IllusionVQA Website

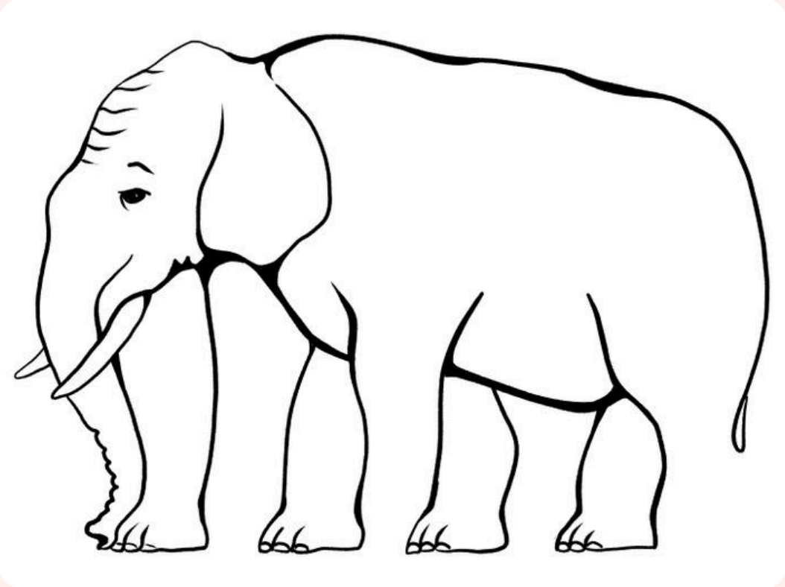
IllusionVQA : A Challenging Optical Illusion Dataset for Vision Language Models

Haz Sameen Shahgir*, Khondker Salman Sayeed*, Abhik Bhattacharjee,
Wasi Uddin Ahmad, Yue Dong, Rifat Shahriyar

UC RIVERSIDE

TL:DR: Vision Language Models (VLM) struggle with understanding and locating optical illusions whereas humans have near-perfect accuracy. We believe it's because current VLMs can't think deliberately about the images they see.

IllusionVQA-Comprehension



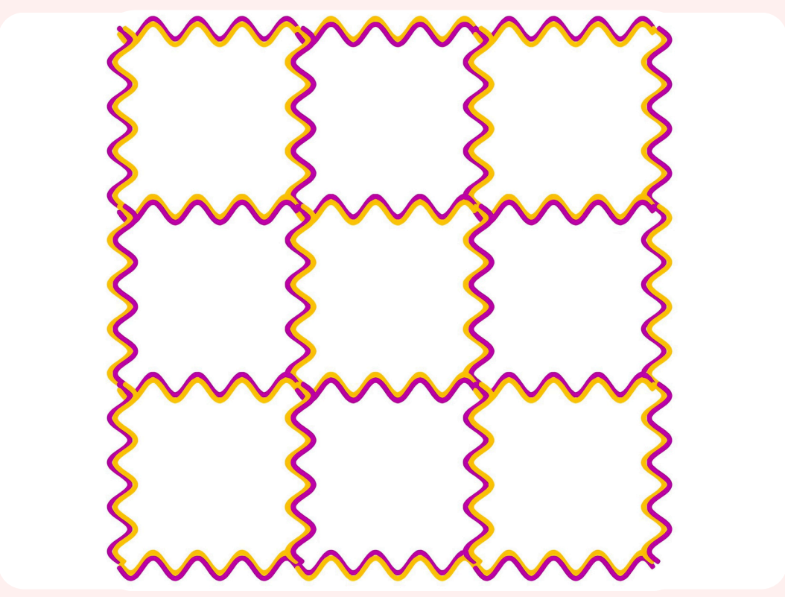
Q: What is unusual about this line drawing of an elephant?

A. The elephant has five or six legs

B. The elephant is using its trunk as a fifth leg

C. The elephant is merging with the background in some regions

D. The elephant has six legs while the rest of its body is normal



Q: How many pale yellow regions are in this image?

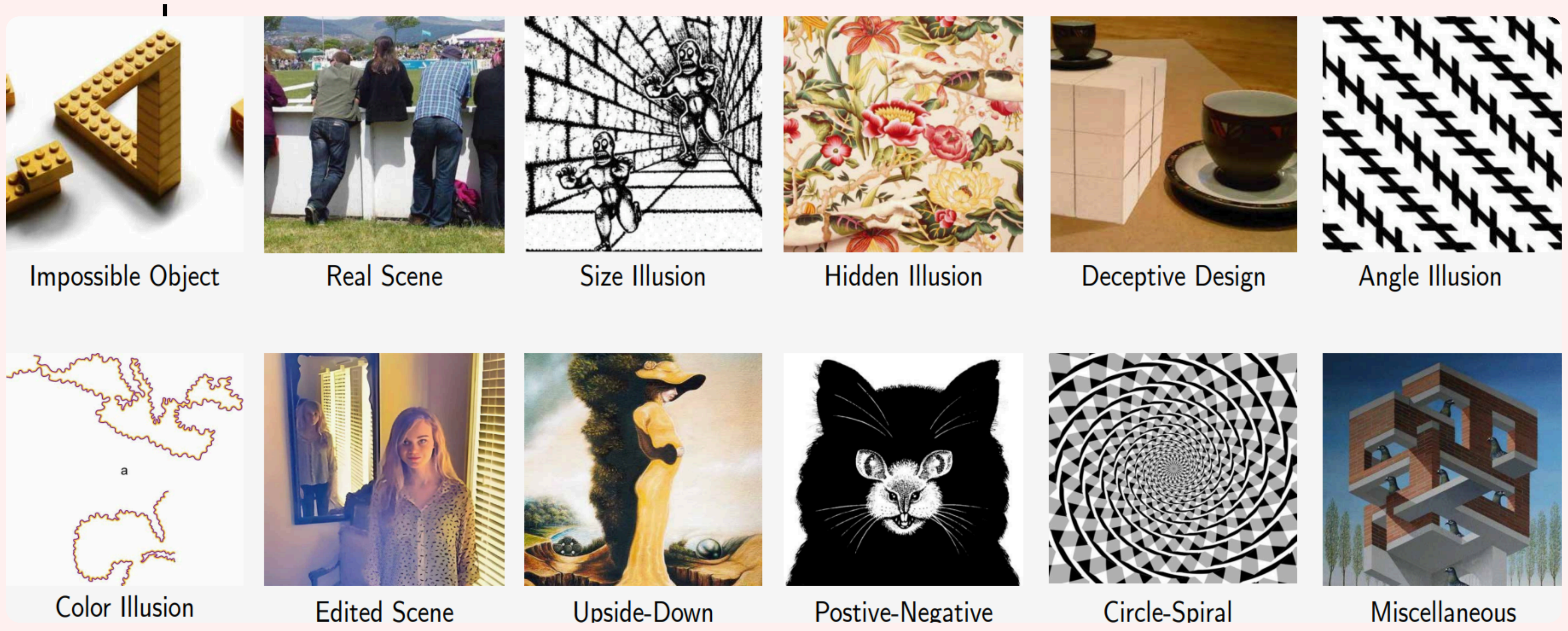
A. Four, the four corner regions

B. Zero, all the regions are white

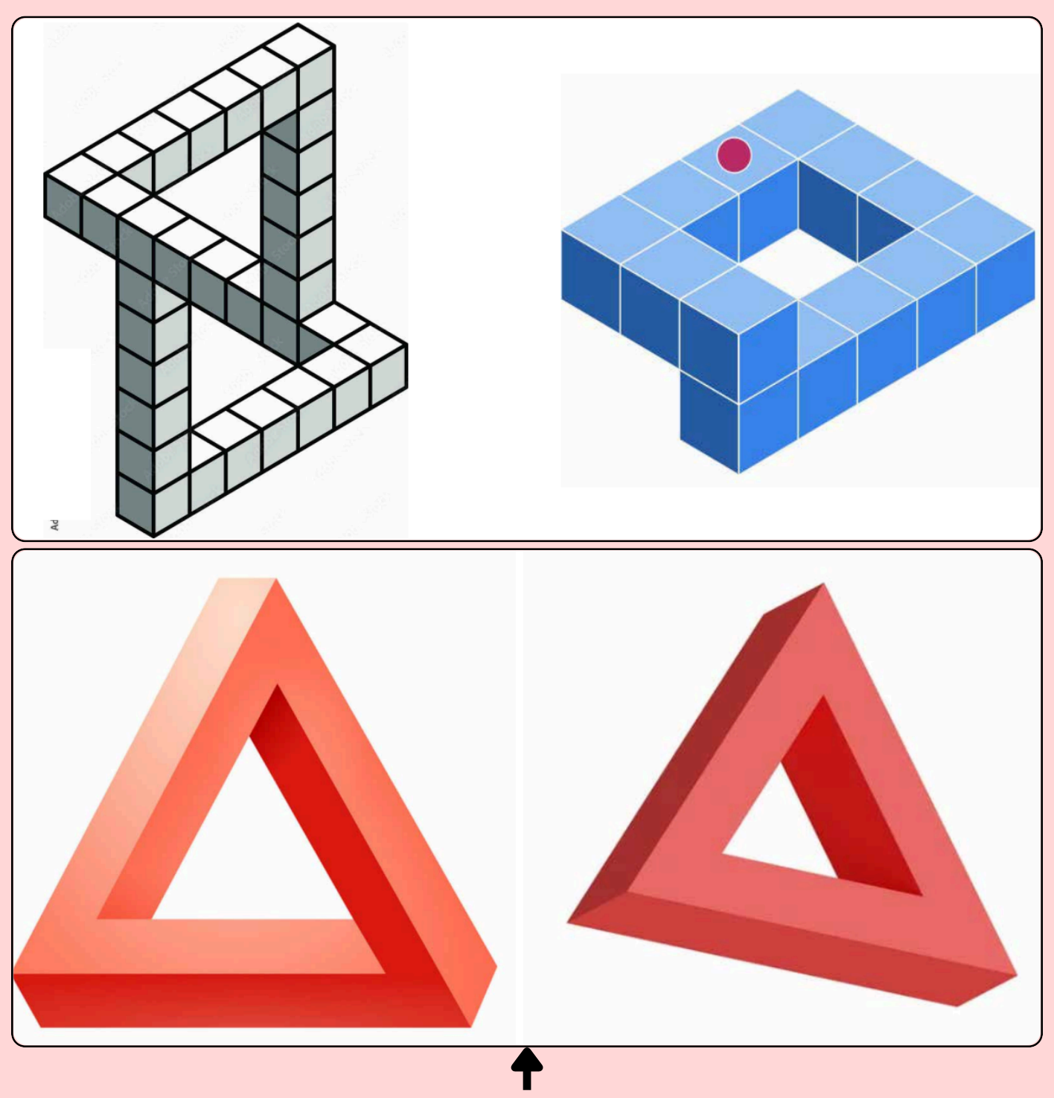
C. Five, the center and the four corner regions

D. One, only the center

E. Nine, all the regions are pale yellow



IllusionVQA-Soft-Localization



Which object is geometrically impossible?

A. Left

B. Right

C. Both

D. Neither

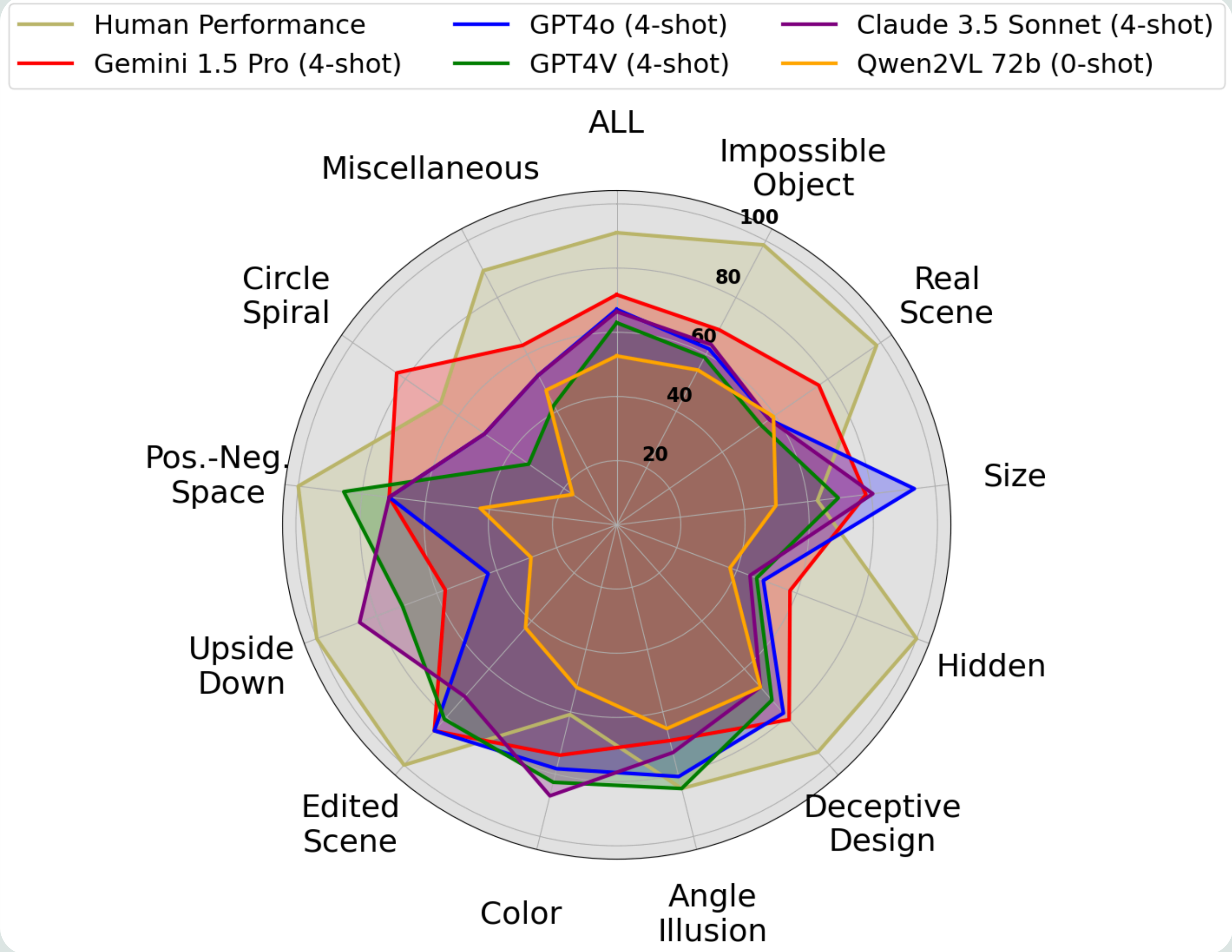
`permute(40 impossible, 20 ordinary)`

IllusionVQA-Soft-Localization:

- 1000 VQA from combining 40 impossible and 20 ordinary geometric objects

IllusionVQA-Comprehension:

- 370 high-quality illusions after filtering 3500 web-scraped images.
- 435 handcrafted VQA pairs.
- Wrong options are adversarially curated (VLM answers, misinterpretations, etc.).



Results

	Comprehension (Acc.)		Soft-Localization (Acc.)		
	0-shot	4-shot	0-shot	4-shot	4-shot+CoT
Human Performance	91.03		100		
Gemini-1.5-Pro	65.98	**71.72** ↑	47.3	**53.8** ↑	50.7 ↓
GPT4o	62.53	67.12 ↑	45	49.1 ↑	53.3 ↑
Claude-3.5-Sonnet	59.08	66.44 ↑	45.9	47.4 ↑	39.5 ↓
Gemini-1.5-Flash	54.02	59.31 ↑	42.2	49.8 ↑	45.9 ↓
Qwen2-VL-72B	52.64	n/a	41.1	n/a	n/a
InternVL2-8B	45.06	n/a	28.3	n/a	n/a
Phi-3.5-V-4.2B	41.38	34.71 ↓	24.9	24.9 -	27 ↑

Key Takeaways:

1. Humans outperform VLMs in illusion comprehension. VLMs are consistently better in only two categories: **Size** and **Color**.
2. Most small, open-source VLMs do not support **interleaved image-text input**. Phi-3.5-V shows inconsistent 4-shot performance.
3. **Text-based** Chain-of-Thought (CoT) reasoning is challenging to do on optical illusions.

Big Ideas:

1. Illusions Are Logical Puzzles: Humans need **15 seconds** to work out each optical illusion while VLMs answer instantaneously. We must move beyond text-based strategies, such as CoT, for VLMs.
2. I am not a robot ✓: Understanding Real Scene, Deceptive Design, and Angle Illusions is crucial for **embodied robotics**. Conversely, soft-localization illusions can serve as a **CAPTCHA** for malicious web bots.