VLMs have Tunnel Vision:

Evaluating Nonlocal Visual Reasoning in Leading VLMs

Shmuel Berman

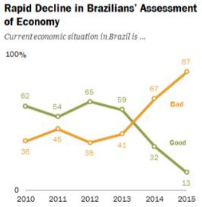
Jia Deng

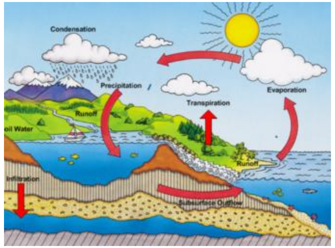
PRINCETON
VISION & LEARNING LAB

Robust Vision

- Image QA performance suggests that VLMs are capable of robust generalist vision

	Claude 3.5 Sonnet[4]	GPT-4o[5]	Qwen2-VL-72B[6]	Phi 4 Multimodal[7]
AI2D[1]	94.7%	94.2%	88.4%	82.3%
ChartQA[2]	90.8%	85.7%	88.3%	81.4%
DocVQA[3]	95.2%	92.8%	96.5%	93.2%





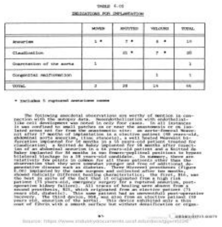


TABLE 6.05

INDICATIONS FOR IMPLANTATION

	WOVEN	KNITTED	VELOURS	TOTAL
Aneurism	1 *	7 *	6 *	14
Claudication		21 *	7 *	28
Coarctation of the aorta	1			1
Congenital malformation			1	1
TOTAL	2	28	14	44

* Includes 5 ruptured aneurisms cases

The following anecdotal observations are worthy of mention in connection with the autopsy data. Neoendothelization with endothelial-like cell development was noted in only four cases. In all instances it was confined to small patches on or near the anastomosis or on isolated areas not far from the anastomotic site: an aorto-femoral Weavenit after 17 months of implantation in a elective patient (48 years-old, abdominal aorta aneurism, iliac stenosis), a well healed Meadox bifurcation implanted for 58 months in a 55 years-old patient treated for claudication; a Knitted de Bakey implanted for 58 months after resection of an abdominal aneurism in a 64 years-old patient and a Knitted de Bakey implanted for 84 months in two femoro-popliteal positions to bypass bilateral blockage in a 38 years-old candidate. In summary, there are relatively few points in common for all these patients other than the observation that they were somewhat younger and free of additional degenerative disease such as diabetes. Three Microvel prostheses (Figure 6.08) implanted by the same surgeon and collected after two months, showed radically different healing characteristics; the first, 816, was the best in spite of the fact that it originated from a high risk patient (75 years-old, emergency surgery for a ruptured aneurism, post-operative kidney failure). All traces of healing were absent from a second prosthesis, 821, which originated from an elective patient (71 years old, diabetic). This same patient had an uneventful post-operative recovery. The third device, 808, was also from an elective patient (59 years old, aneurism of the aorta). This device exhibited only a thin coat of fibrin with a smooth surface but without densification or orga-

85

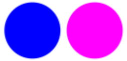
Source: https://www.industrydocuments.ucsf.edu/docs/[illegible]

[1] Kembhavi et al. [ECCV 2016]
[2] Masry et al. [ACL 2022]
[3] Mathew et al. [WACV 2021]
[4] Anthropic et al. [Anthropic 2024]
[5] Hurst et al. [arXiv 2024]
[6] Wang et al. [arXiv 2024]
[7] Abouelenin et al. [arXiv 2025]

Strong Chart Performance ≠ Robust Vision

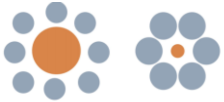
- Despite strong chart performance, these models fail simple perception tests

VLMs are Blind

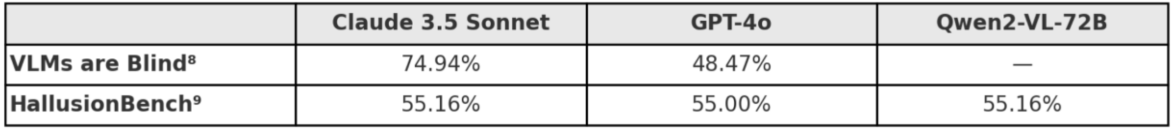


“Are these circles touching?”

HallusionBench



“Are the orange circles the same size?”



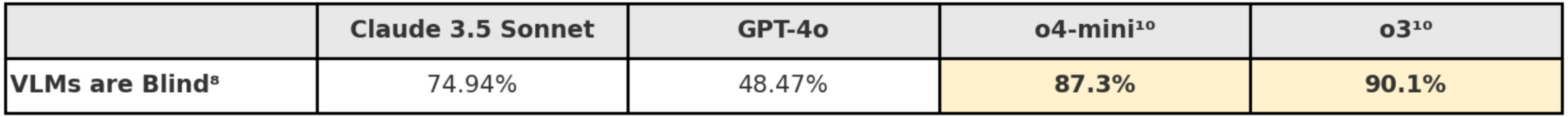
	Claude 3.5 Sonnet	GPT-4o	Qwen2-VL-72B
VLMs are Blind[8]	74.94%	48.47%	—
HallusionBench[9]	55.16%	55.00%	55.16%

[8] Rahmanzadehgervi et al. [ACCV 2024]

[9] Guan et al. [CVPR 2024]

Strong Chart Performance ≠ Robust Vision

- Newer models improve on primitive perception benchmarks
- Do they have robust vision?

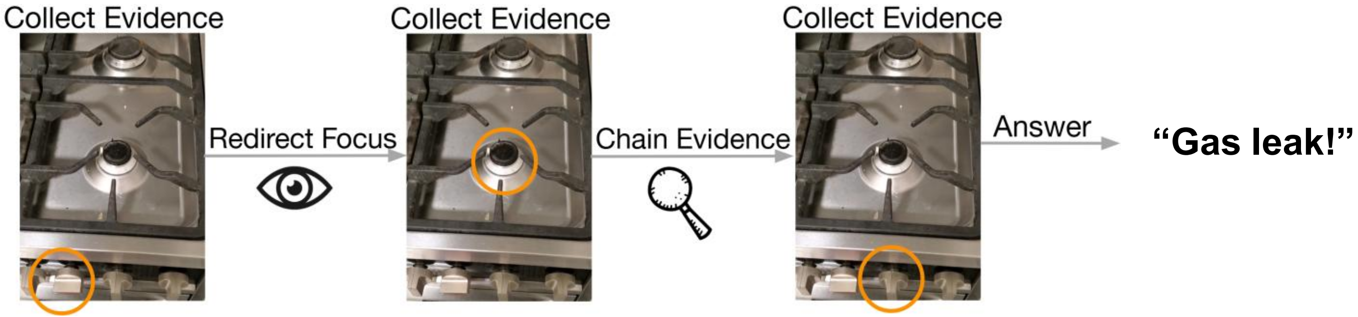


	Claude 3.5 Sonnet	GPT-4o	o4-mini[10]	o3[10]
VLMs are Blind[8]	74.94%	48.47%	**87.3%**	**90.1%**

[8] Rahmanzadehgervi et al. [ACCV 2024]
[10] Ramesh et a. [OpenAI 2025]

How do humans process images?

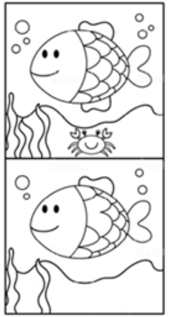
- Humans gather evidence from multiple regions, and use the image itself to redirect focus
- We call this **nonlocal visual reasoning**



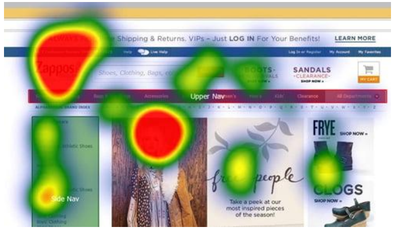
Nonlocal Visual Reasoning Skills

- We define tasks that test three skills

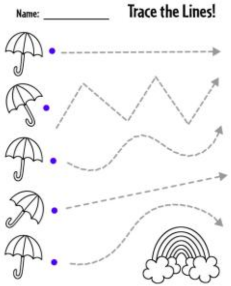
Comparative Perception



Saccadic Search



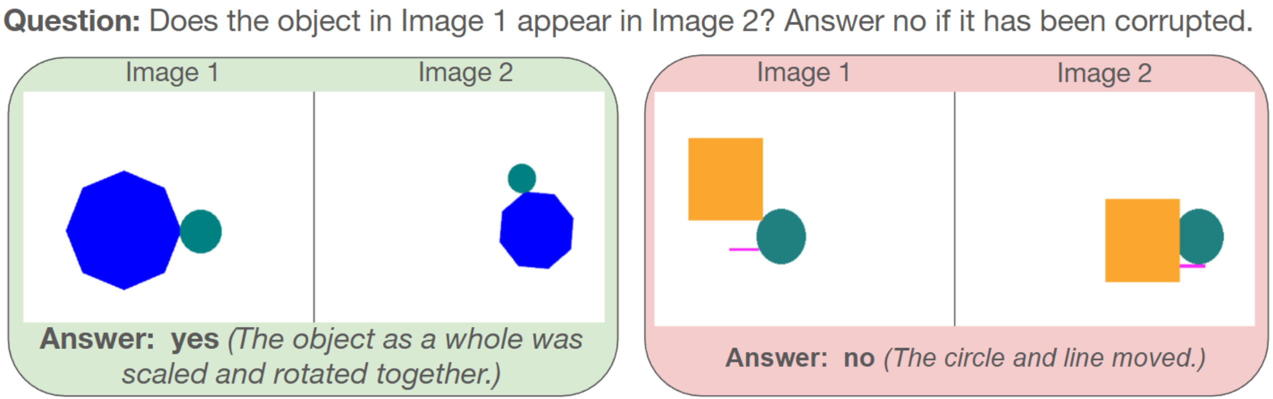
Smooth Visual Search



- Our tasks are designed to minimize necessary background knowledge and require the tested skill

Comparative Perception

- The ability to compare two similar objects
- We evaluate three variants: Standard, Unconnected, Pixel-Perfect



Question: Does the object in Image 1 appear in Image 2? Answer no if it has been corrupted.

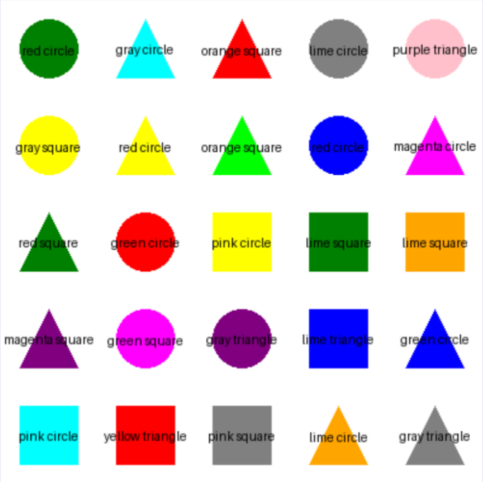
Answer: yes *(The object as a whole was scaled and rotated together.)*

Answer: no *(The circle and line moved.)*

Saccadic Search

- The ability to make discrete, evidence-driven jumps across an image
- We evaluate 2, 3, and 4 jumps

Question: "Starting at the red square, follow the text labels for 3 steps. What color do you end on?"

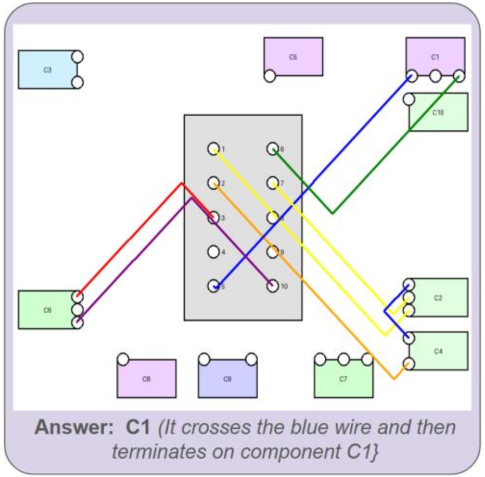


Answer: green *(Red square, yellow triangle, red circle and end with green circle.)*

Smooth Visual Search

- The ability to trace a line or a contour
- We evaluate over the wire colors: Single Color, Standard, Unique Colors

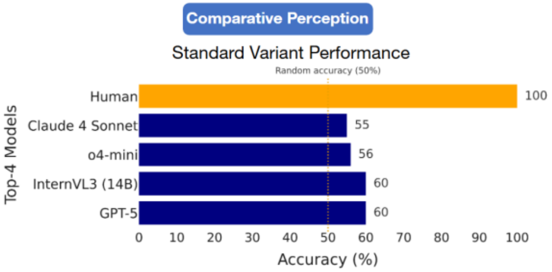
Question: "Which component does the wire from port 6 on the breadboard connect to?"



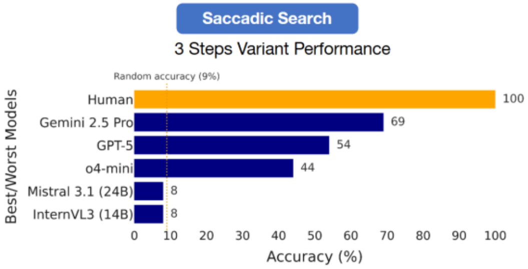
Answer: **C1** *(It crosses the blue wire and then terminates on component C1}*

Main Results

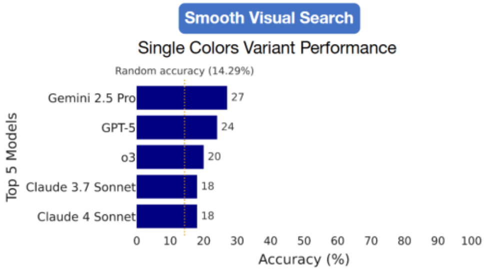
- Models fail catastrophically when comparing connected objects
- VLMs struggles to make discrete steps
- Cannot trace lines without color cues or other heuristics


Comparative Perception

Standard Variant Performance


Saccadic Search

3 Steps Variant Performance


Smooth Visual Search

Single Colors Variant Performance

Conclusion

- Current models achieve strong benchmark scores despite lacking the human-like skills typically used to solve those tasks
- Visual reasoning is not robust even in frontier VLMs
- To the extent VLMs can visually reason, they perform it inconsistently and fail in non-intuitive ways

Paper

Website