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

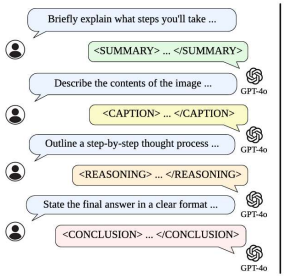
Guowei Xu*, Peng Jin*, Ziang Wu*, Hao Li, Yibing Song, Lichao Sun, Li Yuan

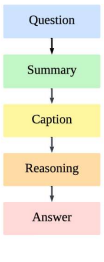
Source Code | Model | Dataset

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.

Structured Thinking — **Spontaneous, Systematic Reasoning**

LLaVA-CoT decomposes the answer generation process into **four structured reasoning stages**: summary, caption, reasoning, conclusion.

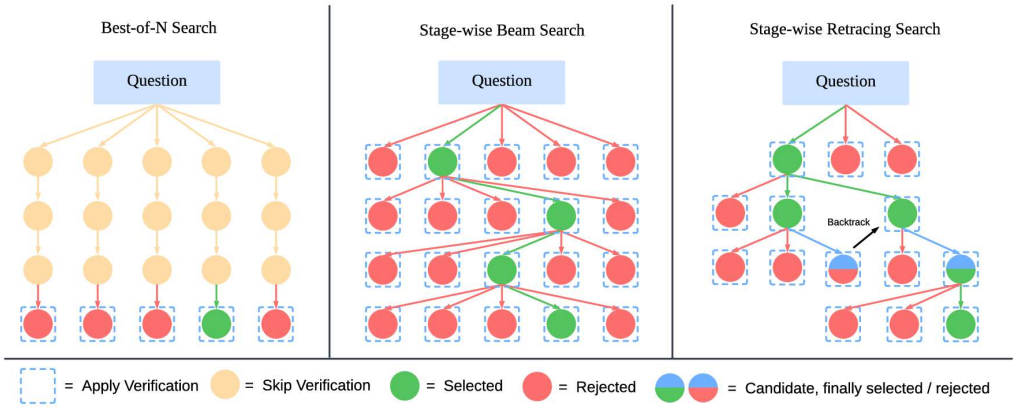




Dataset	Type	Size
ShareGPT4V [9]	General VQA	31.3k
ChartQA [41]	General VQA	17.2k
A-OKVQA [48]	General VQA	16.1k
AI2D [28]	Science-Targeted VQA	11.4k
GeoQA+ [7]	Science-Targeted VQA	11.4k
ScienceQA [37]	Science-Targeted VQA	5.6k
DocVQA [42]	General VQA	4.0k
PISC [31]	General VQA	1.0k
CLEVR [27]	General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

- **Summary:** A brief outline in which the model summarizes the task.
- **Caption:** A description of the relevant parts of an image.
- **Reasoning:** A detailed analysis in which the model systematically considers the question.
- **Conclusion:** A concise summary of the answer.

Test-time Scaling: Stage-wise retracing search (🏆 SWIRES) incorporates a retracing mechanism into the reasoning process.



- **Step 1:** At each stage, generate **M candidate responses**.
- **Step 2:** Check whether at least one of the generated responses surpasses a **reward threshold**.
- **Step 3:** If so, select the **top N** responses to advance, each producing **M/N new candidates** to keep M candidates in total.
- **Step 4:** If no candidate exceeds the threshold, the **algorithm backtracks to the previous stage**, regenerates responses, and produces new candidates. The search halts once a new response surpasses the threshold. Otherwise, retracing continues (up to C times).
- **Step 5:** Finally, the response with the highest reward is chosen.

ICCV OCT 19-23, 2025 HONOLULU HAWAII

💻 **LLaVA-CoT is open source!**

LLaVA-CoT