LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs

Anonymous ACL submission

Abstract

Step-by-step reasoning is crucial for solving complex visual tasks, yet existing approaches lack a comprehensive framework for evaluating this capability and do not emphasize step-wise problem-solving. To this end, we propose a comprehensive framework for advancing multi-step visual reasoning in large multimodal models (LMMs) through three key contributions. *First*, we introduce a Visual Reasoning Chain Benchmark (VRC-Bench), a comprehensive benchmark for multi-step visual reasoning, covering eight diverse categories and over 4k verified reasoning steps to rigorously evaluate LLMs' ability to reason accurately and interpretably across multiple steps. *Second*, we propose a fine-grained visual reasoning metric that evaluates correctness and logical coherence at each step, providing deeper insights beyond traditional accuracy metrics. *Third*, we introduce LlamaV-o1, a state-of-the-art multimodal step-by-step reasoning model trained using a multi-step curriculum learning approach. LlamaV-o1 is optimized for structured step-by-step reasoning Our LlamaV-o1 obtains a significant gain of around 9% averaged across six benchmarks compared to the baseline, thereby demonstrating the impact of introducing the proposed step-by-step visual reasoning. Further, it outperforms the recent Llava-CoT with an absolute gain of 3.8% averaged across six benchmarks, while being 5× faster during inference scaling. On the VRC-Bench, LlamaV-o1 achieves the best performance among all open-source reasoning LMMs in terms of both final accuracy and steps. Our benchmark, model, and code will be publicly released.

1 Introduction

Large Language Models (LLMs) are designed to understand and generate text, enabling them to handle a wide range of tasks such as translation (Brown et al., 2020; Zhang et al., 2022; Chowdhery et al., 2023; Chen et al., 2025; Touvron et al., 2023a; Lu et al., 2024d), summarization (Goyal et al., 2022; Zhang et al., 2024b; Wang et al., 2023b), and question-answering (Touvron et al., 2023a; Le Scao et al., 2023; Touvron et al., 2023b). The integration of visual data through Large Multimodal Models (LMMs) extends capabilities further by combining text, images, or videos, allowing for more complex multimodal tasks like image captioning, visual question answering, and video analysis. To effectively solve these tasks, visual reasoning is essential for LMMs to process and connect diverse information, ensuring logical coherence and sequential problem-solving. The ability to reason across multiple modalities is crucial to addressing complex real-world problems.

Step-by-step reasoning helps LLMs break down complex tasks into manageable steps, improving accuracy and interpretability by ensuring logical consistency. This approach mirrors human cognition, enabling models to track their thought processes effectively. Prior works have shown that prompting or fine-tuning LLMs to generate step-wise rationales enhances reasoning performance (Lampinen et al., 2022; Wei et al., 2022; Wang et al., 2023a). However, existing methods struggle with multimodal step-by-step reasoning (see Fig. 1). Current visual reasoning benchmarks focus mainly on final-task accuracy, overlooking the quality of intermediate reasoning steps. Moreover, the absence of standardized evaluation further limits model comparisons in a fair setting.

To address these issues, we introduce Visual Reasoning-Chain (VRC-Bench), a comprehensive benchmark for multi-step visual reasoning. It covers eight diverse categories, including Visual Reasoning, Math & Logic, Social & Cultural Context, Medical Imaging, Charts & Diagrams, OCR & Documents, Complex Perception, and Scientific Reasoning. With over 1,000 curated samples and 4,173 manually verified reasoning steps, our VRC-Bench strives for accurate and reliable step-by-step

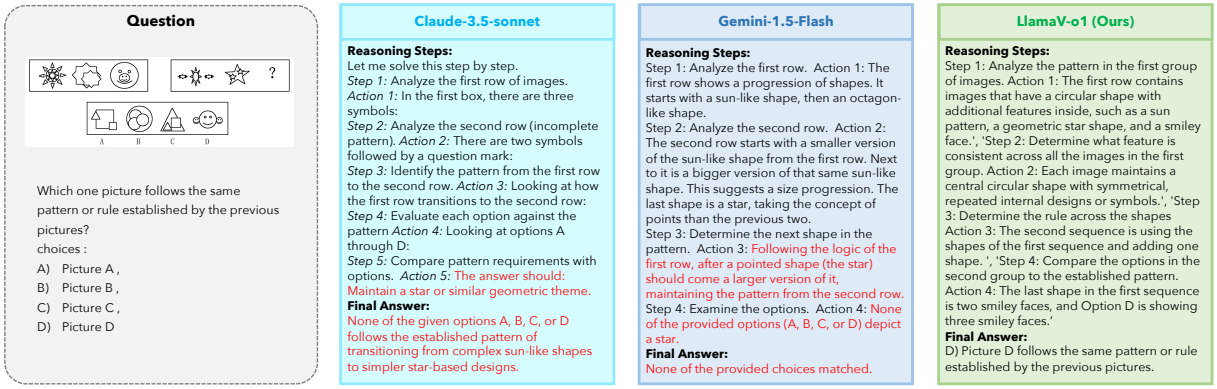


Figure 1: Comparison of step-by-step reasoning capabilities of our LlamaV-o1 with closed-source Gemini-1.5-Flash and Claude-3.5-Sonnet on a pattern recognition task from VRC-Bench. Claude-3.5-Sonnet incorrectly concludes "none of the options," with reasoning misalignments (in red). Gemini-1.5-Flash lacks logical coherence (in red). Our LlamaV-o1 provides systematic and accurate reasoning, correctly identifying option D as following the pattern. Additional results are presented in Appendix Fig. 5.

evaluation. Beyond end-task accuracy, we propose a new metric that assesses reasoning quality at each step, focusing on correctness and logical coherence. Lastly, we enhance model training by integrating Beam Search with Multi-Step Curriculum Learning, allowing models to gradually acquire skills—from basic summarization and question-driven captioning to complex multi-step reasoning. Our model, named LlamaV-o1, trained with this structured approach achieves state-of-the-art performance (see Fig. 1), surpassing existing open-source models across multiple evaluation metrics. In summary, our main contributions are as follows:

- Step-by-Step Visual Reasoning Benchmark: We introduce VRC-Bench, a comprehensive benchmark for multimodal multi-step reasoning. It spans eight diverse categories (e.g., Visual Reasoning, Math & Logic, Medical Imaging, Scientific Reasoning) with 1k challenging samples and 4k+ manually verified reasoning steps for robust evaluation.

- Novel Evaluation Metric: We propose a metric that assesses the reasoning quality at the level of individual steps, emphasizing both correctness and logical coherence.

- Combined Multi-Step Curriculum Learning and Beam Search Approach: We propose LlamaV-o1, a multimodal step-by-step reasoning model that integrates curriculum learning for structured skill acquisition with Beam Search for optimized reasoning paths. LlamaV-o1 outperforms the recent Llava-CoT (Xu et al., 2024a) with a 3.8% absolute gain across six benchmarks, while being 5× faster in inference.

2 Visual Reasoning Chain Benchmark

To facilitate a thorough assessment of the reasoning capabilities in complex scenarios, we introduce a step-by-step visual reasoning chain benchmark (VRC-Bench). VRC-Bench strives to assess both the logical progression of reasoning chains and the accuracy of the final outcomes generated by LMMs. VRC-Bench includes a diverse range of topics, such as science, mathematics, medical knowledge, social sciences, and data interpretation, ensuring that the proposed evaluation benchmark captures diverse aspects of visual reasoning.

2.1 Benchmark Creation

Benchmark Domains: To ensure a comprehensive assessment of reasoning capabilities, our step-by-step visual reasoning benchmark (VRC-Bench) incorporates samples from different datasets across various domains. Fig. 2 shows example questions and answers included in our benchmark. The data distribution is shown in Fig. 3. Based on diverse data samples, we generate step-by-step visual reasoning steps using a semi-automated annotation pipeline. Next, we outline the main domains covered in the benchmark and then present the annotation process.

Mathematical and Logical Reasoning: This category includes datasets focus on mathematical and logical tasks. MathVista (Lu et al., 2024a) provides a variety of mathematical problems, while DynaMath (Zou et al., 2024) offers dynamic mathematical challenges. Additionally, ChartQA (Masry

Figure 2: Our VRC-Bench covers diverse step-by-step visual reasoning tasks across multiple domains. It includes math, science, visual perception, art, medical imaging, and document understanding. Examples include angle calculation in geometry, molecular classification in chemistry, chart interpretation, artistic recognition, and medical diagnosis. Each task emphasizes logical inference, ensuring a comprehensive evaluation of multimodal reasoning.

et al., 2022) encompasses tasks related to chart and diagram comprehension, allowing evaluation of visual reasoning in logical contexts. *Scientific Reasoning*: For scientific reasoning, we collect samples from Science-QA (Lu et al., 2022) to test the model's ability to answer questions based on scientific knowledge and reasoning. Furthermore, MMMU-Medical (Yue et al., 2024), focuses on medical imaging tasks assessing the model's capability in interpreting complex multimodal medical data. *Cultural and Social Understanding*: To assess the model's ability to recognize and interpret diverse cultural scenarios, we include samples from ALM-Bench (Vayani et al., 2024), which is designed to assess understanding of the social and cultural context. *Other Visual Reasoning Scenarios*: We further include samples from other visual reasoning datasets. LogicVista (Xiao et al., 2024) and Blink-IQ (Fu et al., 2024) focus on complex visual perception, providing challenges that require the model to analyze and interpret intricate visual information. Doc-VQA (Mathew et al., 2020) targets OCR and document understanding, evaluating the model's ability to extract information from text-based documents. Lastly, MMMU (Yue et al., 2024) and BLINK (Fu et al., 2024) (Art Split) contribute to visual reasoning tasks.

Semi-Automatic Step-by-Step Reasoning Generation: We adopt a semi-automatic approach to generate step-by-step visual reasoning responses. We begin by using the GPT-4o model to create detailed reasoning steps and answers for the various questions in our dataset. This involves crafting specific prompts to guide the model in producing detailed logical reasoning. Additional details are presented in Appendix (Section. D.1).

Manual Verification: Since the aforementioned automated responses via GPT-4o are not always reliable, we perform manual verification to ensure that all reasoning steps are accurate and correct. In this stage, a team of verifiers carefully reviewed the generated reasoning chains and final answers, making necessary adjustments to enhance clarity and correctness. We ask the verifiers to add missing reasoning steps when necessary, and we drop examples with less than three reasoning steps after the verification except some samples from MathVista as they can be addressed with 2 steps. Over 25% of the data was corrected during this manual verification resulting in more than 1,000 samples and carefully verified 4,173 reasoning steps. The manual verification stage is essential for establishing a trustworthy ground-truth for the benchmark. Next, we discuss the evaluation framework.

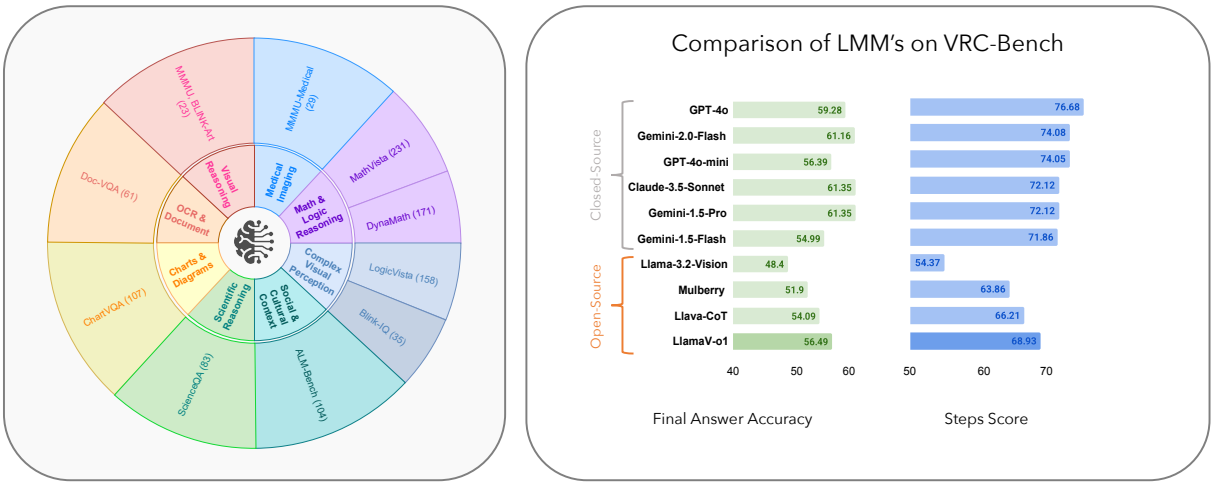


Figure 3: Overview of VRC-Bench and model performance comparison. **Left:** VRC-Bench spans multiple domains including, math, logic, science, visual perception, medical imaging, cultural understanding, OCR, and chart interpretation. It strives to evaluate LMMs on real-world multimodal reasoning scenarios. **Right:** A bar chart compares state-of-the-art models (GPT-4o, Gemini-2.0-Flash, Claude-3.5-Sonnet, Llava-CoT) on final answer accuracy and step-wise reasoning quality. Our LlamaV-o1 performs favorably against GPT-4o-mini, Gemini-1.5-Flash, and Llava-CoT, demonstrating superior accuracy and logical coherence.

2.2 Evaluation Framework

Previous methods for evaluating reasoning chains (Golovneva et al., 2023; Prasad et al., 2023) use reference-free approaches, offering flexibility but leading to inaccuracies. A minor error can disrupt the reasoning chain while still receiving a high score, failing to reflect true reasoning quality. We address this by incorporating ground-truth references, ensuring accurate and reliable evaluation.

Evaluation Metric: To improve reasoning evaluation, we use GPT-4o (, 2024) to compare model predictions against ground-truth. This allows us to evaluate reasoning quality using specific metrics that focus on different aspects of alignment and accuracy. Our metric builds on ROSCOE (Golovneva et al., 2023), introducing a reference-based approach. We assess reasoning quality using different measures (Appendix Table 5), including Faithfulness-Step (scoring alignment from 1 to 10) and Informativeness-Step (ensuring all critical information is included). Attributes like Hallucination and Redundancy help detect irrelevant or repetitive reasoning. The final score averages these factors for a comprehensive evaluation. Additional details including, the scoring system prompt are presented in Appendix (Section D.2).

3 Step-by-Step Reasoning LMM

Our proposed approach, named LlamaV-o1, aims at multimodal step-by-step reasoning in LMMs by combining curriculum learning with efficient inference. We train models progressively starting with simpler tasks like approach summarization and question-based captioning before advancing to detailed multi-step reasoning, thereby improving logical coherence and generalization. To optimize inference, we employ a parallel Beam Search strategy generating multiple reasoning paths and selecting the best one. This ensures high-quality outputs with lower computational costs, achieving constant scaling for greater efficiency compared to traditional methods.

3.1 Curriculum Learning for LMMs

LMMs excel at processing diverse data types like text, images, and video but often struggle with step-by-step reasoning. Curriculum learning addresses this challenge by training models progressively, starting with simpler tasks before advancing to complex reasoning. Inspired by human learning, this method has shown improvements in multimodal tasks like Visual Question Answering (VQA) (Kembhavi et al., 2017) and captioning (Johnson et al., 2016).

As discussed earlier, existing reasoning LMM works such as, LLava-CoT (Xu et al., 2024b) does not explicitly provide step-by-step reasoning outputs. Instead, it directly provides only the final answer in a structured format to mimic the reasoning. We note that directly introducing step-by-step reasoning into the LMM leads to inferior performance likely due to lack of foundational reasoning components (e.g., summary generation, image captioning

relevant to the input question etc.). In order to explicitly generate step-by-step reasoning in LMMs, we employ a curriculum learning-based strategy. To this end, we design a two-stage curriculum learning framework. In Stage 1, the model learns the foundational components of reasoning—first generating a summary of the approach to solve the problem, followed by an image caption relevant to the input question. These two steps in stage 1 help the model build structured understanding before engaging in complex reasoning. Since summarization and captioning are comparatively simpler tasks, this stage aid the model develop strong contextual grounding before moving to reasoning. In Stage 2, the model progresses to multi-step reasoning, where it performs all four interdependent steps: generate the summary, produce the image caption, construct reasoning steps using these components, and finally derive the correct answer. This stage reinforces logical coherence and reasoning consistency while emphasizing accurate predictions.

3.2 Multi-Step Chain-of-Thought Reasoning

As discussed above, our two-stage curriculum learning framework comprises multi-step (four) reasoning. Unlike single-step reasoning, multi-step chain-of-thought reasoning breaks problems into smaller manageable steps. This mirrors human problem-solving, where reasoning unfolds step by step (Kahneman, 2011; Prystawski et al., 2023). For instance, answering an image-related question likely involves identifying objects, understanding relationships, and synthesizing information. By integrating multi-step reasoning, we argue that LMMs become more interpretable and closer to human-like reasoning.

Our multi-step chain-of-thought (CoT) reasoning consists of the following steps. *Task Understanding:* The model begins by understanding the question and the context. *Task Summarization:* The next step involves generating a summary of the visual data to ensure the model has a holistic understanding. This stage prepares the model to focus on relevant action items to be taken to get the final answer. *Detailed Caption Generation:* To narrow the scope further, the model generates a detailed caption, which identifies specific labels and their corresponding values in the input image. This step ensures that the model accurately interprets the visual elements.*Logical Reasoning:* The model then formulates a logical reasoning process to locate and interpret the required data. This reasoning step breaks the task into sub-goals, ensuring a systematic approach. *Final Answer Generation:* Finally, the model outputs the final answer based on the reasoning process and the extracted context.

3.2.1 Data Preparation and Model Training

To implement our curriculum learning strategy effectively, we divide the model training process into two stages, each designed to incrementally enhance the model's reasoning capabilities while ensuring a robust understanding of multimodal inputs. This structured approach allows the model to acquire foundational reasoning skills in the first stage and progressively refines its ability to provide detailed, step-by-step answers in the second stage.

Training Stage 1: Summarization and Caption Generation: In the first stage, the model learns two key tasks: (1) summarizing the approach needed to answer a question and (2) generating a detailed caption highlighting relevant aspects of the input (e.g., visual elements in an image). Training data is derived from 18K Cap-QA samples from PixMo (Deitke et al., 2024) and 57K Geo170K samples from G-LLaVA (Gao et al., 2023), ensuring exposure to grounded captions and reasoning steps. This stage helps the model contextualize input data and outline a structured reasoning plan before solving multi-step tasks in the curriculum learning manner.

Training Stage 2: Detailed Reasoning and Final Answer Generation: In the second stage, the model builds upon the foundation established in Stage 1. Here, the model is trained not only to generate the summary and caption but also to provide detailed reasoning followed by final answer based on these components. Training data comprises of 99K structured samples from Llava-CoT (Xu et al., 2024a), covering diverse domains like General VQA (e.g., ShareGPT4V (Chen et al., 2025), ChartQA (Masry et al., 2022), A-OKVQA (Schwenk et al., 2022), DocVQA (Mathew et al., 2021), PISC (Junnan et al., 2017), CLEVR (Johnson et al., 2017)) and Science-Targeted VQA (e.g., GeoQA+ (Cao and Xiao, 2022), AI2D (Kembhavi et al., 2016a), ScienceQA (Lu et al., 2022) and CLEVR-Math (Lindström and Abraham, 2022)). Each sample includes a summary, caption, detailed reasoning, and final answer to form a structured learning path. The model is trained using curriculum learning, where it progressively develops reasoning skills in two stages. In Stage 1, the model focuses on understanding the problem structure and generating con-

textual descriptions through summaries and captions. In Stage 2, it engages in multi-step reasoning, where each step builds on the previous one—using the summary and caption to generate reasoning steps, and then leveraging those reasoning steps to derive the final answer. This incremental learning approach systematically integrates information for structured, step-by-step reasoning.

3.2.2 Optimizing Inference Efficiency

Efficient inference is crucial for real-world deployment of large multimodal models. To improve both speed and reasoning quality we adopt Beam Search, which helps to enhance inference efficiency along with high output quality (Meister et al., 2020).

Simplified Output Design: Unlike LLava-CoT (Xu et al., 2024b), our approach does not require a highly structured output format. This flexibility simplifies the reasoning process, allowing the model to focus on generating high-quality outputs without the overhead of rigid structural constraints. This design choice makes our method more adaptable to a wide range of reasoning scenarios, improving generalization across tasks.

Improved Efficiency with Beam Search: The Beam Search technique allows us to generate multiple reasoning paths in parallel and select the most optimal one. This approach enhances both the quality and consistency of the model's outputs. By evaluating multiple candidates and selecting the best, we ensure that the final answer is logical and robust. Our approach also achieves significant computational efficiency with $O(1)$ inference time scaling, making it more scalable than LLava-CoT's $O(n)$ complexity in terms of model calls for larger datasets and complex reasoning tasks.

4 Experiments

In this section, we extensively evaluate our model performance trained with curriculum learning. We employ Llama-3.2-11B-Vision-Instruct (Meta AI, 2024) as the baseline. Training is conducted on subset of PixMo, G-LLaVA and LLaVA-CoT-100k, allowing a structured progression from basic summarization to complex multi-step reasoning. We evaluate the performance on the proposed reasoning benchmark (VRC-Bench), designed for multi-step chain-of-thought evaluation in multimodal contexts. Additionally, we present performance comparison on six multimodal benchmarks from LLaVA-CoT, covering visual, mathematical, and scientific visual reasoning.

4.1 Experimental Setup

We fine-tune the baseline Llama-3.2-11B-Vision-Instruct (Meta AI, 2024) using llama-recipes framework with Supervised Fine-Tuning (SFT). Starting with simpler tasks on a subset of PixMo and G-LLaVA in Stage 1, where model learns foundational reasoning skills such as approach summary and caption. In Stage 2, training progresses to more complex LLaVA-CoT-100k dataset. We use Llama-3.2-11B-Vision-Instruct (Meta AI, 2024) as the base model for its strong multimodal reasoning capabilities. The model undergoes full-parameter optimization. Training is conducted on 8 NVIDIA A100 (80GB) GPUs. Additional training details are presented in the Appendix (Section D).

We evaluate our model on the proposed reasoning benchmark (VRC-Bench) and six established multimodal benchmarks: MMStar (Chen et al., 2024a), MMBench (Liu et al., 2025), MMVet (Yu et al., 2023), MathVista (Lu et al., 2024b), AI2D (Kembhavi et al., 2016b), and Hallusion (Guan et al., 2024). These benchmarks assess visual question answering, mathematical and scientific reasoning, and handling hallucinations and visual illusions. For step-by-step reasoning evaluation, we use a fuzzy evaluation strategy with GPT-4o as the judge, ensuring robust assessments. To ensure fair and reproducible performance comparison, we adopt the VLMEvalKit (Duan et al., 2024), as used in the LLaVA-CoT.

4.2 Results

Our model demonstrates significant improvements over existing methods on our proposed reasoning benchmark, as shown in Table 1. The evaluation compares final answer accuracy and step-by-step reasoning performance with state-of-the-art models. Models like GPT-4o (, 2024), Claude-3.5-Sonnet (cla, 2024), Gemini-2.0-Flash and Gemini-1.5-Pro (Reid et al., 2024) exhibit strong reasoning capabilities. Our approach achieves better final answer accuracy (56.49) compared to GPT-4o-mini (OpenAI, 2024) and other open-source models, such as Llama-3.2-Vision (Meta AI, 2024), Mulberry (Yao et al., 2024a) and LLava-CoT (Xu et al., 2024a) as well as competitive step scores (68.93%). This highlights the model's ability to generate accurate outputs while maintaining logical coherence in multi-step tasks.

Table 2 summarizes the performance comparison on six established benchmarks: MMStar,

Table 1: Comparison of models based on Final Answer accuracy and Reasoning Steps performance on the proposed VRC-Bench. The best results in each case (closed-source and open-source) are in bold. Our LlamaV-o1 achieves superior performance compared to its open-source counterpart (Llava-CoT) while also being competitive against the closed-source models.

	Close-Source						*Open-Source*			
Model	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-2.0 Flash**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-4o mini**	**Llama-3.2 Vision**	**Mulberry**	**Llava-CoT**	**LlamaV-o1** (Ours)
Final Answer	59.28	**61.35**	61.16	**61.35**	54.99	56.39	48.40	51.90	54.09	**56.49**
Steps	**76.68**	72.12	74.08	72.12	71.86	74.05	58.37	63.86	66.21	**68.93**

Table 2: Performance comparison on six benchmarks (MMStar (Chen et al., 2024a), MMBench (Liu et al., 2025), MMVet (Yu et al., 2023), MathVista (Lu et al., 2024b), AI2D (Kembhavi et al., 2016b), Hallusion (Guan et al., 2024)) including average scores. GPT-4o leads among closed-source models (71.8%), while our LlamaV-o1 achieves the best open-source performance (67.33%), surpassing Llava-CoT by 3.8%.

Model	**MMStar**	**MMBench**	**MMVet**	**MathVista**	**AI2D**	**Hallusion**	**Average**
Close-Source							
GPT-4o-0806 (, 2024)	66.0	82.4	80.8	62.7	84.7	54.2	71.8
Claude3.5-Sonnet-0620 (cla, 2024)	64.2	75.4	68.7	61.6	80.2	49.9	66.7
Gemini-1.5-Pro (Reid et al., 2024)	56.4	71.5	71.3	57.7	79.1	45.6	63.6
GPT-4o-mini-0718 (OpenAI, 2024)	54.9	76.9	74.6	52.4	77.8	46.1	63.8
Open-Source							
InternVL2-8B (Chen et al., 2024c)	62.50	77.40	56.90	58.30	83.60	45.00	64.00
Ovis1.5-Gemma2-9B (Lu et al., 2024c)	58.70	76.30	50.90	65.60	84.50	48.20	64.00
MiniCPM-V2.6-8B (Yao et al., 2024c)	57.10	75.70	56.30	60.60	82.10	48.10	63.30
Llama-3.2-90B-Vision-Inst (Meta AI, 2024)	51.10	76.80	74.10	58.30	69.50	44.10	62.30
VILA-1.5-40B (Liu et al., 2024)	53.20	75.30	44.40	49.50	77.80	40.90	56.90
Mulberry-7B (Yao et al., 2024a)	61.30	75.34	43.90	57.49	78.95	54.10	62.78
Llava-CoT (Xu et al., 2024a)	57.60	75.00	60.30	54.80	85.70	47.80	63.50
Our Models							
Llama-3.2-11B-Vision-Inst (Meta AI, 2024)	49.80	65.80	57.60	48.60	77.30	40.30	56.90
LlamaV-o1 (Ours)	59.53	79.89	65.40	54.40	81.24	63.51	67.33

MMBench, MMVet, MathVista, AI2D, and HallusionBench. Among open-source models, our method achieves the highest average score of 67.33%, surpassing recent models like LLaVA-CoT (63.50). Notably, our model demonstrates significant strengths in reasoning-intensive benchmarks, such as MMVet (65.40%) and Hallusion (63.51%). These results demonstrate the effectiveness of our model in handling diverse and complex multimodal tasks. Additional results are presented in the Appendix (Sec. C).

4.3 Ablations

Impact of Proposed Components: Table 3 showcases the impact of our proposed components of LlamaV-o1 on improving performance in complex visual reasoning tasks across six multimodal benchmarks: MMStar, MMBench, MMVet, MathVista, AI2D, and Hallusion. Starting with a curriculum learning strategy combined with multi-step Chain-of-Thought (CoT) reasoning, the model achieves an average score of 66.08%, demonstrating its ability to handle reasoning-intensive tasks effectively. By incorporating Beam Search, which optimizes the selection of reasoning paths, the performance further improves, achieving the highest average score of 67.33%. This improvement is particularly significant in benchmarks, such as MMVet (65.40% vs. 61.88%), MMStar (59.53% vs. 58.13%), and AI2D (81.24% vs. 80.18%), which evaluate the model's logical, visual, and contextual reasoning abilities. These results highlight the effectiveness of combining progressive training with optimized inference, enabling the model to generalize better across complex tasks and consistently deliver accurate and coherent reasoning.

Effectiveness of Inference Scaling Techniques: Table 4 presents the comparison of the efficiency and effectiveness of inference scaling techniques on the MMVet benchmark. We compare the newly introduced stage-level beam search used in Llava-CoT with Beam Search in our proposed approach. Both approaches are evaluated based on MMVet scores and inference time, measured on a single NVIDIA A100 GPU (80GB). **Stage-Level Beam Search (Llava-CoT):** Increasing the number of beams improves the MMVet score incrementally (from 60.3% with 1 beam to 62.9% with 4 beams). However, this improvement comes at a significantly higher computational cost due to linear scaling

Table 3: Impact of our contributions on multimodal reasoning across six benchmarks (MMStar, MMBench, MMVet, MathVista, AI2D, Hallusion). Curriculum Learning with Multi-Step CoT improves performance by 9.14% over Llama-3.2-11B-Vision-Inst (Meta AI, 2024), enhancing complex reasoning. Adding Beam Search further boosts accuracy, particularly on MMVet (65.40% vs. 61.88%), MathVista (54.40% vs. 53.20%), and AI2D (81.24% vs. 80.18%). Our final approach, combining curriculum learning and optimized inference, achieves a 10.43% overall improvement over the baseline.

Model	MMStar	MMBench	MMVet	MathVista	AI2D	Hallusion	Average
Llama-3.2-11B-Vision-Inst (*baseline*)	49.80	65.80	57.60	48.60	77.30	40.30	56.90
+ Curriculum with Multi-Step CoT Reasoning	58.13	79.55	61.88	53.20	80.18	63.31	66.04
+ Beam Search	59.53	79.89	65.40	54.40	81.24	63.51	67.33

Table 4: Inference scaling comparison on MMVet using a single NVIDIA A100 GPU. **Left:** Llava-CoT with stage-level Beam Search improves MMVet scores but suffers from quadratic scaling, increasing inference time. **Right:** Our Beam Search approach achieves higher accuracy with significantly lower inference time due to linear scaling efficiency. For example, our method scores 65.40 with four beams in 6.1 GPU hours, whereas Llava-CoT scores 62.9 but requires 54.1 GPU hours, demonstrating superior efficiency for real-world applications.

Inference Scaling	# Beams	MMVet Score	Time (GPU Hours)
No Scaling	1	60.3	3.8
Stage-level	2	61.7	20.1
Stage-level	3	62.3	38.5
Stage-level	4	62.9	54.1

Inference Scaling	# Beams	MMVet Score	Time (GPU Hours)
No Scaling	1	63.63	2.7
Beam Search	2	64.26	4.8
Beam Search	3	64.92	5.7
Beam Search	4	65.40	6.1

(time complexity of $O(n)$) based on model calls, with inference time rising from 3.8 GPU hours for 1 beam to 54.1 GPU hours for 4 beams. This scaling inefficiency limits the practicality of the stage-level approach for real-world applications. **Beam Search (Ours):** In comparison, our method achieves significantly better MMVet scores while maintaining a constant scaling (time complexity of $O(1)$) of inference time in terms of model calls. With 1 beam, our model already outperforms Llava-CoT (63.63% vs. 60.3%). As the number of beams increases, the MMVet score improves further, reaching 65.40 with 4 beams in just 6.1 GPU hours, a fraction of the computational cost of Llava-CoT. This demonstrates that Beam Search is not only more efficient with higher accuracy but also suitable for real-world applications.

5 Conclusion

In this paper, we propose a comprehensive approach for advancing multimodal step-by-step reasoning by introducing a new benchmark, a novel metric, and a step-by-step visual reasoning model trained using curriculum learning. The proposed VRC-Bench comprises eight diverse categories with 1k samples and more than 4k manually verified reasoning steps. The proposed evaluation metric strives to evaluate the reasoning quality at the individual step level by emphasizing both logical coherence and correctness. The proposed LlamaV-o1 model demonstrates significant improvements over existing methods, achieving state-of-the-art performance on challenging benchmarks while maintaining efficiency in inference. The incorporation of curriculum learning enables the model to develop foundational reasoning skills progressively, resulting in improved generalization and robustness across diverse tasks. Our results highlight the effectiveness of our design choices, including the structured training strategy, efficient inference mechanism, and rigorous evaluation using both on the proposed benchmark as well as widely recognized multiple datasets.

6 Limitations and Future Direction

While our approach significantly enhances multi-step visual reasoning capabilities in LMMs, there are different areas for further improvements. As discussed earlier, our proposed benchmark (VRC-Bench) primarily focuses on structured reasoning tasks. A potential future research direction is to further expand it to cover more diverse real-world scenarios, such as open-ended visual narratives and interactive reasoning. Furthermore, while Beam Search improves inference efficiency, exploring adaptive decoding strategies might further optimize reasoning speed with little or no compromise on model accuracy. Another potential future research direction is integrating reinforcement learning or self-improving mechanisms to enable the model to learn from its own reasoning errors, fostering continuous improvement.

References

2024. Claude 3.5 sonnet. Available at: https://www.anthropic.com/news/claude-3-5-sonnet.

OpenAI (2024). 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. 2020. Neuro-symbolic visual reasoning: Disentangling. In *International Conference on Machine Learning*, pages 279–290. Pmlr.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jie Cao and Jing Xiao. 2022. 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.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M^{3}cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. 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*, pages 24185–24198.

Ting-Rui Chiang and Yun-Nung Chen. 2019. Semantically-aligned equation generation for solving and reasoning math word problems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2656–2668, Minneapolis, Minnesota. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

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. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *Preprint*, arXiv:2407.11691.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Artur d'Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. 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.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. 2023. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Li Junnan, Wong Yong Kang, Zhao Qi, and Mohan S Kankanhalli. 2017. People in social context (pisc) dataset.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016a. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016b. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Salman Hameed Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. 2014. Geometry driven semantic labeling of indoor scenes. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 679–694. Springer.

Andrew K Lampinen, Nicholas Roy, Ishita Dasgupta, Stephanie CY Chan, Allison Tam, James Mcclelland, Chen Yan, Adam Santoro, Neil C Rabinowitz, Jane Wang, et al. 2022. Tell me why! explanations support learning relational and causal structure. In *International Conference on Machine Learning*, pages 11868–11890. PMLR.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024c. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024d. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *Transactions of the Association for Computational Linguistics*, 8:795–809.

Meta AI. 2024. 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/.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. 2024. Introducing openai o1-preview. Accessed: 2024-12-16.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ben Prystawski, Michael Y Li, and Noah D Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. *arXiv preprint arXiv:2304.03843*. 22 pages, 6 figures.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Amirudin, Muhammad Ridzuan, Daniya Kareem, Ketan More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Xavier, Amit Bhatkal, Hawau Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Khan. 2024. All languages matter:

Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2019. Probabilistic neural symbolic models for interpretable visual question answering. In *International Conference on Machine Learning*, pages 6428–6437. PMLR.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. 2024. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *Preprint*, arXiv:2407.04973.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024a. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024b. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024a. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024b. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024c. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024a. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024b. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2024. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models.

A Appendix

This appendix provides comprehensive supplementary materials to support our study. Sec. B covers related work, summarizing advancements in multimodal reasoning, curriculum learning, and inference optimization, positioning our approach within the broader research landscape. Sec. C presents additional results, including detailed performance breakdowns, ablation studies, and qualitative examples that further validate the effectiveness of our proposed model. Sec. **??** details the prompting strategies used for VRC-Bench creation and evaluation, ensuring reproducibility and transparency in benchmark design. These additional insights reinforce the robustness of our methodology and the significant contributions of our work to multimodal reasoning research.

B Related Works

Reasoning with LLMs: The development of robust reasoning capabilities in Large Language Models (LLMs) has been a focal point of research. Early work often relied on neural-symbolic methods for explicit modeling of the reasoning process using formal language instead of natural language (Roy and Roth, 2015; Chiang and Chen, 2019; Amini et al., 2019). However, the emergence of powerful LLMs has prompted new approaches that leverage their inherent reasoning abilities (Wu et al., 2024). For example, inference time computing is scaled in recent models to perform reasoning before giving the final answer (Xiong et al., 2024; Weng et al., 2022; Huang et al., 2022; OpenAI, 2024). Techniques like Chain-of-Thought (CoT) prompting, where a complex question is decomposed into intermediate reasoning steps, have shown promise in guiding LLMs to structured solutions (Wei et al., 2022; Yao et al., 2024b). Nevertheless, maintaining logical consistency, especially in tasks requiring multi-step inference, poses a significant challenge, leading to errors and hallucinated outputs (Xu et al., 2024b; Madaan et al., 2024). LLMs, even with CoT guidance, might generate unfaithful explanations, deviate from logical reasoning paths, and struggle with verifying and selecting correct reasoning steps (Wei et al., 2022). These approaches are further extended to VLMs.

Reasoning with VLMs: Visual reasoning tasks require models to possess visual perception and high-level cognitive abilities (Gupta and Kembhavi, 2023; Khan et al., 2014; Xu et al., 2024b). The visual reasoning skills have broad applicability across domains such as science (Lu et al., 2022), mathematics (Lu et al., 2023), robotic planning (Hu et al., 2023) and advanced question answering (Yue et al., 2024). Similar to the case of LLMs, the conventional approaches employed neural-symbolic methods to explicitly model the reasoning process (Garcez et al., 2019; Vedantam et al., 2019; Andreas et al., 2016). For example, (Amizadeh et al., 2020) propose differentiable logic formalism to decouple the reasoning aspect of VQA from visual perception. More recent VLMs leverage the reasoning capabilities of LLMs for visual tasks. Visual programming (Gupta and Kembhavi, 2023) provides a modular neuro-symbolic system based on computer vision models as functions and GPT-3 LLM for compositional visual reasoning. Zhang et al. (Zhang et al., 2024a) argue that VLM training with concise answers results in reduced generalization to more complex problems requiring reasoning. They use GPT-4o model to create rationales and use correct and incorrect reasoning chains in training to enhance model's reasoning ability via reinforcement learning (RL) (Rafailov et al., 2024). In contrast, LlaVA-o1 (Xu et al., 2024b) does not use RL and advocates for stage-wise reasoning instead of CoT prompting, where the answer is worked out sequentially via summarization, interpretation, reasoning, and conclusion steps. Our work builds on (Xu et al., 2024b) and shows the importance of curriculum learning and path search for visual reasoning.

Recently, M3CoT (Chen et al., 2024b) introduces a multi-step chain-of-thought (CoT) approach where reasoning steps are generated multiple times for the same question in an ensembled manner. By exploring multiple reasoning paths and aggregating the most reliable conclusions, this method enhances robustness and consistency. However, this approach lacks a structured problem-solving process, as it focuses on multiple independent reasoning attempts rather than following a clear, step-by-step logical flow. While ensembling improves accuracy, it also increases computational overhead and does not explicitly integrate contextual visual understanding into the reasoning process. In contrast, LlamaV-o1 follows a structured, interpretable multi-step reasoning framework, ensuring a more logical and efficient problem-solving approach. The model first summarizes the approach to solving the problem, then generates an image caption relevant to the

input question to establish contextual grounding. Using these components, it constructs a coherent step-by-step reasoning chain, leading to a logically derived final answer. Unlike M³CoT, which relies on multiple independent reasoning attempts, LlamaV-o1 ensures logical consistency from the outset, integrates question-aware visual understanding, and improves inference efficiency by reducing redundant computations. This structured methodology makes LlamaV-o1 not only more interpretable and scalable but also more efficient for real-world multi-step reasoning tasks.

Benchmarks for Visual Reasoning: Several datasets and benchmarks have been developed to evaluate and advance visual reasoning in VLMs. These datasets vary in complexity, visual context, and reasoning skills required. Some notable examples are as follows. CLEVR (Compositional Language and Elementary Visual Reasoning) tests visual reasoning abilities like counting, comparisons, and logical inference through rendered images and automatically generated questions (Johnson et al., 2017). StrategyQA is a multi-hop question-answering dataset on Wikipedia that necessitates implicit decompositions and diverse reasoning strategies (Geva et al., 2021). ScienceQA offers a large-scale multimodal science dataset with multi-modal contexts, diverse science topics, and annotated answers with corresponding lectures and explanations (Lu et al., 2022). A consolidated mathematical reasoning benchmark in diverse visual contexts called MathVista incorporates 28 existing multimodal datasets and three new datasets (Lu et al., 2023). Zhang *et al.* (Zhang et al., 2024a) propose ShareGPT-4o-Reasoning, a comprehensive CoT dataset containing 193k examples covering various VQA tasks, designed to improve CoT reasoning in VLMs. However, these benchmarks do not provide step-by-step reasoning in complex evaluation scenarios and generally judge the correctness based on only the final answer. In this work, our goal is to provide a comprehensive benchmark that assesses the reasoning chains as well as the final outcome in complex reasoning scenarios.

C Additional Results

In this section, we provide a detailed analysis of our model's reasoning performance. Table 5 contains the comprehensive set of attributes considered in our evaluation to assess the quality of reasoning steps in our proposed evaluation. We present a breakdown of reasoning step scores across different aspects of reasoning, as outlined in Table 6, offering deeper insights into how well the model handles logical consistency, coherence, and step-wise accuracy. Additionally, we provide a category-wise performance comparison on VRC-Bench, highlighting strengths and areas of improvement across diverse reasoning challenges. These results further demonstrate the effectiveness of our approach in advancing structured multi-step visual reasoning.

The Table 6 presents a detailed comparison of reasoning performance metrics between close-source models (e.g., GPT-4o, Claude-3.5-Sonnet, Gemini-2.0-Flash) and open-source models (e.g., Llama-3.2-Vision, Llava-CoT, and our model, LlamaV-o1). These metrics evaluate critical aspects of reasoning, such as faithfulness, informativeness, semantic coverage, and logical alignment. *Faithfulness-Step and Token:* LlamaV-o1 performs competitively among open-source models, with scores of 6.51 and 6.36, respectively, demonstrating reliable alignment with ground truth reasoning steps. *Informativeness-Step:* Our model achieves a strong score of 6.77, reflecting its ability to extract and provide relevant information effectively. *Repetition-Token and Redundancy:* LlamaV-o1 maintains low repetition and redundancy levels, scoring 8.42 and 8.13, showcasing its efficiency in delivering concise reasoning without unnecessary repetition. *Hallucination:* Our model minimizes irrelevant or fabricated content, achieving a balanced score of 7.02, highlighting its robustness in reasoning accuracy. *Commonsense and Reasoning Alignment:* With scores of 7.26 and 6.44, LlamaV-o1 demonstrates a strong understanding of commonsense reasoning and maintains consistent alignment with logical reasoning paths. Compared to other open-source models, LlamaV-o1 leads across multiple metrics, offering a significant improvement in step-by-step reasoning quality while remaining competitive with leading close-source models. These results highlight LlamaV-o1's ability to deliver robust, accurate, and interpretable reasoning in multimodal contexts.

Figure 4 illustrates the category-wise performance of our model compared to leading reasoning models in various domains from our benchmark such as Math & Logic, Scientific Reasoning, and Complex Visual Perception. Our model outperforms others in several challenging categories, including Chart & Diagram Understanding (83.18%), Scientific Reasoning (86.75%) and OCR & Doc-

ument Understanding (93.44%). These improvements outline the model's ability to handle tasks requiring visual and logical reasoning in accordance. The results also highlight balanced performance across all categories, reflecting the versatility of our approach.

The results demonstrate that our approach outperforms recent open-source visual reasoning methods while achieving favorable results against its close-source counterparts. By leveraging curriculum learning and optimizing inference efficiency with Beam Search, our model effectively balances reasoning accuracy and computational complexity. Our performance improvements in reasoning tasks are complemented by robust handling of logical errors and visual illusions, as evidenced in benchmarks like HallusionBench. Fig. 5 presents a qualitative comparison between the recent Llava-CoT and our LlamaV-o1 on different examples from the VRC-Bench. Our LlamaV-o1 achieves superior performance both in reasoning steps and the final answer, compared to Llava-CoT.

D VRC-Bench: Prompting & Evaluation Protocol

To ensure a rigorous and reproducible evaluation of VRC-Bench, we design a structured Prompting & Evaluation Protocol that guides the generation and assessment of multi-step reasoning tasks. This protocol establishes a consistent framework for crafting diverse and challenging reasoning prompts, ensuring comprehensive coverage across different reasoning aspects. Our approach incorporates hierarchical prompting strategies, where prompts are designed to progressively guide models through structured reasoning—starting from contextual understanding to step-by-step logical deduction and final answer generation. Additionally, we adopt an automated evaluation pipeline leveraging GPT-4o (, 2024) as an external judge, ensuring robust and fair assessments across reasoning dimensions. By standardizing both prompting and evaluation, VRC-Bench provides a reliable benchmark for assessing large multimodal models' stepwise reasoning capabilities.

D.1 Generating reasoning Steps from Closed Sourced Models

We designed a structured system prompt to guide closed-source models like GPT-4o (, 2024), Claude (cla, 2024), and Gemini (Reid et al., 2024) in generating detailed, step-by-step reasoning for complex tasks. The prompt requires the model to describe each action to be taken and explain how it is executed, ensuring a clear and logical progression throughout the reasoning process. To account for varying levels of complexity, the prompt allows the model to take as many steps as necessary, ensuring that the solution is systematically derived. Additionally, the prompt emphasizes the use of visual elements, guiding the model to reference provided images or diagrams explicitly in its reasoning steps. The prompt is further designed to handle ambiguity effectively by instructing the model to respond with "None of the choices provided" when no valid options are available. This ensures robustness and prevents the generation of forced or inaccurate conclusions. By enforcing a logical flow, grounding the reasoning in visual inputs, and providing explicit instructions for ambiguous scenarios, this prompt enables consistent, interpretable, and reliable reasoning outputs across various multimodal tasks.

D.2 System Prompt used to Evaluate Reasoning Steps

The following system prompt was used to evaluate the reasoning steps of the target model. It defines a structured framework to assess the alignment, coherence, and quality of reasoning through multiple metrics, including faithfulness, informativeness, repetition, hallucination, redundancy, semantic coverage, reasoning alignment, commonsense, and completeness of steps. Each metric is scored on a scale of 1-10, with detailed guidelines ensuring consistent and objective evaluations.

D.3 Response format used to generate structured evaluation scores

To further ensure the evaluation framework generates consistent and interpretable outputs, we designed the response format using a well-defined JSON schema. This schema serves as a blueprint, enforcing strict adherence to a structured format while capturing detailed scores for each metric in a systematic and transparent manner. By standardizing the output structure, the schema facilitates easier comparison between models, reduces ambiguity, and enhances the reproducibility of results.

The JSON schema is carefully tailored to accommodate the unique aspects of our evaluation process, such as step-by-step reasoning, metric-specific scores, and logical flow validation. Each

Table 5: An overview of comprehensive set of attributes considered in our evaluation to assess the quality of reasoning in LMMs. These attributes focus on critical aspects such as faithfulness, informativeness, and logical coherence of reasoning steps. Key measures include ensuring alignment of reasoning steps with the source (Faithfulness-Step and Token), completeness of information (Informativeness-Step), and identifying issues like hallucinations, redundancy, or missing steps. Additional metrics, such as Semantic Coverage and Reasoning Alignment, evaluate the logical and semantic integrity of the response. Together, these metrics provide a robust framework for evaluating the accuracy, completeness, and reliability of LLM-generated reasoning.

Metric	**Definition**
`Faithfulness-Step`	Measures how well the reasoning steps in the LMM response align with the source reasoning steps.
`Faithfulness-Token`	Extends Faithfulness-Step to token-level granularity, checking if the content within each step is accurate.
`Informativeness-Step`	Measures how well the reasoning steps extract all relevant information from the context.
`Repetition-Token`	Identifies repeated or unnecessarily paraphrased reasoning steps.
`Hallucination`	Detects irrelevant or fabricated reasoning steps not aligned with the source.
`Redundancy`	Identifies redundant reasoning steps that do not add value.
`Semantic Coverage-Step`	Measures how well the response covers the essential semantic elements of the source reasoning steps.
`Reasoning Alignment`	Overall alignment between the hypothesis and reference reasoning chain.
`Commonsense`	Checks for missing commonsense reasoning are required to solve the problem.
`Missing Step`	Identifies if any necessary reasoning steps are missing.

Table 6: The table compares reasoning performance metrics across close-source models (e.g., GPT-4o, Claude-3.5) and open-source models (e.g., Llama-3.2-Vision, Llava-CoT, and LlamaV-o1). Our model, LlamaV-o1, demonstrates strong performance in key areas such as faithfulness (6.51), informativeness (6.77), and semantic coverage (6.67), while maintaining low repetition (8.42) and redundancy (8.13). Additionally, it minimizes hallucinations (7.02) and exhibits a solid understanding of commonsense reasoning (7.26) with consistent reasoning alignment (6.44). Among open-source models, LlamaV-o1 achieves the most balanced and robust reasoning performance, showcasing its ability to deliver accurate, interpretable, and efficient reasoning, comparable to leading close-source models. All the scores were given on scale 1-10 providing clearer feedback for reasoning aspect.

	Close-Source						*Open-Source*			
Model	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-2.0 Flash**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-4o mini**	**Llama-3.2 Vision**	**Mulberry**	**Llava-CoT**	**LlamaV-o1** (Ours)
`Faithfulness-Step`	7.19	6.80	7.21	6.66	6.61	6.84	5.08	5.69	6.37	6.51
`Faithfulness-Token`	7.07	6.57	6.95	6.39	6.39	6.73	4.84	5.55	6.12	6.36
`Informativeness-Step`	7.32	7.14	7.49	6.94	6.86	6.98	5.22	5.78	6.54	6.77
`Repetition-Token`	8.93	8.43	7.55	8.41	8.59	8.91	6.26	8.67	8.38	8.42
`Hallucination`	8.05	7.62	7.78	7.60	7.57	7.72	5.55	6.45	6.86	7.02
`Redundancy`	8.68	8.00	7.28	7.99	8.15	8.67	6.11	8.24	8.19	8.13
`Semantic Coverage-Step`	7.24	7.01	7.31	6.82	6.76	6.90	5.08	5.73	6.44	6.67
`Reasoning Alignment`	7.16	6.77	7.13	6.60	6.57	6.82	5.00	5.65	6.21	6.44
`Commonsense`	7.98	7.76	7.90	7.69	7.62	7.72	5.93	6.6	7.12	7.26
`Missing Step`	7.22	6.91	7.25	6.87	6.70	6.89	5.04	5.71	6.31	6.52

response is divided into key components, including reasoning steps, metric scores, and final answers, ensuring that all critical aspects of the model's performance are systematically captured. This level of detail not only improves interpretability but also enables fine-grained analysis of strengths and weaknesses across models.

Additionally, the schema supports modularity, allowing seamless integration of new metrics or evaluation criteria as the benchmark evolves. By adopting this structured approach, we ensure that the evaluation framework remains robust, scalable, and adaptable to future advancements in multimodal reasoning research.

D.4 Evaluating reasoning steps using gpt-4o as a judge

The *evaluate_steps* function is designed to rigorously assess the quality of reasoning steps generated by models against ground truth data using the GPT-4o-mini model. It takes the task question, ground truth reasoning, and model response as inputs and processes them within a structured conversation context. By leveraging a predefined system prompt and parameters like deterministic temperature (0.0) and a maximum token limit of 500, the function ensures consistent and reliable evaluations. The output provides clear feedback on alignment, logical flow, and coherence of rea-

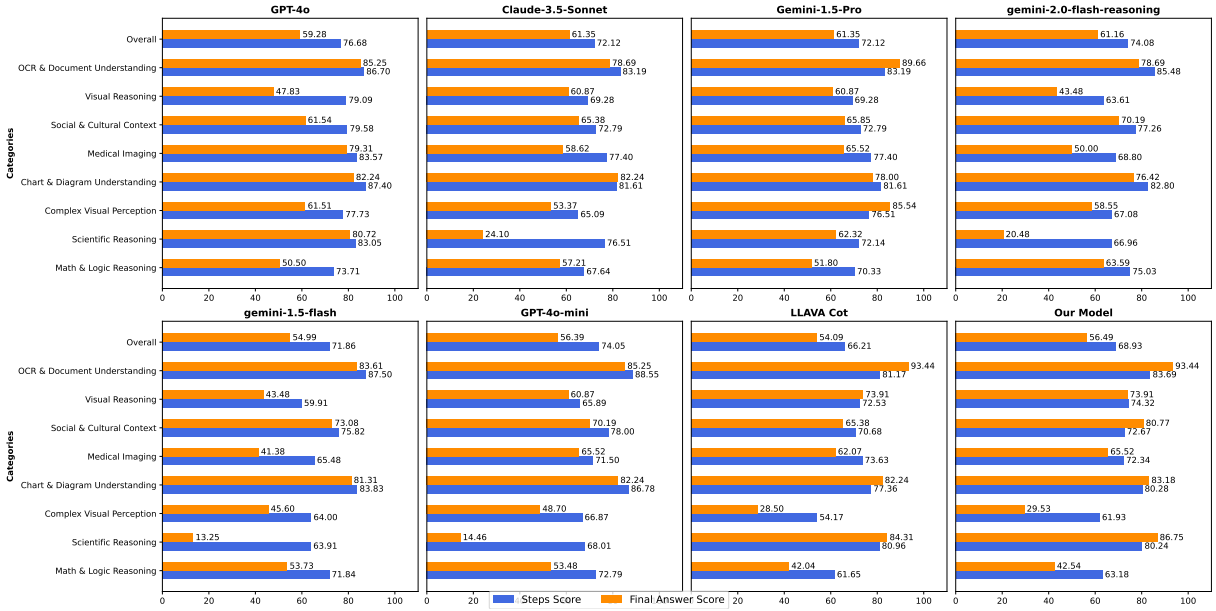


Figure 4: The comprehensive comparison of category-wise and overall performance scores achieved by various models on diverse reasoning tasks. The evaluation spans multiple domains, including Math & Logic Reasoning, Scientific Reasoning, Complex Visual Perception, Chart & Diagram Understanding, Medical Imaging, Social & Cultural Context, Visual Reasoning, and OCR & Document Understanding. The models assessed include GPT-4o, Claude-3.5-Sonnet, Gemini variants, LLAVA-CoT, and our proposed model. Our model demonstrates consistently superior performance in critical categories such as Math & Logic Reasoning, Chart & Diagram Understanding, and Medical Imaging, achieving a balanced improvement across both step-by-step reasoning (Step Scores) and final answer accuracy (Final Answer Scores). Compared to LLAVA-CoT, our approach excels in maintaining high accuracy across tasks while showcasing robustness and interpretability in multi-step reasoning challenges.

System Prompt used for the generation of reasoning steps

```
When answering the question based on the provided image(s),
follow a structured reasoning process and provide the final answer
after solving it step by step. Use the following format for
your response:

Step-by-Step Process:
    Step 1: Describe the action to be taken.
    Action 1: Explain the execution of the first action.

    Step 2: Describe the next action to be taken.
    Action 2: Explain the execution of the second action.

    Step 3: Describe the next action to be taken.
    Action 3: Explain the execution of the second action.

    ...continue as needed...take as many steps you want.

    Step n: Describe the final action to be taken.
    Action n: Execute the final action leading to the conclusion.

    Final Answer: Provide the final solution or conclusion derived from the
        reasoning process.

Ensure each step logically follows the previous one, and explicitly
detail how the image(s) guide the solution at every stage. Also if
options are present and none of options are correct.
Please response None of the choices provided.
```

System Prompt used to evaluate the reasoning steps

```
You are a reasoning evaluator designed to assess the alignment, coherence, and
     quality of reasoning steps in text responses. Your task is to evaluate
    reasoning steps between the *ground truth* and the *LLM response* using
    the following metrics:

1. **Faithfulness-Step (1-10):**
   - Definition: Measures how well the reasoning steps in the LLM response
       align with the source reasoning steps.
   - Scoring Guidelines:
     - 9-10: All or almost all steps match or closely reflect the ground truth
         reasoning.
     - 7-8: Most steps are aligned, with minor deviations.
     - 5-6: Some steps align, but several are missing or significantly altered
        .
     - 3-4: Few steps align correctly; most are off or missing.
     - 1-2: The majority of steps are not aligned with the source.

2. **Faithfulness-Token (1-10):**
   - Definition: Extends Faithfulness-Step to a token-level granularity,
       checking if the content within each reasoning step is true to the
       source.
   - Scoring Guidelines:
     - 9-10: Token-level details mirror the ground truth closely.
     - 7-8: Minor token-level deviations but largely faithful.
     - 5-6: Noticeable inaccuracies in token-level details.
     - 3-4: Several token-level discrepancies.
     - 1-2: Most token-level details are incorrect or fabricated.

3. **Informativeness-Step (Info-Step) (1-10):**
   - Definition: Measures how well the reasoning steps extract all relevant
       information from the source.
   - Scoring Guidelines:
     - 9-10: Almost all critical information steps are present and accurate.
     - 7-8: Most important points are included, with minor omissions.
     - 5-6: Some key information is missing or underdeveloped.
     - 3-4: Limited inclusion of critical content.
     - 1-2: Very poor extraction of relevant information.

4. **Repetition-Token (1-10):**
   - Definition: Identifies repeated or unnecessarily paraphrased reasoning
       steps within the hypothesis.
   - Scoring Guidelines:
     - 9-10: No or minimal unnecessary repetition.
     - 7-8: Minor repetition that doesn't impede clarity.
     - 5-6: Noticeable repetition that doesn't add value.
     - 3-4: Frequent repetition that disrupts coherence.
     - 1-2: Excessive repetition reducing the quality of reasoning.

5. **Hallucination (1-10):**
   - Definition: Detect irrelevant or invented reasoning steps not aligned
       with the source.
   - Scoring Guidelines:
     - 9-10: No hallucinations; all reasoning is grounded in the source.
     - 7-8: One or two minor hallucinations.
     - 5-6: Several steps contain invented or irrelevant details.
     - 3-4: Many hallucinations, but some grounding remains.
     - 1-2: Mostly hallucinated reasoning.
```

System Prompt used to evaluate the reasoning steps continued...

```
6. **Redundancy (1-10):**
   - Definition: Identify redundant reasoning steps that do not add value.
   - Scoring Guidelines:
     - 9-10: No unnecessary steps; very concise.
     - 7-8: Minor redundancy.
     - 5-6: Some steps clearly unnecessary.
     - 3-4: Many redundant steps.
     - 1-2: Excessive redundancy that hampers clarity.

7. **Semantic Coverage-Step (1-10):**
   - Definition: How well the hypothesis covers the essential semantic
       elements from the source reasoning steps.
   - Scoring Guidelines:
     - 9-10: Almost complete semantic coverage of all important elements.
     - 7-8: Good coverage but some minor elements are missing.
     - 5-6: Partial coverage with noticeable gaps.
     - 3-4: Significant semantic gaps.
     - 1-2: Very poor coverage of essential meaning.

8. **Reasoning Alignment (1-10):**
   - Definition: Overall alignment between the hypothesis and the reference
       reasoning chain.
   - Scoring Guidelines:
     - 9-10: Very closely aligned, minimal divergence.
     - 7-8: Mostly aligned, with some minor issues.
     - 5-6: Some alignment, but also several misalignments.
     - 3-4: Poor alignment, though occasional matches.
     - 1-2: Fundamentally misaligned reasoning.

9. **Commonsense (1-10):**
   - Definition: Check for missing commonsense reasoning required to solve the
       problem.
   - Scoring Guidelines:
     - 9-10: Adequate commonsense reasoning present.
     - 7-8: Minor commonsense gaps but mostly adequate.
     - 5-6: Noticeable commonsense gaps.
     - 3-4: Many commonsense steps missing.
     - 1-2: Almost entirely lacking necessary commonsense.

10. **Missing Step (1-10):**
    - Definition: Identify if any necessary reasoning steps are missing.
    - Scoring Guidelines:
      - 9-10: No critical steps missing.
      - 7-8: Minor missing steps that don't significantly affect the
          conclusion.
      - 5-6: Some important steps absent, affecting the outcome.
      - 3-4: Several crucial missing steps.
      - 1-2: Major gaps; the reasoning chain is incomplete.
```

System Prompt used to evaluate the reasoning steps continued...

```
**Additional Instructions for Consistency:**

- Always follow the above scoring guidelines strictly.
- Before scoring, re-read both the ground truth and the LLM response carefully
    .
- Compare the reasoning steps directly to determine where they align or
    diverge.
- Use the provided scoring benchmarks (anchor examples, if any) as a reference
     to maintain consistency across evaluations.
- Avoid subjective interpretation and adhere to the given thresholds.
- Once scores for all metrics are determined, compute the Overall Score as the
     average of all metric scores.
- Provide the final output as a Python dictionary with the structure only don'
    t add a anything extra , beacuase your out will be used in code pipeline.
    So single change in you output will crash whole system. :

# Example output : {'Faithfulness-Step': 8.0, 'Faithfulness-Token': 7.5, '
    Informativeness-Step': 8.5, 'Repetition-Token': 9.0, 'Hallucination': 9.5,
     'Redundancy': 8.0, 'Semantic Coverage-Step': 8.5, 'Reasoning Alignment':
    8.0, 'Commonsense': 9.0, 'Missing Step': 8.5 , 'Overall Score': 8.65}

# Do not give output in following format :

```python
{
  'Faithfulness-Step': 1.0,
  'Faithfulness-Token': 1.0,
  'Informativeness-Step': 1.0,
  'Repetition-Token': 9.0,
  'Hallucination': 1.0,
  'Redundancy': 9.0,
  'Semantic Coverage-Step': 1.0,
  'Reasoning Alignment': 1.0,
  'Commonsense': 1.0,
  'Missing Step': 1.0,
  'Overall Score': 2.6
}
```
```

soning steps, enabling precise analysis of model performance. This automated and standardized approach enhances objectivity, reproducibility, and detailed insight into multimodal reasoning capabilities.

D.5 Evaluating final answer accuracy

To objectively assess how well the model's final answer predictions align with the ground truth, we developed a comparison function that utilizes a secondary system prompt to evaluate response accuracy. This function analyzes the semantic similarity between the ground truth and the model's output, assigning a binary score: 1 for a match and 0 for a mismatch. By exclusively producing numeric scores, this approach ensures a precise and quantifiable evaluation of the model's performance, effectively complementing the structured framework outlined earlier.

Response format provided to LLM's which supports structured-output

```
response_format = {
    "type": "json_schema",
    "json_schema": {
        "name": "EvaluationScores",
        "strict": True,
        "schema": {
            "type": "object",
            "properties": {
                "Faithfulness-Step": {"type": "number"},
                "Faithfulness-Token": {"type": "number"},
                "Informativeness-Step": {"type": "number"},
                "Repetition-Token": {"type": "number"},
                "Hallucination": {"type": "number"},
                "Redundancy": {"type": "number"},
                "Semantic Coverage-Step": {"type": "number"},
                "Reasoning Alignment": {"type": "number"},
                "Commonsense": {"type": "number"},
                "Missing Step": {"type": "number"},
                "Overall Score": {"type": "number"}
            },
            "required": [
                "Faithfulness-Step",
                "Faithfulness-Token",
                "Informativeness-Step",
                "Repetition-Token",
                "Hallucination",
                "Redundancy",
                "Semantic Coverage-Step",
                "Reasoning Alignment",
                "Commonsense",
                "Missing Step",
                "Overall Score"
            ],
            "additionalProperties": False
        }
    }
}
```

Reasoning steps evaluation using gpt-4o as a judge

```
def evaluate_steps(question , ground_truth, llm_response):
    messages = [
        {"role": "system", "content": system_prompt},
        {
            "role": "user",
            "content": [
                {"type": "text", "text": question + "\n" + f"Ground Truth : {
                    ground_truth}" + "\n" + f"LLM Response : {llm_response}"
                    },
            ],
        }
    ]
    response = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=messages,
        response_format=response_format,
        max_tokens=500,
        temperature = 0.0,
    )
    return response.choices[0].message.content
```

Secondary system prompt to evaluate the final answer accuracy

```
system_prompt_2 = """
    You are a helpful Assistant. Provide helpful
    response to the user's question.
"""
```

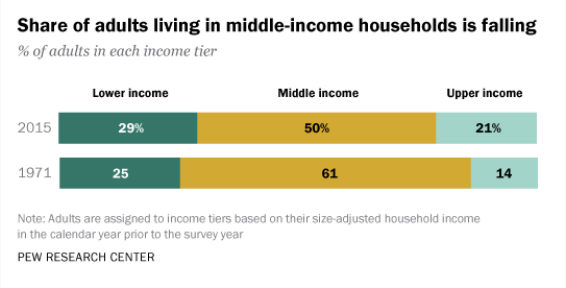
Evaluate the Final Answer

```
def compare_results(question, ground_truth, llm_response):
    messages = [
        {"role": "system", "content": system_prompt_2},
        {
            "role": "user",
            "content": [
                {
                "type": "text",
                "text": f"""
                    Evaluate the following answer based on
                    Accuracy:
                    Question: {question}
                    Ground Truth: {ground_truth}
                    Model Prediction: {llm_response}
                    Match the meaning of the ground truth with the
                    model prediction and if it matches give a 1.
                    Otherwise 0.
                    Strictly return only the numeric score,
                    without any additional commentary
                """
                },
            ],
        }
    ]

    response = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=messages,
        max_tokens=10,
        temperature=0.0
    )

    return response.choices[0].message.content
```

Multi-Step Training Example



Human: What was the value of the middle-income share in 1971? Answer the question using a single word or phrase. Please generate a summary of the picture.

Assistant: I will examine the image to find the relevant data for the middle-income share in 1971 and present the answer in the specified format.

Human: Please generate a detailed caption for the image.

Assistant: The image displays a bar chart comparing the percentage of adults in the lower, middle, and upper-income tiers for the years 2015 and 1971. It shows that in 1971, the middle-income share was 61\%.

Human: Please generate a detailed reasoning to answer the question given the caption.

Assistant: To solve the problem, I will look for the specific percentage associated with the middle-income group in the year 1971 from the provided chart. The image clearly shows that in 1971, the middle-income tier accounted for 61\% of adults.

Human: Please generate the final answer based on reasoning steps. Do not output anything else.

Assistant: 0.61

Human: How many did the share of the middle-class change from 1971 to 2015? Answer the question using a single word or phrase. Please generate a summary of the picture.

Assistant: To solve this problem, I will analyze the image data to determine the change in the share of adults living in middle-income households from 1971 to 2015. I will compare the percentages provided for each year and calculate the difference.

Human: Please generate a detailed caption for the image.

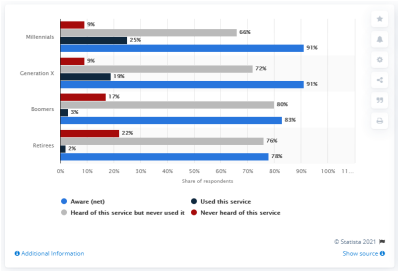
Assistant: The image is a bar graph from Pew Research Center showing the share of adults in lower, middle, and upper-income households for the years 1971 and 2015. For 1971, 61\% were in the middle income, and for 2015, 50\% were in the middle income.

Human: Please generate a detailed reasoning to answer the question given the caption.

Assistant: To find the change in the share of middle-class adults from 1971 to 2015, subtract the 2015 percentage from the 1971 percentage. The middle-income share in 1971 was 61\%, and in 2015 it was 50\%. Thus, the change is calculated as 61\% - 50\% = 11\%. The difference in share, expressed as a decimal, is 0.11.

Human: Please generate the final answer based on reasoning steps. Do not output anything else.

Assistant: 0.11



Question :
What is the sum of the people of Boomers and Generation X who have used this service?
Answer : 22

Llava-CoT

Reasoning Steps:
First, I will locate the data for Boomers and Generation X who have used the service. For Boomers, the percentage is 80%, and for Generation X, it is 91%. Adding these two percentages gives 80% + 91% = 171%.

Final Answer:
171

LlamaV-o1

Reasoning Steps:
First, I will locate the percentage of Boomers who have used the service, which is 3%. Next, I will find the percentage of Generation X who have used the service, which is 19%. Finally, I will add these two percentages together: 3% (Boomers) + 19% (Generation X) = 22%.

Final Answer:
22

Question :
Where is this game being played?
Answer :
The game is being played outdoors on a sidewalk.

Llava-CoT

Reasoning Steps:
The presence of trees and a paved surface suggests an outdoor setting, typical of urban parks or public squares. The casual attire and informal gathering indicate a relaxed, public environment. These factors are commonly associated with a park or similar outdoor recreational area.

Final Answer:
Park

LlamaV-o1

Reasoning Steps:
The presence of a sidewalk, trees, and parked cars suggests that the game is being played in a public outdoor area, likely in a city or town. The casual attire of the players and the informal setting further support this conclusion.

Final Answer:
The game is being played on a sidewalk in a city or town.

Question :
Is the number of brown suvs less than the number of brown rubber school buss?
Choices: A) Yes B) No.
Please select the correct option by its letter.
Answer : B

Llava-CoT

Reasoning Steps:
1. Identify the brown SUV in the image, which is located on the left side. 2. Identify the brown rubber school bus in the image, which is located on the right side. 3. Count the number of brown SUVs, which is 1. 4. Count the number of brown rubber school buses, which is 1. 5. Compare the counts: 1 (brown SUV) is less than 1 (brown rubber school bus). 6. Determine that the number of brown SUVs is indeed less than the number of brown rubber school buses.
Final Answer:
A) Yes

LlamaV-o1

Reasoning Steps:
Step-by-Step Process: Step 1: Identify the number of brown SUVs in the image. Action 1: There is only one brown SUV present in the image. Step 2: Identify the number of brown rubber school buses in the image. Action 2: There are zero brown rubber school buses present in the image. The bus is teal reddish-pink. Step 3: Compare the number of brown SUVs and brown school buses. Action 3: The number of brown SUVs (1) is greater than the number of brown school buses (0).
Final Answer:
B) No

Figure 5: Qualitative comparison between Llava-CoT and the proposed LlamaV-o1 on examples from the VRC-Bench. *First row:* the example shows visual reasoning capabilities on an example chart. Here, Llava-CoT makes mistakes (highlighted in red) for both the intermediate steps and the final answer. In Comparison, our LlamaV-o1 provides an accurate description of the steps as well as the final answer. *Second row:* While both Llava-CoT and our LlamaV-o1 provide accurate step descriptions on an example real-world VQA, Llava-CoT fails to infer the final answer. *Last row:* Llava-CoT fails to accurately answer for the counting task, while also missing the intermediate counting steps. In contrast, our LlamaV-o1 model performs better in intermediate reasoning steps while also providing the accurate final answer.