Chain-of-region: Visual Language Models Need Details for Diagram Analysis

Paper Link: https://openreview.net/pdf?id=M6fYrICcQs


Xue Li


Yiyou Sun


Wei Cheng


Yinglun Zhu

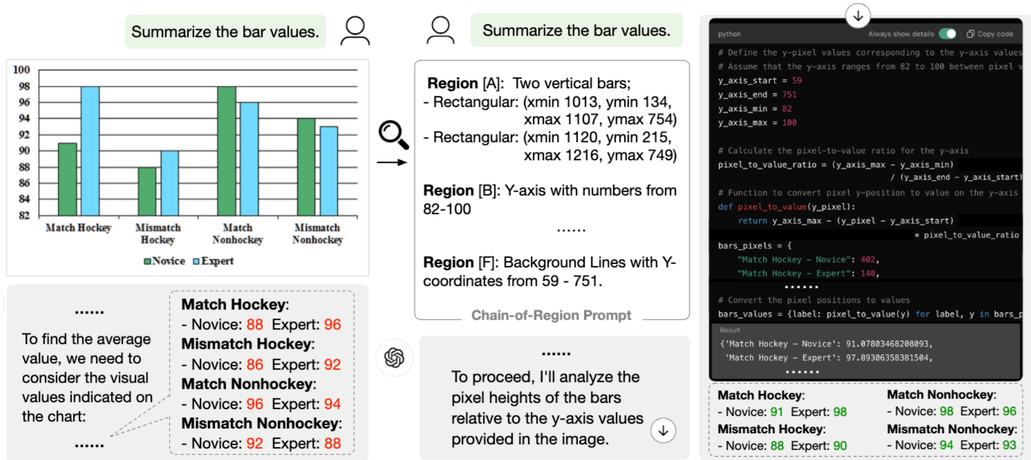

Haifeng Chen

NEC Laboratories America

UC RIVERSIDE

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

Introduction



(a) Original VLM fails to capture visual details (b) Augmented VLM with Chain-of-region (CoR) Prompt

Background

- Recent advances in Visual Language Models (VLMs), such as **GPT-4V** and **Gemini**, have shown great success in multimodal tasks like Navigation, Search, Scientific QA
- Despite these, **critical perceptual limitations** remain:
  - Struggle with **fine visual details**
  - Misled by **visual distractors**
  - Poor understanding of **visual relationships**
  - Prone to **hallucination** of non-existent objects

CoR Advantages

1. **White-box**, interpretable steps
2. **Cost-effective**, CPU-only, fast
3. **Plug-and-play** with existing VLMs
4. **Outperforms** DL-based methods

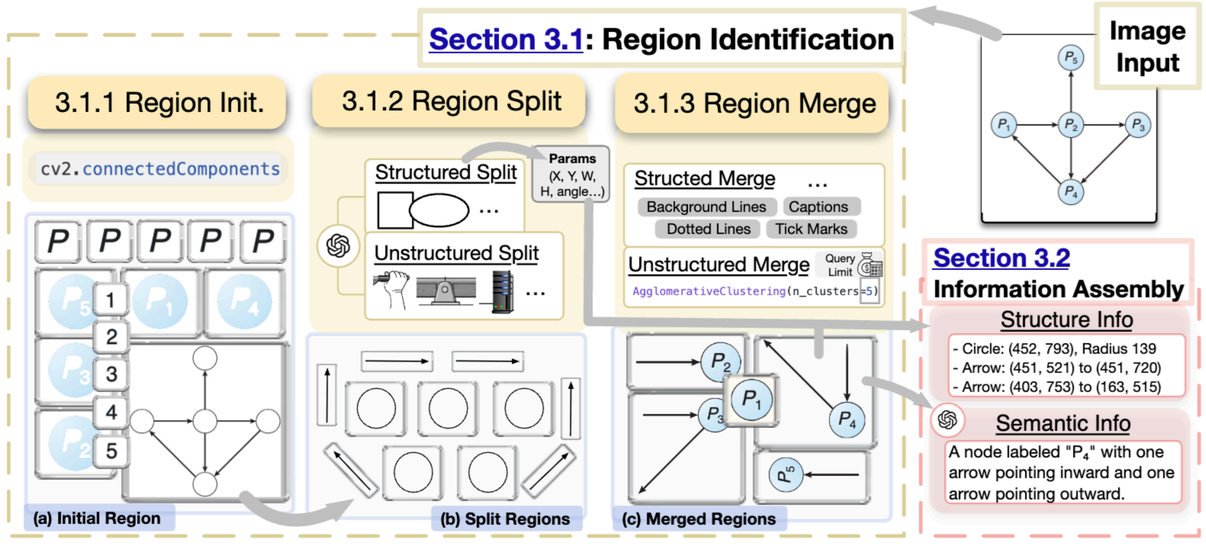
Experiments

Table 1: **Main results for QA on scientific diagrams across the MMMU dataset.** `Bio.`, `Chem.`, `Geo.`, `Arch.`, `CS`, `Elec.`, `Mater.` and `ME` are short aliases for Biology, Chemistry, Geography, Architecture & Engineering, Computer Science, Electronics, Energy & Power, Materials, and Mechanical Engineering respectively.

Method	Science					Tech and Engineering						Average
	Bio.	Chem.	Geo.	Math	Physics	Arch.	CS	Elec.	Energy	Mater.	ME	
GPT4o-mini	40.9	35.5	52.3	47.2	40.0	25.8	33.4	34.6	30.6	42.0	31.9	37.6
+ Zero-shot CoT	38.4	32.6	52.1	38.7	45.7	29.2	35.1	23.1	34.7	42.4	44.4	37.9
+ Few-shot CoT	31.3	37.7	50.9	45.1	42.3	30.4	45.5	31.0	47.5	38.7	35.0	39.6
+ SAM2	38.5	40.0	54.8	52.2	47.9	40.5	34.7	35.3	25.3	45.5	41.9	41.5
+ SoM	31.2	33.6	40.0	43.5	45.1	35.2	27.0	17.9	41.2	34.2	29.9	34.4
+ CoR (Ours)	42.0	34.6	51.8	51.4	52.8	45.1	43.3	33.5	48.7	44.9	51.9	**45.4**
GPT4-Turbo	41.5	34.8	51.0	41.6	49.9	47.0	57.4	34.6	41.5	45.0	32.5	43.3
+ Zero-shot CoT	43.3	25.1	59.6	46.5	58.3	44.8	50.8	31.3	46.8	31.4	50.0	44.3
+ Few-shot CoT	40.0	37.8	51.8	52.8	52.4	51.6	57.9	47.6	45.2	32.2	51.7	47.4
+ SAM2	38.3	39.9	48.4	63.4	48.0	44.2	50.6	48.6	66.9	37.9	49.3	48.7
+ SoM	41.0	30.3	44.8	47.6	44.0	42.3	37.5	15.1	42.1	26.9	24.7	36.0
+ CoR (Ours)	42.9	43.5	56.0	56.8	61.0	51.9	55.7	52.4	65.2	46.0	57.0	**53.5**
GPT4o	41.3	33.6	45.4	48.8	51.6	28.8	41.7	24.6	45.1	31.1	22.2	37.7
+ Zero-shot CoT	47.7	46.1	47.4	52.2	66.6	44.3	50.6	29.3	41.5	42.6	47.5	46.9
+ Few-shot CoT	47.8	38.6	58.0	44.9	56.1	51.3	35.2	34.2	41.8	36.5	42.3	44.2
+ SAM2	44.7	42.3	41.5	51.6	64.4	42.0	48.8	38.0	48.9	48.9	45.8	47.0
+ SoM	48.9	37.3	44.7	55.1	49.8	25.9	38.9	28.1	57.5	33.9	24.6	40.4
+ CoR (Ours)	49.8	46.7	59.1	53.9	72.3	49.7	49.9	34.7	65.7	41.0	59.1	**52.9**

Takeaway: CoT helps improve better performance in multimodal scientific question answering.

Framework Overview



Examples and Comparison with SAM2