NumerosityVLM: A Cognitively Inspired Benchmark for Interpreting Numerosity Representations in Vision–Language Models

Yiming Fu[1] · Fangjun Li[1] · Xiujin Liu[2] · Ruidong Ma[5] · Hang Yu[3] · Zhichen Lu[4] · Kanwei He[1] · Alessandro Di Nuovo[5] · Angelo Cangelosi[1] · Zhegong Shangguan[1]*

[1] The University of Manchester [2] University of Michigan [3] Tufts University [4] ENSTA, Institut Polytechnique de Paris [5] Sheffield Hallam University

1. Motivation

While VLMs have shown strong visual understanding capabilities, their numerosity perception remains poorly understood, particularly in terms of error mechanisms and underlying representations.

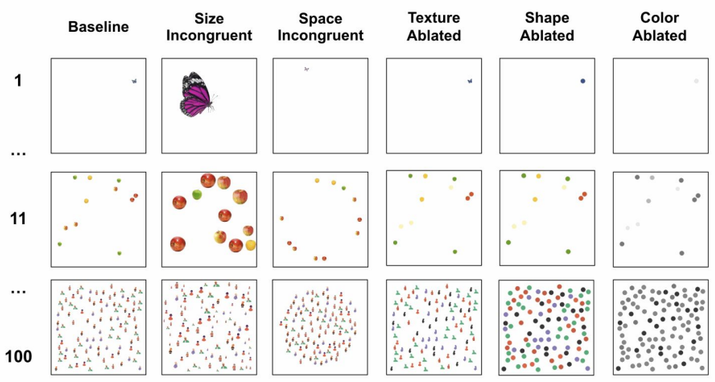
2. Core Contributions

Cognitively Inspired Benchmark: Systematically controls visual factors and progressively removes visual shortcuts to isolate numerosity reasoning in VLMs.

Key Findings: Statistical attribution and representation analysis reveal strong associations between VLM counting performance and model language backbone.

3. Benchmark Design

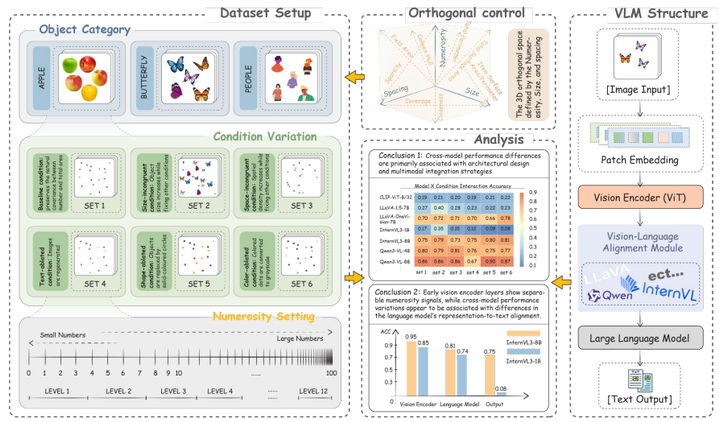
10,800 synthetic images · **6** controlled conditions · **12** numerosity levels 1 → 100 · **3** object categories



Orthogonal Control Conditions:
Baseline fixes ISA; Size-Incongruent fixes TSA; Space-Incongruent fixes hull perimeter.

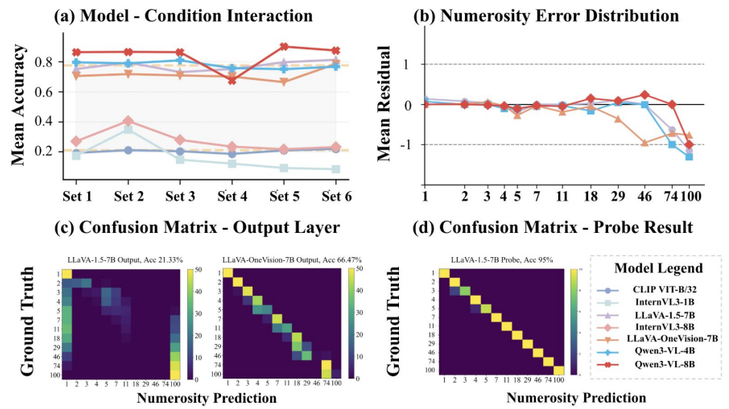
Visual Cue Ablation Conditions:
Baseline-matched coordinates/areas; remove texture -> shape -> color progressively.

4. Evaluation Pipeline



STIMULI 6 conditions · **PROMPTS** open / closed · **7 VLMs** 4 families · **METRICS** ACC + errors · **DIAGNOSIS** bias + probes

6. Systematic Bias Analysis



Visual Bias. Size-Incongruent improves performance in both tiers, while Space-Incongruent shows no significant difference.
Cue Ablation. Removing texture significantly reduces performance in both tiers. Tier 1 then rebounds as shape and color are removed, whereas Tier 2 remains largely unchanged.
Numerosity Bias. Accuracy is generally higher below 20, with systematic underestimation emerging above 50.

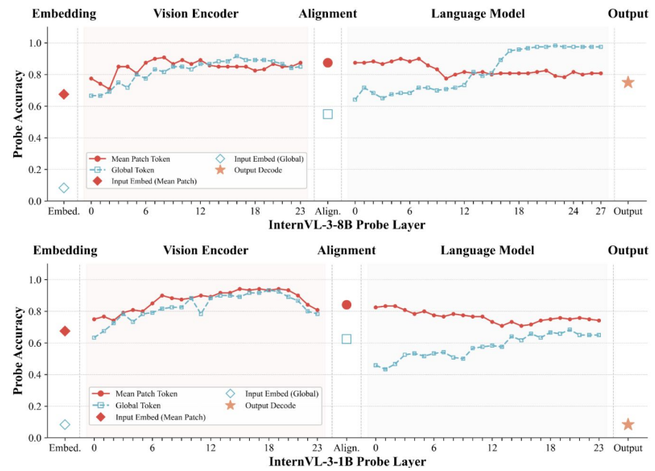
5. Statistical Attribution Analysis

3-way ANOVA: Model architecture is the dominant factor influencing counting performance differences in VLMs.

0.325	**0.005**	**0.001**
Model partial ω²	Condition partial ω²	Subset partial ω²

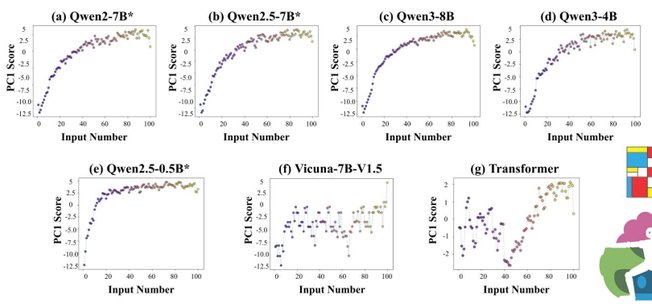
7. Representation Analysis

Layer-wise Analysis of Numerosity Representations



Linearly separable numerosity signals emerge in early visual layers, while cross-tier differences are primarily reflected in the language modeling and decoding stages.

Language-backbone Number Geometry



PCA Result: More ordered, log-compressed number geometry tracks stronger counting performance.

European Research Council
Established by the European Commission
TALK
PRIMI
Performance in Robots Interaction via Mental Imagery

Contact: zhegong.shangguan@manchester.ac.uk