

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

Guowei Xu^{♣*} Peng Jin^{♣,♦*} Ziang Wu^{♣*} Hao Li[♣] Yibing Song[♣] Lichao Sun[■] Li Yuan^{♣,♦}

[♣]Shenzhen Graduate School, Peking University

[♣]Institute for Interdisciplinary Information Sciences, Tsinghua University

[♦]Rabbitpre AI & PKU Shenzhen AIGC Joint Lab [♣]DAMO Academy, Alibaba Group

[■]Computer Science and Engineering, Lehigh University

xgw23@mails.tsinghua.edu.cn yibingsong.cv@gmail.com yuanli-ece@pku.edu.cn

Abstract

Large language models have demonstrated substantial advancements in reasoning capabilities. However, current Vision-Language Models (VLMs) often struggle to perform systematic and structured reasoning, especially when handling complex visual question-answering tasks. In this work, we introduce LLaVA-CoT¹, a large VLM designed to conduct autonomous multistage reasoning. Unlike chain-of-thought prompting, LLaVA-CoT independently engages in sequential stages of summarization, visual interpretation, logical reasoning, and conclusion generation. This structured approach enables LLaVA-CoT to achieve marked improvements on reasoning-intensive tasks. To accomplish this, we construct the LLaVA-CoT-100k dataset, integrating samples from various visual question answering sources and providing structured reasoning annotations. Besides, we propose a test-time stage-wise retracing search method (SWIRES), which enables effective and efficient test-time scaling. Remarkably, with only 100k training samples and test-time scaling, LLaVA-CoT not only outperforms its base model by 9.4% on a wide range of multimodal reasoning benchmarks, but also surpasses the performance of larger and even closed-source models, such as Gemini-1.5-pro, GPT-4o-mini, and Llama-3.2-90B-Vision-Instruct. The code, dataset, and pre-trained weights are publicly available at <https://github.com/PKU-YuanGroup/LLaVA-CoT>.

1. Introduction

Large language models, represented by OpenAI o1 [68] and Deepseek R1 [15], demonstrate strong capabilities for systematic and in-depth reasoning, validating the effectiveness of test-time scaling for language models [51]. However, vision is equally important for enabling models to fully understand the world and extend their cognitive abilities [6, 65].

*G. Xu, P. Jin, and Z. Wu contribute equally.

¹Our LLaVA-CoT is built upon Llama-3.2-Vision model [43].

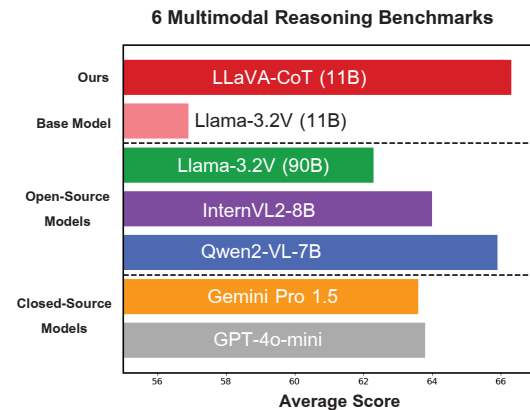


Figure 1. Performance of LLaVA-CoT and other models across six multimodal reasoning benchmarks. Although LLaVA-CoT is fine-tuned from the Llama-3.2-11B-Vision-Instruct [43] model (which has the lowest average score), it outperforms many larger open-source models and even some closed-source models. Detailed benchmark results are shown in Table 5.

Therefore, developing a multimodal model that integrates both language and vision while facilitating effective, systematic, and deep reasoning holds substantial significance.

Early open-source vision language models (VLMs) mainly employ a direct prediction approach, generating brief answers immediately in response to a question [26, 33, 35]. The main limitation of this direct-response paradigm is its lack of a structured reasoning process, making it less effective for tasks demanding logical reasoning [66]. Recent studies have shown that incorporating Chain-of-Thought (CoT) reasoning encourages the model to reason step by step, significantly improving its question-answering capabilities [56]. However, even with CoT reasoning, most VLMs frequently produce errors or hallucinated outputs during the reasoning progress [29, 34, 53].

Our findings suggest that a significant cause of these issues is the insufficiently systematic and structured nature of

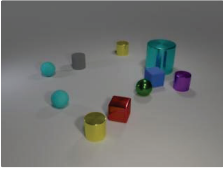
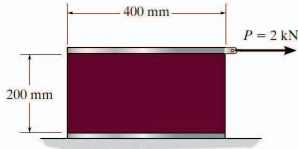
Example 1: Reasoning Problems	Answer from Llama-3.2-11B-Vision-Instruct:	Answer from LLaVA-CoT:
 Question: Subtract all tiny shiny balls. Subtract all purple objects. How many objects are left? Options: A. 4, B. 8, C. 2, D. 6	To determine the correct answer, we need to follow the instructions provided. 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. Given that we have already removed all the tiny shiny balls and there were no purple objects to begin with, the correct answer is C. 2, as initially, there were 2 purple objects, but they are not to be subtracted according to the instruction.	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. Then, I will identify and subtract the purple objects. There is one purple cylinder. 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>
Example 2: Science Problems	Answer from Llama-3.2-11B-Vision-Instruct:	Answer from LLaVA-CoT:
 Question: An acetel polymer block is fixed to the rigid plates at its top and bottom surfaces. If the top plate displaces 2 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	To determine the shear modulus of the polymer, we can use the formula for shear modulus in a small angle analysis: $G = \frac{F}{\Delta x} \frac{L}{A}$ where G is the shear modulus, F is the applied force, A is the area of the block, and Δx is the shear strain. 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 Δx. Since the displacement is horizontal and the original length of the block is 400 mm, the shear strain Δx is: $\Delta x = \frac{F}{L} \frac{A}{G} = \frac{2000 \text{ N}}{400 \text{ mm}} \frac{100 \text{ mm}^2}{G} = 0.005$ Next, we calculate the area A of the block. The width of the block is given as 100 mm, and assuming the height of the block is also 200 mm (as given in the image analysis), the area A is: $A = 100 \text{ mm} \times 200 \text{ mm} = 20000 \text{ mm}^2$ Now, we can substitute these values into the formula for shear modulus: $G = \frac{F}{\Delta x} \frac{L}{A} = \frac{2000 \text{ N}}{0.005} \frac{400 \text{ mm}}{20000 \text{ mm}^2} = 20 \text{ MPa}$ Therefore, the correct answer is C: 20 MPa.	Summary stage: What's the problem? What should I do? <SUMMARY> To solve the problem, I will determine the shear modulus of the acetel 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 acetel polymer fixed between two rigid plates. The top plate moves horizontally by 2 mm when a force $P = 2 \text{ kN}$ 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: $G = \frac{F}{\Delta x} \frac{L}{A}$ where: - F is the applied force (2 kN), - A is the area of the block's top surface ($100 \text{ mm} \times 400 \text{ mm} = 40,000 \text{ mm}^2$), - Δx is the horizontal displacement (2 mm), - L is the height of the block (200 mm). First, convert the force to Newtons: $F = 2 \text{ kN} = 2000 \text{ N}$. Next, calculate the shear modulus: $G = \frac{2000}{2} \frac{200}{40000} = 20 \text{ MPa}$ </REASONING> Conclusion stage: What is the final answer? <CONCLUSION> B. 5 MPa </CONCLUSION>

Figure 2. **Comparison of the base model and LLaVA-CoT.** As shown, the base model Llama-3.2-11B-Vision-Instruct exhibits obvious flaws in reasoning, with several errors occurring throughout the reasoning process. In contrast, LLaVA-CoT begins by outlining the problem, interprets relevant information from the image, proceeds with a step-by-step reasoning process, and ultimately reaches a well-supported conclusion. More examples of these comparisons can be found in Appendix F.

the reasoning process in existing VLMs. Specifically, by referring **systematic**, we mean that the model does not generate a direct reasoning chain but instead engages in multitasking reasoning. **Structured**, on the other hand, refers to the model’s ability to clearly identify the reasoning stage it is in and understand the primary task to be addressed at each stage. We observe that VLMs often initiate responses without adequately organizing the problem and the available information. Moreover, they frequently deviate from a logical reasoning toward conclusions, presenting a conclusion prematurely and subsequently attempting to justify it. Given that language models generate responses token-by-token, once an erroneous conclusion is introduced, the model typically continues along a flawed reasoning path. Examples of these issues can be found in Appendix A.

To mitigate these problems, we propose LLaVA-CoT. LLaVA-CoT undergoes supervised fine-tuning on structured training data and employs stage-wise retracing search (SWIRES) at test time. Specifically, our annotated training data includes four distinct stages: summary, caption, reasoning, and conclusion, enabling the model to systematically address questions in a multi-stage manner. Each stage serves a unique purpose in the reasoning process.

- **Summary:** A brief outline in which the model summa-

rizes the forthcoming task.

- **Caption:** A description of the relevant parts of an image, focusing on elements related to the question.
- **Reasoning:** A detailed analysis in which the model systematically considers the question.
- **Conclusion:** A concise summary of the answer, providing a final response based on the preceding reasoning.

To maintain clarity throughout the reasoning process, LLaVA-CoT marks each stage with a dedicated tag (e.g., <SUMMARY> . . . </SUMMARY>) to denote the beginning and end of each stage. To achieve these, we construct the LLaVA-CoT-100k dataset by generating responses using GPT-4o [3] and then train our model using supervised fine-tuning. After training, the model is capable of seamlessly transitioning between different stages without requiring any additional test-time intervention or prompting.

At test time, LLaVA-CoT can further enhance its reasoning capability through scaling. Unlike conventional scaling methods, LLaVA-CoT employs a stage-wise retracing search approach, which generates multiple candidate responses at each reasoning stage (e.g., summary, caption) and retains the most promising ones using a reward model. Moreover, if all candidate responses at a given stage are suboptimal, this suggests that the output from the previous

stage may have been inaccurate. In such cases, the model retraces to the preceding stage and attempts to regenerate its response. This retracing mechanism provides the model with an opportunity to revise its own answers, effectively improving error correction during the reasoning process.

We conduct experiments on several multimodal reasoning benchmarks, including MMStar [10], MMBench [36], MMVet [62], MathVista [38], AI2D [28], and HallusionBench [20], and observed that LLaVA-CoT offers two primary advantages: First, enabling the model to perform structured reasoning independently substantially outperforms traditional CoT prompting, particularly in complex reasoning tasks that require systematic analysis. Second, our stage-wise retracing search method is scalable and improves performance reliability, making it more effective in achieving stable and accurate results. Our contributions are summarized as follows:

- We introduce LLaVA-CoT, a visual language model designed for systematic reasoning, demonstrating outstanding performance on tasks that require structured thinking and reasoning.
- We demonstrate that LLaVA-CoT, using stage-wise retracing search, is test-time scalable. This means that with increased computational resources, the performance of our approach can be further enhanced, making it applicable to more complex scenarios.
- Extensive experiments on various benchmarks demonstrate that our method achieves superior performance relative to many larger and closed-source models, underscoring the effectiveness of LLaVA-CoT for multimodal reasoning.

2. Related Works

2.1. Visual reasoning with large language models

Visual reasoning demands a model’s visual perception capability and high-level cognition ability [27, 40]. Traditional vision-language models employ neural symbolic approaches [5, 12] to explicitly model the visual reasoning process. With the development of LLMs, vision-language models leverage the advanced reasoning abilities of LLMs to interpret visual tasks [35, 61]. Some vision-language models enhance visual reasoning by optimizing the visual encoding strategy [26, 32, 35] to produce cognition-focused visual tokens. VISPROG [21] positions the LLM as a decision-making agent, enhancing visual reasoning by invoking task-specific visual modules. Hu et al. [23] improves reasoning capabilities through sequential instruction tuning. Additionally, instructing learning techniques for language models, including prompt tuning [63], in-context learning, and supervised fine-tuning [50], also contribute to improvements in visual reasoning capabilities.

2.2. Chain-of-thought in large language models

Chain-of-thought prompting [56] offers a step-by-step reasoning trajectory when LLM faces hard questions including commonsense reasoning [17, 47], logical reasoning [30, 58], etc. Specifically, CoT prompting decomposes the question into a group of reasoning steps and builds a chain to guide the model to generate the results of complex problems step-by-step [13]. Recent works have demonstrated that CoT reasoning also substantially improves VLM’s capability on reasoning tasks. For instance, Prism [45] prompts LLMs by dividing the process into a perception stage and a reasoning stage. MSG [8] pioneers the use of forced Chain-of-Thoughts, establishing a new direction for structured prompting techniques. Distilling CoT [22] and Visual Program Distillation [24] distill CoT responses into VLMs. Visual CoT [49] enhances interpretability by generating bounding boxes for relevant regions alongside the answer. Compared to these approaches, LLaVA-CoT employs structured CoT to further enhance reasoning performance and experiments demonstrate that structured CoT outperforms direct CoT.

2.3. Test time scaling

Existing methods for test-time scaling primarily include majority voting [25], best-of- N search [4, 55], and beam search [19, 52]. Majority voting is effective for problems with standard answers but is not well-suited for open-ended tasks. Best-of- N search generates N complete responses and selects the best one; however, evaluating the accuracy of full responses can be challenging. Beam search generates multiple candidate sentences, selects the best ones, and iteratively refines the output. However, determining when to perform beam search is difficult to control. Previous works primarily conduct search after generating a fixed number of tokens or sentences [67], lacking an appropriate granularity. To address this, we propose stage-wise beam search, which performs search after the completion of an entire reasoning stage, ensuring that the search granularity aligns with semantic stages. Building on this, we further introduce stage-wise retracing search (SWIRES), which enhances the model’s ability to reflect and correct errors at test time.

3. Proposed Method

Our LLaVA-CoT facilitates a progressive, step-by-step reasoning process that enhances the reasoning capabilities of Vision-Language Models (VLMs) and allows for effective inference time scaling [51]. Using structured thinking, LLaVA-CoT achieves a systematic and efficient reasoning process. The proposed stage-wise retracing search method (SWIRES) enables it to outperform existing methods in test time scalability. This design ensures both robustness and accuracy in complex tasks requiring reasoning, which sep-

arates it from traditional approaches.

3.1. Enhancing Reasoning Capability through Structured Thinking

In this paper, our goal during training time is to develop a visual language model capable of engaging in systematic and in-depth reasoning.

3.1.1. Reasoning Stages

Our proposed model, LLaVA-CoT, decomposes the answer generation process into four structured reasoning stages:

- **Summary Stage.** In this initial phase, LLaVA-CoT provides a high-level summary interpretation of the question, outlining the primary aspects of the problem it intends to address.
- **Caption Stage.** LLaVA-CoT offers a concise overview of the visual elements relevant to the question, helping to understand multimodal input.
- **Reasoning Stage.** Building on the initial summary, LLaVA-CoT conducts structured, logical reasoning to derive a preliminary answer.
- **Conclusion Stage.** In this final stage, LLaVA-CoT synthesizes an answer based on the preceding reasoning. Here, the output from the conclusion stage is the direct response provided to the user, while the prior three stages are internal "hidden stages" representing LLaVA-CoT's reasoning process. The output at this stage adapts to the user's requirements: for instance, if the user requests a brief answer, the conclusion will be concise; if detailed explanations are desired, the conclusion provides a thorough, comprehensive response.

Each stage is initiated at the model's discretion, without any additional prompting, and all stages are completed by the model in a single inference pass. To achieve this, we provide the model with four pairs of special tags: `<SUMMARY></SUMMARY>`, `<CAPTION></CAPTION>`, `<REASONING></REASONING>`, and `<CONCLUSION></CONCLUSION>`. These tags correspond to summarizing the response approach, describing relevant image content, conducting reasoning, and preparing a final answer, respectively.

Upon training, LLaVA-CoT is capable of seamlessly transitioning between different stages without any external intervention, and we have not observed any instances where the model fails to adhere to the designated stage format. This structured approach enables the model to independently manage its reasoning process, improving its adaptability and performance on complex reasoning tasks.

3.1.2. Data Preparation and Model Training

Most existing VQA datasets lack detailed reasoning processes needed to train the LLaVA-CoT model. Therefore, we compile a new dataset, integrating samples from

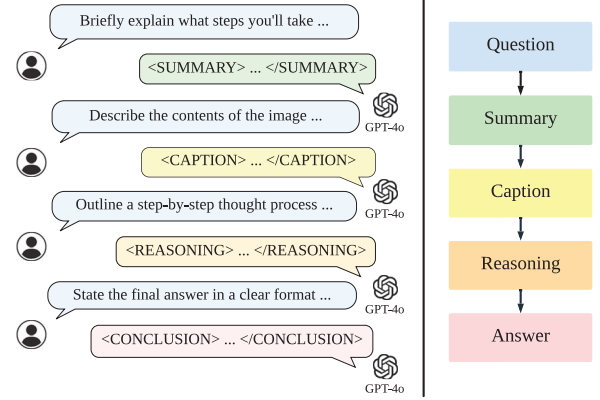


Figure 3. **Process flow for generating the LLaVA-CoT-100k dataset.** We prompt GPT-4o to generate responses in separate stages, and filter its outputs to ensure quality.

Dataset	Type	Size
ShareGPT4V [9]	General VQA	31.3k
ChartQA [41]	General VQA	17.2k
A-OKVQA [48]	General VQA	16.1k
AI2D [28]	Science-Targeted VQA	11.4k
GeoQA+ [7]	Science-Targeted VQA	11.4k
ScienceQA [37]	Science-Targeted VQA	5.6k
DocVQA [42]	General VQA	4.0k
PISC [31]	General VQA	1.0k
CLEVR [27]	General VQA	0.5k
CLEVR-Math [14]	Science-Targeted VQA	0.5k

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

several widely used VQA datasets, resulting in a total of 99k image QA pairs (each pair may include one or multiple rounds of questioning). As shown in Figure 3, since no multimodal model currently exists that can directly produce systematic, structured reasoning, we use GPT-4o [3] to generate detailed reasoning processes, including summary, caption, reasoning, and conclusion, and compile these into the LLaVA-CoT-100k dataset, which we plan to release for public use. Details of the generation process and examples of the generated data are provided in Appendix B. We include data from both general-purpose VQA datasets and science-targeted VQA datasets specified below:

General VQA Datasets. We include several general-purpose VQA datasets with distinct focuses. **ShareGPT4V** [9] provides multi-turn question-answering data from GPT-4V [59] interactions. **ChartQA** [41] focuses on interpreting charts and graphs. **A-OKVQA** [48] emphasizes external knowledge beyond visible content. **DocVQA** [42] involves document-based questions requiring textual comprehension. We also include **PISC** [31] to understand social relationships, and **CLEVR** [27] to address object properties, spatial relationships, and counting tasks.

Science-Targeted VQA Datasets. These datasets include **GeoQA+** [7] for geometric reasoning, along with **AI2D**

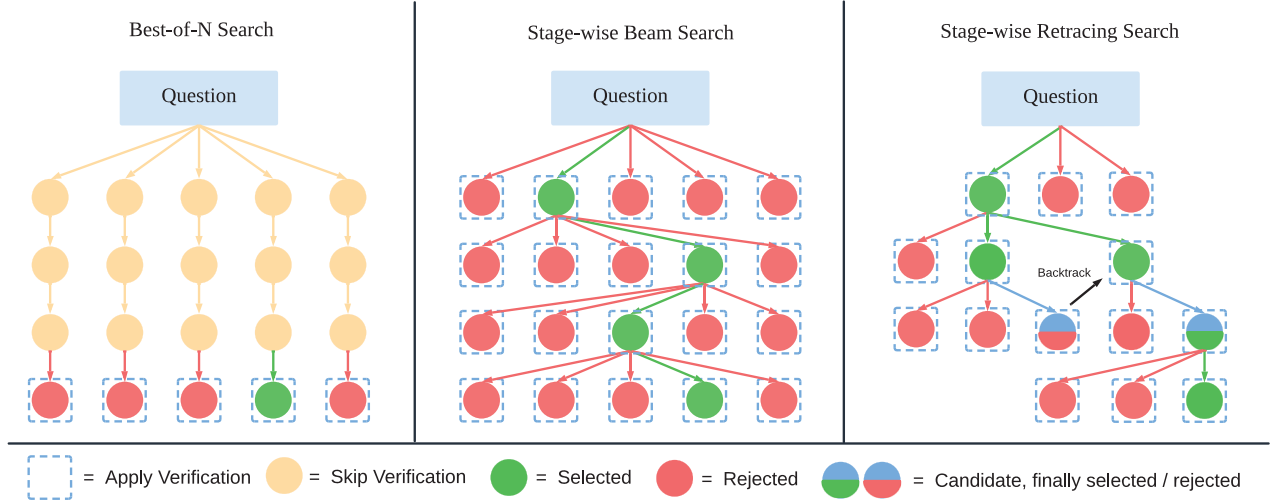


Figure 4. **An illustration of inference approaches.** Best-of-N search generates N complete responses and selects the best one among them. Stage-wise beam search generates M candidate responses for each reasoning stage and selects the top N to proceed to the next stage. Stage-wise retracing search, when generating responses for a given stage, retraces to the previous reasoning stage for regeneration if all candidate responses are of low quality, thereby enhancing the model’s self-reflection and error correction capabilities.

[28] and ScienceQA [37], which target scientific questions. CLEVR-Math [14], an extension of CLEVR, focuses on arithmetic analysis in visual contexts. Table 1 shows the number of QA pairs selected from each dataset.

Model Training. The LLaVA-CoT-100k dataset we construct can be used to further conduct Supervised Fine-Tuning (SFT) on any existing model to enhance reasoning capabilities. In this work, we select the Llama-3.2-11B-Vision-Instruct [43] model as the base model, and perform a full parameter fine-tuning by using the LLaVA-CoT-100k dataset. The training is conducted on a single node with 8 H100 GPUs. Details on the specific training parameters, including training epochs, learning rate, and optimization settings, are provided in Appendix C.

3.2. Test-Time Scaling via Stage-wise Retracing

After training, to further enhance the VLM’s reasoning ability during test time while also improving its error correction capability, we explore test-time scaling based on our model. Specifically, existing beam search methods improve upon best-of-N search but typically search at fixed intervals (i.e., after a predetermined number of sentences or tokens), which is not flexible to process visual questions at different complexity. To address this issue, we propose a stage-wise beam search method, which effectively solves the challenge of determining appropriate search step lengths. During our beam search, we incorporate a retracing mechanism that enhances the model’s overall performance and improves its error correction capability.

3.2.1. Stage-wise beam search

Stage-wise beam search is an improved version of traditional beam search, where the search step is set at the end of each reasoning stage (e.g., summary, caption). As shown in Figure 4, at each stage, the model generates M candidate responses and selects the top N based on a reward model. Each selected response then generates $\frac{M}{N}$ candidates in the next stage, and this process repeats. This approach ensures that the search step adapts to different task complexities and problem types, providing a flexible search granularity.

However, the stage-wise beam search has a critical problem. Since the reward model selects the highest-scoring response based only on the current stage, it may suffer from local optima or selection biases. For instance, if the caption stage produces a suboptimal response, refining the reasoning stage alone is insufficient to obtain accurate answers. To mitigate this, our model needs to retrace to a previous reasoning stage for self-reflection and error correction. To address this, we enhance stage-wise beam search and propose the stage-wise retracing search (SWIRES).

3.2.2. Stage-wise retracing search (SWIRES)

Our SWIRES design incorporates a retracing mechanism into the reasoning process. Its core procedure is shown as follows, with detailed implementation and hyperparameters provided in Appendix D.

- At each reasoning stage, generate M candidate responses.
- Check whether at least one of the generated responses surpasses a predefined reward threshold (see Appendix D for details on threshold settings).
- If at least one response exceeds the threshold, select the top N responses with the highest reward and proceed to

the next stage. Each of the N selected responses generates $\frac{M}{N}$ new candidates, maintaining M candidates in the next stage.

- If none of the responses exceed the threshold, it suggests that the previous stage’s output may not be optimal, leading to ineffective search in the current stage. In this case, the algorithm retains the current stage’s search results as candidates but retraces to the previous stage to regenerate new responses for the previous stage. These regenerated responses for the previous stage are then used to generate M new responses for the current stage. If any of these new responses surpass the threshold, the search for this stage terminates immediately, the model selects the top N among all candidates, and proceeds to the next stage. Otherwise, the generated responses are added to the candidate pool, and retracing continues (up to a maximum of C times).
- After completing the final reasoning stage, the answer with the highest reward among all generated responses is selected as the final answer.

Figure 4 intuitively illustrates the workflow of the SWIRES algorithm and its differences from other methods. Empirically, we observe that the summary stage typically produces high-quality outputs. Therefore, retracing search is applied starting from the caption stage.

4. Post-Training Performance

In this section, we compare LLaVA-CoT with the base model, Llama-3.2-11B-Vision-Instruct, on six commonly used multimodal benchmarks to demonstrate the effectiveness of our approach during the training phase. Following this comparison, we conduct ablation studies to evaluate the contribution of each component within our method.

4.1. Experimental Setup

We selected six widely used and challenging benchmarks for our experiments: MMStar [10], MMBench V1.1 [36], MMVet [62], MathVista [38], AI2D [28], and HallusionBench [20]. MMStar, MMBench, and MMVet primarily evaluate the general visual question-answering capabilities of models, while MathVista, and AI2D focus on models’ proficiency in mathematical and scientific reasoning. HallusionBench specifically assesses the models’ handling of language hallucinations and visual illusions. For MMBench, we use the V1.1 version of the test set, MathVista is evaluated using the testmini set, and the remaining datasets each have a single test set. To ensure fairness and reproducibility, all evaluations are conducted using VLMEvalKit [16], an open-source evaluation toolkit for large vision-language models. The performance metrics of all baseline models are derived from VLMEvalKit’s testing results [1].

4.2. Benchmark Results

We found that LLaVA-CoT achieves significant performance improvements, despite using only 100k data. According to Table 2, compared to the base model, Llama-3.2-11B-Vision-Instruct, LLaVA-CoT demonstrates notable improvements across general VQA, mathematical reasoning, scientific VQA, and hallucination control tasks, with an average benchmark score increase of **5.8%**, thereby validating the effectiveness of our approach.

4.3. Ablation Study

Effectiveness of LLaVA-CoT-100k Compared to Original Datasets. To demonstrate the effectiveness of our improved LLaVA-CoT-100k dataset, we present a comparison between LLaVA-CoT and the model trained on the original Q&A pairs across different benchmarks in Table 2. Although the model trained directly on the original Q&A pairs shows some overall improvement on the base model, its average performance remains significantly lower. In particular, on the MMVet benchmark, which requires more detailed responses, its performance is even worse than the base model. This result underscores the importance of the multi-stage format of our LLaVA-CoT-100k dataset for training models capable of advanced reasoning.

Structured Tags are Essential for Enhanced Performance. To examine whether the four tags we introduced improve the model’s performance, we compare LLaVA-CoT with the model trained on the LLaVA-CoT-100k dataset with structured tags removed. As shown in Table 2, our results show a significant drop in performance when the tags are removed, indicating that the structured tagging facilitates reasoning and improves model performance. To the best of our knowledge, LLaVA-CoT is the first attempt to successfully enhance a model’s reasoning ability and overall performance through a structured reasoning with tags.

Performance Gains Primarily in Reasoning-Intensive Areas. To analyze the specific areas in which LLaVA-CoT has improved compared to the base model, we conduct a detailed assessment of the model’s performance across different skills on the MMStar benchmark. MMStar is designed to evaluate six key capabilities: coarse perception, fine-grained perception, instance reasoning, logical reasoning, math, and science & technology. In Table 3, we compare the base model with LLaVA-CoT. Our analysis reveals that LLaVA-CoT demonstrates notable improvements in tasks requiring systematic reasoning, such as instance reasoning, logical reasoning, math, and science & technology, while showing relatively smaller gains in coarse perception and fine-grained perception. This suggests that our method can mainly improve reasoning capabilities of the model.

Model	MMStar	MMBench	MMVet	MathVista	AI2D	Hallusion	Average
Base Model							
Llama-3.2-11B-Vision-Instruct	49.8	65.8	57.6	48.6	77.3	40.3	56.6
Our Models							
LLaVA-CoT (with Direct Training)	54.3	76.2	49.9	49.5	81.2	42.9	59.0
LLaVA-CoT (w/o Structured Tags)	55.7	74.2	57.0	54.1	79.1	45.0	60.9
LLaVA-CoT	57.6	75.0	60.3	54.8	78.7	47.8	62.4

Table 2. **Experimental results of different models on the benchmark.** Here, LLaVA-CoT (with Direct Training) refers to the model trained directly on the original VQA dataset’s Q&A pairs, while LLaVA-CoT (w/o Structured Tags) represents the model trained on the LLaVA-CoT-100k dataset with the structured tags removed. LLaVA-CoT refers to the model trained on the complete LLaVA-CoT-100k dataset including the structured tags and use retracing during test time.

Model	CP	FP	IR	LR	Math	Science & Technology	Average
Base Model							
Llama-3.2-11B-Vision-Instruct	66.0	46.4	57.6	50.8	45.2	32.8	49.8
Our Models							
LLaVA-CoT (with Direct Training)	68.4	48.0	65.6	52.0	51.6	40.0	54.3
LLaVA-CoT (w/o Structured Tags)	68.4	48.0	60.0	55.2	64.4	38.0	55.7
LLaVA-CoT	68.8	46.8	63.2	58.0	64.0	44.8	57.6

Table 3. **Performance of different models on the MMStar benchmark across various skill areas.** Here, CP represents coarse perception, FP represents fine-grained perception, IR represents instance reasoning, and LR represents logical reasoning. As shown in the table, our model demonstrates substantial improvement over the base model in instance reasoning, logical reasoning, math, and science & technology, indicating that structured reasoning can significantly enhance the model’s reasoning capabilities.

5. Test Time Scaling

In this section, we aim to compare the effectiveness of our stage-wise retracing approach with best-of-N and stage-wise beam search shown in Figure 4 under comparable computational constraints. For all of the three approaches, we used InternLM-XComposer2.5-Reward [64] as the reward model to judge the quality of generation. In Figure 5, the scaling strategies of the three methods differ as follows: the retracing method scales by increasing the maximum number of backtracking iterations, the stage-wise beam search method scales by expanding the number of candidates, and the best-of-N method scales by increasing the value of N, i.e., the number of final response options available for selection.

5.1. Approximate-Scale Comparison Analysis

As shown in Figure 5, from a vertical perspective, under similar test conditions, our SWIRES method performs better than stage-wise beam search, which in turn outperforms best-of-N. The observed results can be attributed to the following reasons: best-of-N approach operates at a coarse level, handling entire responses at once. If there are mistakes in the middle of the process, they can carry forward and lead to wrong answers. Stage-wise beam search improves this by working at a finer level, adjusting each stage separately. Therefore it enhances reasoning quality and performs better than Best-of-N. However, errors may still remain in prior steps and further affect subsequent steps. The SWIRES method builds on stage-wise beam search by

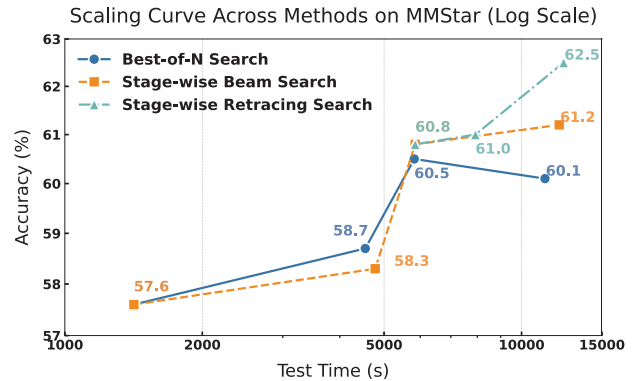


Figure 5. **Test time scaling curve on MMStar.** The experimental setup mirrors that used in the previous section, with evaluations conducted on MMStar using VLMEvalKit on a single A800 node. The time-axis we use here is in logarithmic scale.

adding a backtracking step, thus enhancing both the model performance.

5.2. Scaling Trend Analysis

As shown in Figure 5, compared to the baseline model’s accuracy of 57.6 on MMStar without any test-time scaling, all three methods exhibit the scaling trend. However, our SWIRES method demonstrates the strongest scaling effect as computation time increases. Specifically, both stage-wise beam search and best-of-N search plateau around 10,000 seconds, with best-of-N search even showing a slight decline beyond this point. In contrast, our SWIRES method continues to scale beyond this time scale, highlighting its su-

Model	MMStar	MMBench	MMVet	MathVista	AI2D	Hallusion	Average
Base Model							
Llama-3.2-11B-Vision-Instruct	49.8	65.8	57.6	48.6	77.3	40.3	56.6
Our Models							
LLaVA-CoT	57.6	75.0	60.3	54.8	78.7	47.8	62.4
LLaVA-CoT (w/ scaling)	62.5	77.6	64.9	57.7	81.0	49.1	65.5

Table 4. Experimental results during inference time. LLaVA-CoT (w/ scaling) denotes the model with stage-wise retracing.

Model	Size	MMStar-R	MMBench-R	MMVet-R	MathVista	AI2D	Hallusion	Average
Closed-Source Models								
GPT-4o-0806 [3]	–	66.0	82.4	80.8	62.7	84.7	54.2	71.8
GLM-4v-Plus [18]	–	68.8	78.7	78.1	68.8	85.1	55.6	72.5
Claude3.5-Sonnet-0620 [2]	–	64.2	75.4	68.7	61.6	80.2	49.9	66.7
Gemini-1.5-Pro [46]	–	56.4	71.5	71.3	57.7	79.1	45.6	63.6
GPT-4o-mini-0718 [44]	–	54.9	76.9	74.6	52.4	77.8	46.1	63.8
Larger Size Open-Source Models								
Llama-3.2-Vision-Instruct [43]	90B	51.1	76.8	74.1	58.3	69.5	44.1	62.3
VILA-1.5-40B [33]	40B	53.2	75.3	44.4	49.5	77.8	40.9	56.9
Deepseek-VL2[57]	MoE, 27B	62.3	80.2	60.3	63.9	83.8	45.3	66.0
Comparable Size Open-Source Models								
Qwen2-VL-7B [54]	8B	59.0	77.6	63.7	61.4	83.0	50.4	65.9
InternVL2-8B [11]	8B	62.5	77.4	56.9	58.3	83.6	45.0	64.0
Ovis1.5-Gemma2-9B [39]	11B	58.7	76.3	50.9	65.6	84.5	48.2	64.0
MiniCPM-V2.6 [60]	8B	57.1	75.7	56.3	60.6	82.1	48.1	63.3
Prism [45]	7B	41.0	59.2	47.6	35.7	65.7	40.5	48.3
VisCoT-7b-336 [49]	7B	28.2	50.7	17.1	24.7	40.7	28.8	31.7
Base Model								
Llama-3.2-Vision-Instruct [43]	11B	46.6	64.9	63.8	48.6	77.3	40.3	56.9
Our Models								
LLaVA-CoT	11B	57.5	73.1	66.7	54.8	78.7	47.8	63.1
LLaVA-CoT (w/ scaling)	11B	63.0	75.8	71.4	57.7	81.0	49.1	66.3

Table 5. Experimental results of LLaVA-CoT and state-of-the-art models on reasoning benchmarks.

perior scaling capability compared to the baseline methods. In Table 4, we compare the performance of LLaVA-CoT before and after applying the SWIRES method on benchmarks.

6. Comparison to State-of-the-Art VLMs

As shown in Table 5, we compare LLaVA-CoT with other state-of-the-art open-source and closed-source vision language models (VLM) across six benchmarks that require advanced reasoning capabilities. MMStar-R, MMBench-R, and MMVet-R are benchmarks derived from MMStar, MMBench V1.1, and MMVet, respectively, with tasks requiring only perception and OCR removed. These filtered benchmarks retain tasks that demand reasoning, with further details on the selection criteria in Appendix E. MathVista, AI2D, and HallusionBench inherently focus on reasoning, so we retained all tasks within these benchmarks.

Our results show that, despite our base model (Llama-3.2-11B-Vision-Instruct) being the weakest performer among the listed models, LLaVA-CoT consistently outperforms many open-source models of similar or even larger sizes, such as Qwen2-VL-7B [54], Deepseek-VL2 [57], and Llama-3.2-90B-Vision-Instruct [43], and VILA-1.5-

40B [33]. Remarkably, LLaVA-CoT even surpasses certain closed-source models like GPT-4o-mini [44] and Gemini-1.5-pro [46]. This comparison validates the advantages of our method, particularly in benchmarks that heavily depend on reasoning skills, and highlights LLaVA-CoT as a competitive model in the domain of VLM reasoning tasks.

7. Conclusion

In this paper, we present LLaVA-CoT, a vision language model that performs structured, autonomous reasoning in multiple stages. By introducing four distinct stages, LLaVA-CoT achieves a systematic reasoning process. Our contributions are twofold. First, the creation of the LLaVA-CoT-100k dataset with detailed reasoning annotations supports training on systematic, structured responses. Second, the proposal of a stage-wise retracing search method enables effective test time scaling. Overall, LLaVA-CoT presents a new approach to enhancing the reasoning capabilities of multimodal models. Future research could explore the application of reinforcement learning to further improve complex multimodal reasoning.

Acknowledgement

This work was supported in part by the Natural Science Foundation of China. (No. 62202014, 62332002, 62425101)

References

- [1] Detailed results on openvlm leaderboard. <https://opencompass.openxlab.space/assets/OpenVLM.json>. 6
- [2] Claude 3.5 sonnet, 2024. Available at: <https://www.anthropic.com/news/claude-3-5-sonnet>. 8
- [3] OpenAI (2024). Gpt-4o system card, 2024. 2, 4, 8, 3
- [4] Afra Amini, Tim Vieira, and Ryan Cotterell. Variational best-of-n alignment, 2024. 3
- [5] Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pages 279–290. Pmlr, 2020. 3
- [6] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook. *arXiv preprint arXiv:2307.13721*, 2023. 1
- [7] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 4
- [8] Franz Louis Cesista. Multimodal structured generation: Cvpr’s 2nd mmfm challenge technical report, 2024. 3
- [9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, 2024. 4
- [10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. 3, 6
- [11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 8
- [12] Minkyu Choi, Harsh Goel, Mohammad Omama, Y Yang, S Shah, and S Chinchali. Towards neuro-symbolic video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Milan, Italy, pages 9–13. Springer, 2024. 3
- [13] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, 2024. 3
- [14] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. In *International Joint Conference on Learning and Reasoning, 16th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy 2022)*, Windsor, UK, September 28-30, 2022, pages 155–170. Technical University of Aachen, 2022. 4, 5
- [15] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wan-jia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 1
- [16] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-

- source toolkit for evaluating large multi-modality models, 2024. 6
- [17] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. 3
- [18] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuntao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 8
- [19] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. 3
- [20] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, and et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385, 2024. 3, 6
- [21] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 3
- [22] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes, 2023. 3
- [23] Hanxu Hu, Simon Yu, Pinzhen Chen, and Edoardo M. Ponti. Fine-tuning large language models with sequential instructions, 2024. 3
- [24] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models, 2024. 3
- [25] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022. 3
- [26] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1, 3
- [27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4
- [28] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. pages 235–251, 2016. 3, 4, 5, 6
- [29] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamille Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. 1
- [30] Hao Li, Jinfa Huang, Peng Jin, Guoli Song, Qi Wu, and Jie Chen. Weakly-supervised 3d spatial reasoning for text-based visual question answering. *IEEE Transactions on Image Processing*, 32:3367–3382, 2023. 3
- [31] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. People in social context (pisc) dataset, 2017. Data set. 4
- [32] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. 3
- [33] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699, 2024. 1, 8
- [34] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingyu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning, 2023. 1
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023. 1, 3
- [36] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 3, 6
- [37] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 4, 5
- [38] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

- In International Conference on Learning Representations (ICLR), 2024. 3, 6
- [39] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. arXiv:2405.20797, 2024. 8
- [40] Mikołaj Mańkiński and Jacek Mańdziuk. A review of emerging research directions in abstract visual reasoning. Information Fusion, 91:713–736, 2023. 3
- [41] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Findings of the Association for Computational Linguistics: ACL 2022, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 4
- [42] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 2199–2208, 2021. 4
- [43] Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024. 1, 5, 8, 3
- [44] OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024. 8
- [45] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms, 2024. 3, 8
- [46] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 8
- [47] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. Commonsense reasoning for natural language processing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 27–33, 2020. 3
- [48] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge, 2022. 4
- [49] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning, 2024. 3, 8
- [50] Ming Shen. Rethinking data selection for supervised fine-tuning. arXiv preprint arXiv:2402.06094, 2024. 3
- [51] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. 1, 3
- [52] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2014. 3
- [53] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023. 1
- [54] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024. 8
- [55] Xiaofei Wang, Jinhua Li, and Yifan Zhang. Improved value alignment in large language models using variational best-of-n techniques, 2024. 3
- [56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022. 1, 3
- [57] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 8
- [58] Siheng Xiong, Yuan Yang, Ali Payani, James C Kerce, and Faramarz Fekri. Teilp: Time prediction over knowledge graphs via logical reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 16112–16119, 2024. 3
- [59] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v(ision), 2023. 4
- [60] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024. 8
- [61] Wangbo Yu, Chaoran Feng, Jiye Tang, Xu Jia, Li Yuan, and Yonghong Tian. Evagaussians: Event stream assisted gaussian splatting from blurry images. arXiv preprint arXiv:2405.20224, 2024. 3
- [62] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In International conference on machine learning. PMLR, 2024. 3, 6
- [63] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. Why johnny can’t prompt: how non-ai experts try (and fail) to design llm prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–21, 2023. 3
- [64] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan,

- Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. [arXiv preprint arXiv:2501.12368](#), 2025. [7](#), [2](#)
- [65] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019. [1](#)
- [66] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning, 2024. [1](#)
- [67] Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. Marco-o1: Towards open reasoning models for open-ended solutions, 2024. [3](#)
- [68] Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, and et al. Evaluation of openai o1: Opportunities and challenges of agi, 2024. [1](#)