Vision Language Models fail to translate detailed visual features into words

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, Anh Totti Nguyen

AUBURN UNIVERSITY · UNIVERSITY OF ALBERTA

https://aub.ie/blindtest

Summary

- Large language models with vision capabilities *still* **struggle** with **low-level vision tasks** that are trivial to humans.
- This failure is due to failure in translating **detailed visual information into tokens in the LLM**.

Motivation

- Gemini can **solve 42.9%** of the questions in MMMU benchmark **without seeing the input image**. Are we on the right way…Chen et al. 2024
- Most VQA benchmarks **do not exclusively** test **vision** capabilities.
- Text-only LLMs can reach **> 80%** of SotA on DocVQA, TextVQA, ChartQA, AI2D (images are serialized). Analyzing…Hedge et al. 2023

Benchmark findings

58.07%

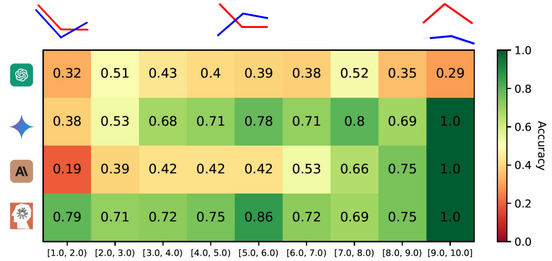
- VLMs cannot reliably
  - Tell if two circles are touching
  - Count the number of times two lines intersect
  - Follow paths from one point to another
  - Reliably count how many rows and columns are in a table
  - Tell which letter is being circled

Examples from **BlindTest** benchmark with VLMs' responses

	P1		P2		P3		P4		P5		P6		P7	
GPT-4o	1	✗	Yes	✗	o	✗	6	✓	5	✗	3×4	✓	1	✓
Gemini-1.5	1	✗	No	✓	w	✗	5	✗	3	✓	3×4	✓	2	✗
Sonnet-3	1	✗	Yes	✗	o	✗	5	✗	4	✗	4×4	✗	2	✗
Sonnet-3.5	0	✓	No	✓	l	✓	6	✓	3	✓	3×4	✓	1	✓

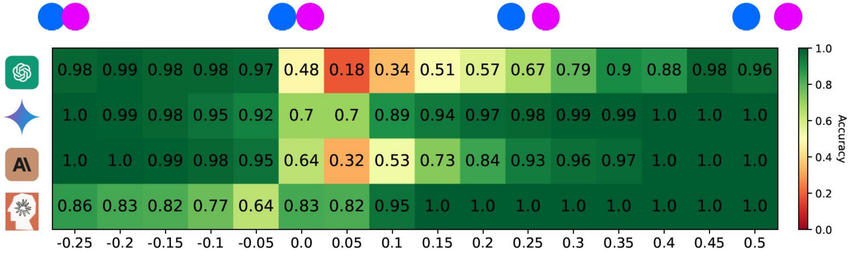
P1: How many times do the blue and red lines touch each other? Answer with a number in curly brackets, e.g., {5}.
P2: Are the two circles overlapping? Answer with Yes/No.
P3: Which character is being highlighted with a red oval? Please provide your answer in curly brackets, e.g. {a}
P4: How many circles are in the image? Answer with only the number in numerical format.
P5: How many squares are in the image? Please answer with a number in curly brackets e.g., {10}.
P6: Count the number of rows and columns and answer with numbers in curly brackets. For example, rows={5} columns={6}.
P7: How many single-color paths go from A to D? Answer with a number in curly brackets e.g. {3}.

Results

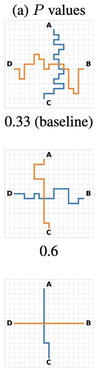


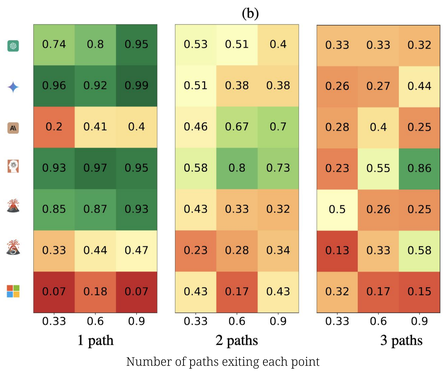
Average distance between 3 points on the lines

VLMs can count intersections more accurately when two lines are distanced away

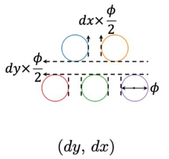


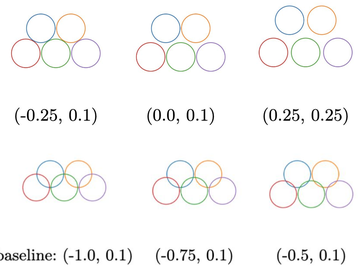
VLMs cannot reliably tell if two circles overlap or not

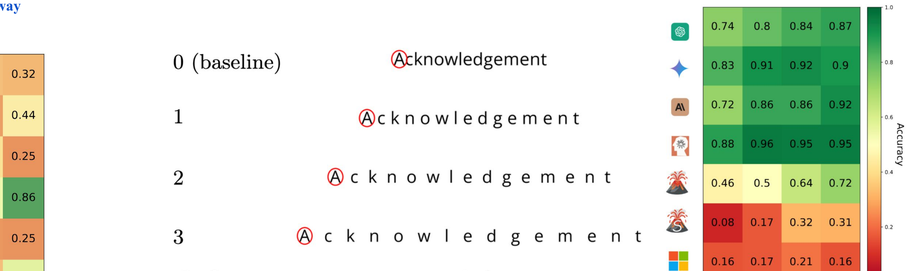




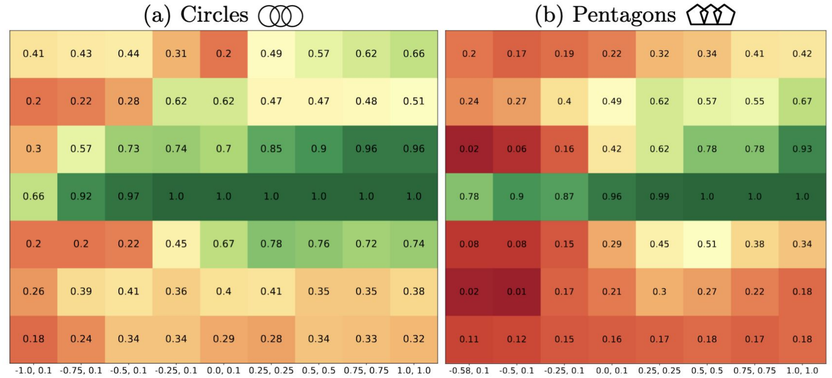
VLMs can count simplified, more straight paths







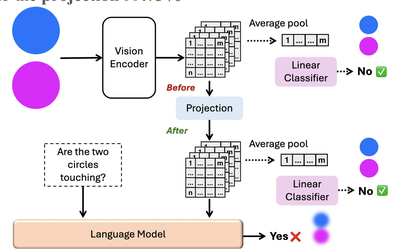
VLMs perform better as the spacing between letter increases



VLMs can more accurately count disjoint shapes

Where do VLMs fail?

- We **probe** the **frozen features** at **4** levels in LLaVA-OneVision 0.5B:
  1. **Before** the **projection 99.82%** 3. **After** the **first decoder 97.11%**
  2. **After** the **projection 99.73%** 4. At the **last decoder 99.52%**

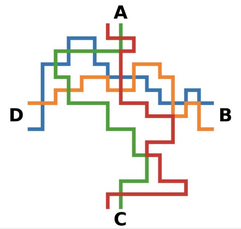


The information necessary for classification is largely preserved in the VLM's latent space; however, it cannot be effectively decoded.

Slow-thinking

- Although, the **performance bottleneck** is in the LLM's ability to **interpret the visual features**, the slow-thinking, i.e., **reasoning models, are on par with their non-reasoning counterparts**.

Gemini 2.0 **Flash 72.75%** vs. Gemini 2.0 **Flash-Thinking 71.59%**



Prompt: How many single-color paths go from C to A?
GT: 2
Thoughts: The user is asking to count the …
… single-color paths from C to A is 1. ✗

Conclusion

- **BlindTest** exposes a **low-level visual shortcoming** in SotA VLMs, i.e., failing to generate **correct answers from detailed visual features**.
- Our findings suggest that **slow-thinking capabilities** do **not improve the VLMs** performance on **BlindTest**.