Semantic Richness or Geometric Reasoning? The Fragility of VLM's Visual Invariance

Jason Qiu[1*], Zachary Meurer[1*], Xavier Thomas[1*], Deepti Ghadiyaram[1]

[1]*Boston University,* [*]*equal contribution*

Project page: Test your VLMs here!

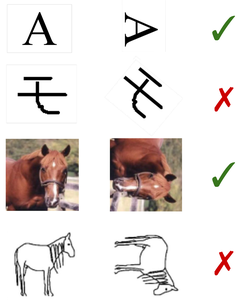
Motivation

*Do **VLMs reason** about these images **the same way**?*

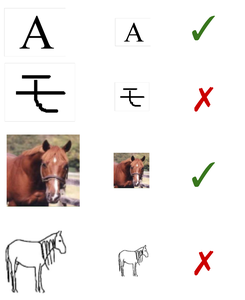


Key finding:** While probing different visual tasks - **semantic richness changes, the spatial question does not, but the performance does.

Rotation: If I rotate the first image, can I get the second image?



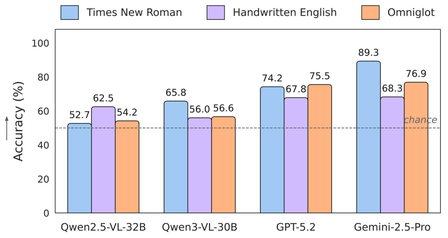
Scale: Compare the two images. Do they show the same character/object?



✓ = VLM got it correct; ✗ = VLM got it wrong

Results

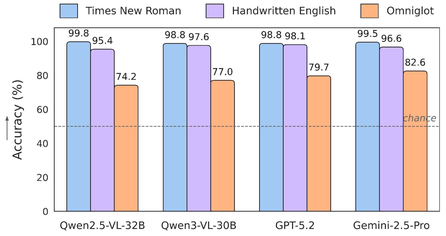
*Can VLMs reason about **rotation** consistently?*



- **No!** Rotation recognition remains challenging for VLMs
- **Familiarity matters:** Gemini-2.5-Pro scores 89.3% on English typed fonts but only 68.3% on handwritten

*Can VLMs reason about **scale** consistently?*

- **No!** VLMs do far better on English characters than Omniglot scripts
- **Even Gemini-2.5-Pro drops ~17 %** from Times New Roman to Omniglot



*Do VLMs favor **semantic** over **spatial** reasoning?*

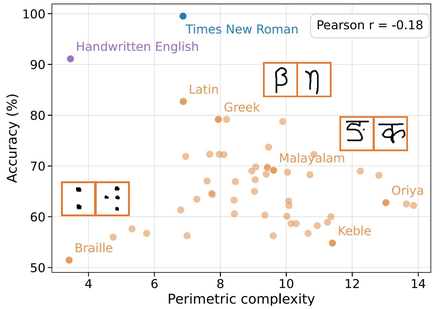


- **Yes!** Near perfect accuracy for the semantic object identification task
- **On the same images, the spatial rotation task performance drops**

Object Identification prompt: Is the object in these images a {class name}?

Analysis

*Does **visual complexity** dictate **performance**?*



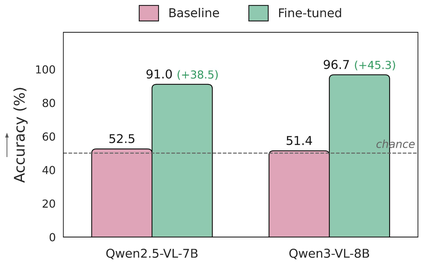
Perimetric Complexity = $\frac{P^2}{4\pi A}$; where P = Perimeter, A = Area

- **No!** We find that perimetric complexity is **only weakly correlated** (r=-0.18) with Qwen2.5-VL-7B performance on the scale task
- Additionally, we find humans score 100% across different scripts! **The difficulty lies in the model, not the task.**

*Role of **Data Domain** and **Semantic Grounding***



- Fine-tune VLM only on describing sketch images
- **Purely semantic**, with no geometric supervision
- Rotation accuracy improves on sketch images, **without a single rotation example in training**



Takeaways 💡 **:**

- **VLMs lean on data familiarity**, not pure visual reasoning
- **Failures persist** across rotation, scale, domains, and model families

Datasets, results, and analysis at xthomasbu.github.io/visual_invariance