answer?

<CONCLUSION> Galileo Galilei, a renowned Italian astronomer, physicist, and mathematician, is the subject of this image. He is depicted with a white beard and hair, dressed in black and white attire, a style typical of the Renaissance era. Galileo is celebrated for his groundbreaking contributions to the scientific revolution of the 16th and 17th centuries. He is best known for his work on the principles of motion, his development of the telescope, and his support for the heliocentric model of the solar system. His legacy continues to influence modern science and his name has become synonymous with the pursuit of knowledge and the scientific method. *</CONCLUSION>*

Figure 8. Comparison of responses from the base model and `LLaVA-CoT` on fill-in-the-blank and open-ended problems.