Clinical and Psychological Visual Assessments of Visual Language Models

Gene Tang, Sam Schwarzkopf, and Kate Storrs

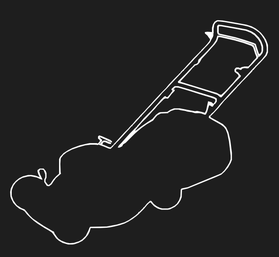
THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

Scan me for more details

Introduction

Visual Language Models (VLMs) have demonstrated remarkable capabilities in visual understanding through human-like interactions. They excel in tasks ranging from basic image recognition across diverse domains to college-level challenges requiring complex image understanding [1-2]. However, emerging evidence suggests that VLMs' perception is still fundamentally different from human perception [3-5]. While VLMs succeed in high-level semantic processing, they often struggle with fine-grained, non-semantic, low- and mid-level visual understanding. A key limitation in VLMs evaluation is the lack of systematic perceptual testing compared to human visual assessments. Neuropsychological and psychophysical batteries were designed to assess specific processes across the visual hierarchy. They provide a robust testing framework alongside human baseline data [6]. The current study leveraged these assessments to examine the similarities and discrepancies between VLMs and human visual perception.

Demo

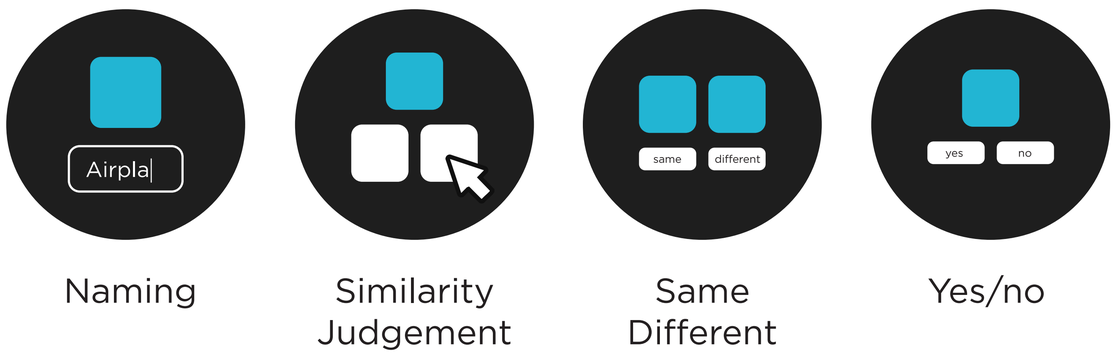


Methods

3 VLMs tested on 51 subtests from 6 batteries.

GPT-4o Claude 3.5 Sonnet (new) Gemini 1.5 pro

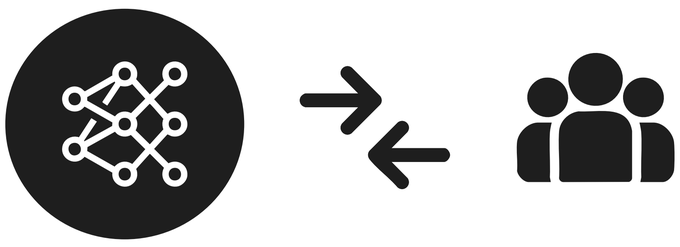
4 Task Types



Naming | Similarity Judgement | Same Different | Yes/no

Prompts were formatted to closely match the original clinical instructions. Model responses were recorded and scored based on clinical criteria for neuropsychological tasks and prior psychological findings for psychophysical tasks.

Performance Evaluations



Clinical tests' normative data

Own data collected via an online study

Discussion

VLMs outperformed humans in simple high-level object naming tasks.

Possible explanation: Extensive semantic representation trained alongside image data.

Performance declines with global image distortions or when dealing with occlusions.

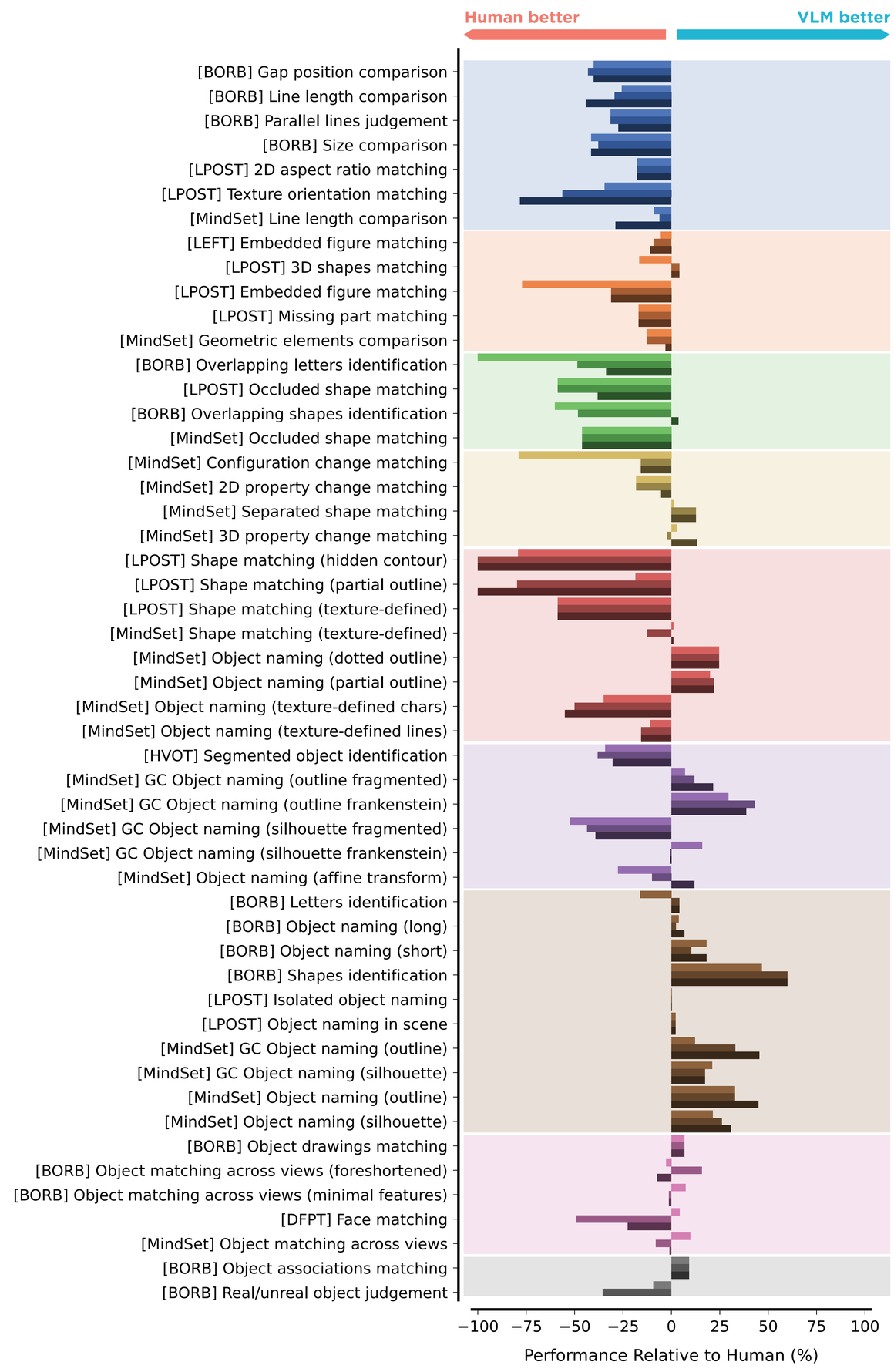
Possible explanation: Overreliance on visual features, but less so on context.

VLMs also struggled with low- and mid-level perceptual tasks, especially when stimuli often lack nameable objects.

Possible explanation: Underrepresented abstract objects/images in the training data result in a limited capacity to semantically represent such stimuli. This impacts their ability to "understand" these images (e.g., matching, or discerning differences).

Results

VLMs Performance Relative to Human



Visual Processes

- Simple image element judgements
- Complex figure judgements
- Occlusion and overlap
- Property Biases
- Robustness across different image cues
- Robustness across different configurations
- Straight forward object recognition
- View invariance
- Semantic association and categorisation

Models GPT-4o Claude 3.5 Sonnet (new) Gemini 1.5 pro