

Massimo Rizzoli*, Simone Alghisi*, Olha Khomyn,
Gabriel Roccabruna, Mahed Mousavi, Giuseppe Riccardi

Signals and Interactive Systems Lab, University of Trento, Italy



Introduction

VLMs have shown **competitive performance** in several vision-language tasks

Results may be misleading: e.g., object classification has been shown to **improve** with **random or spurious non-visual descriptions**

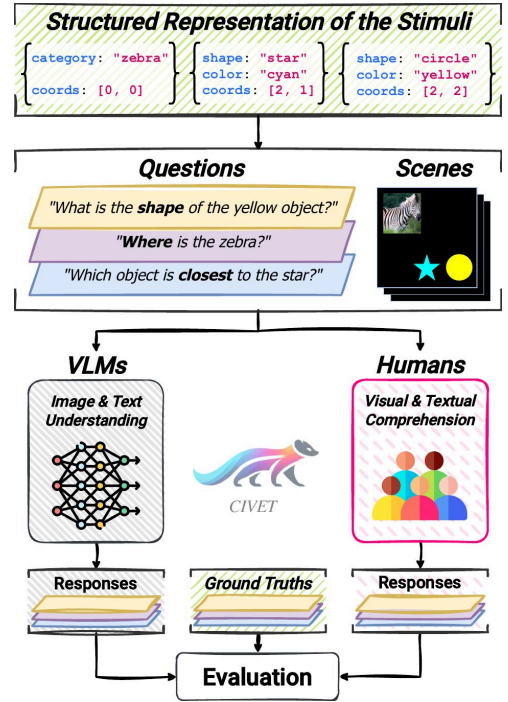
Do VLMs understand or are they exploiting biases in the data?

We propose **CIVET**, a framework for systematic evaluation of understanding

We evaluate 5 sota VLMs in recognizing **basic properties and relations** of objects

We generate **exhaustive** sets of stimuli **free from annotation noise** and **biases**

CIVET Framework



Findings

VLMs have **limited understanding** of visual scenes:

- 1 recognize some properties, but **struggle** to identify **basic relations**
- 2 **performance** heavily **depends** on object **position**
- 3 VLMs **fall short of human-level accuracy**

Can VLMs recognize properties?

VLMs **recognize shape and color**, but **struggle on sheen and position**

Model	Shape	Color	Sheen	Position
Random Baseline	25	17	50	11
LLaVA-NeXT 7B	98	88	50	42
LLaVA-NeXT 13B	97	76	64	47
Molmo-O 7B	100	98	59	62
Qwen2-VL 7B	99	99	60	61
CLIP	95	95	49	14

Can VLMs identify basic relations?

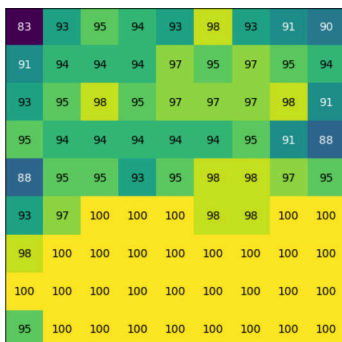
VLMs **struggle** to **identify basic relations** between objects

Model	Relative		
	Position	Distance	Size
Random Baseline	13	50	33
LLaVA-NeXT 7B	24	54	30
LLaVA-NeXT 13B	38	59	33
Molmo-O 7B	24	76	30
Qwen2-VL 7B	46	83	54
CLIP	20	51	49

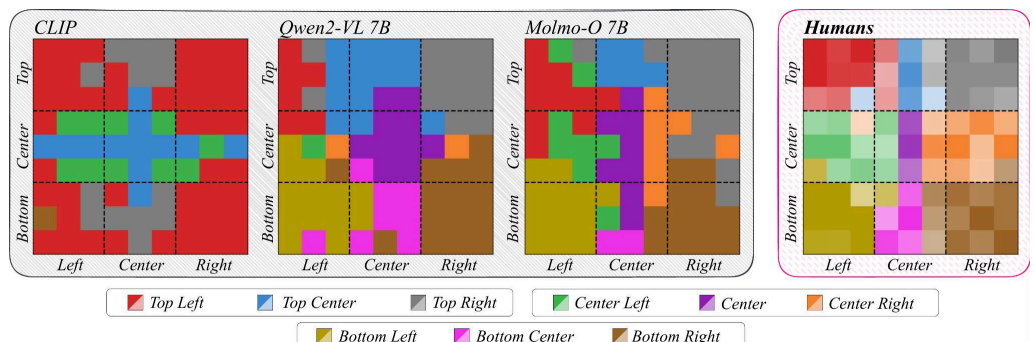
Is their performance robust to variations in object positioning?

Performance **heavily depends** on the **object's position**

VLMs do not reach **human-level accuracy** when asked to **locate** a simple **shape on a black background**



Shape Accuracy at Cell-level for LLaVA-NeXT 13B



VLMs and Humans responses for "Where is the yellow star?"