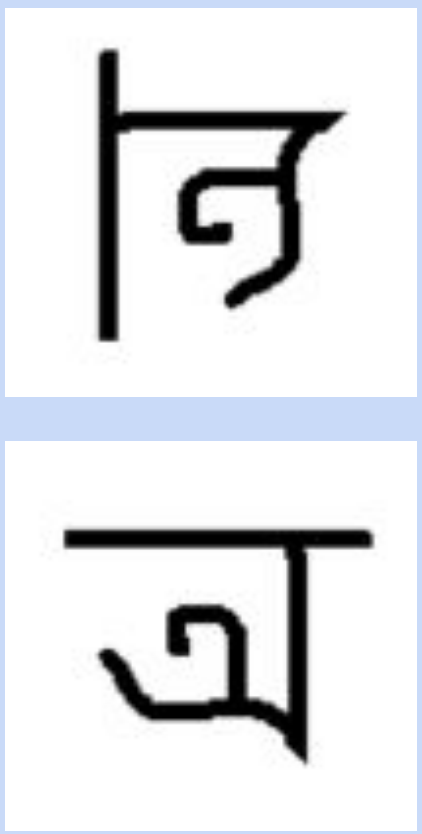
Blind to Shape, Bound to Semantics: A VLM's Dilemma

Zach Meurer[1], Jason Qiu[1], Xavier Thomas[1], Thomas Fel[2], Deepti Ghadiyaram[1]
Boston University[1], Harvard University[2]

Motivation

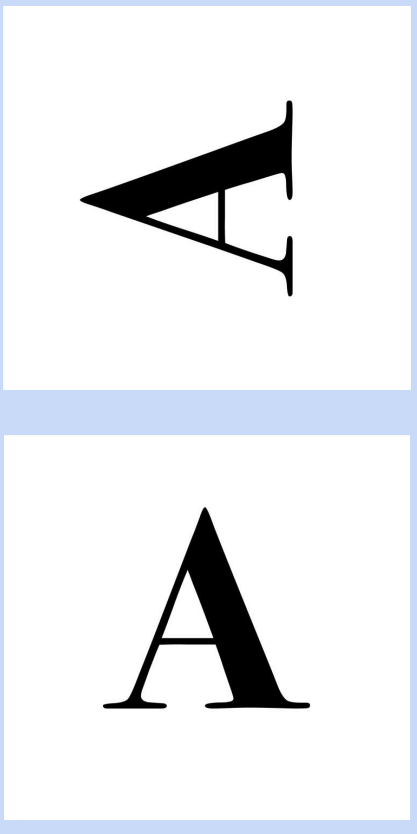
Do Vision Language Models truly understand what they see?

Prompt: Are these the same objects? One could be a rotated version of the other.



Response: No, these are not the same object. While they are…

Prompt: Are these the same objects? One could be a rotated version of the other.



Response: Yes, these are the same object. Based on the…

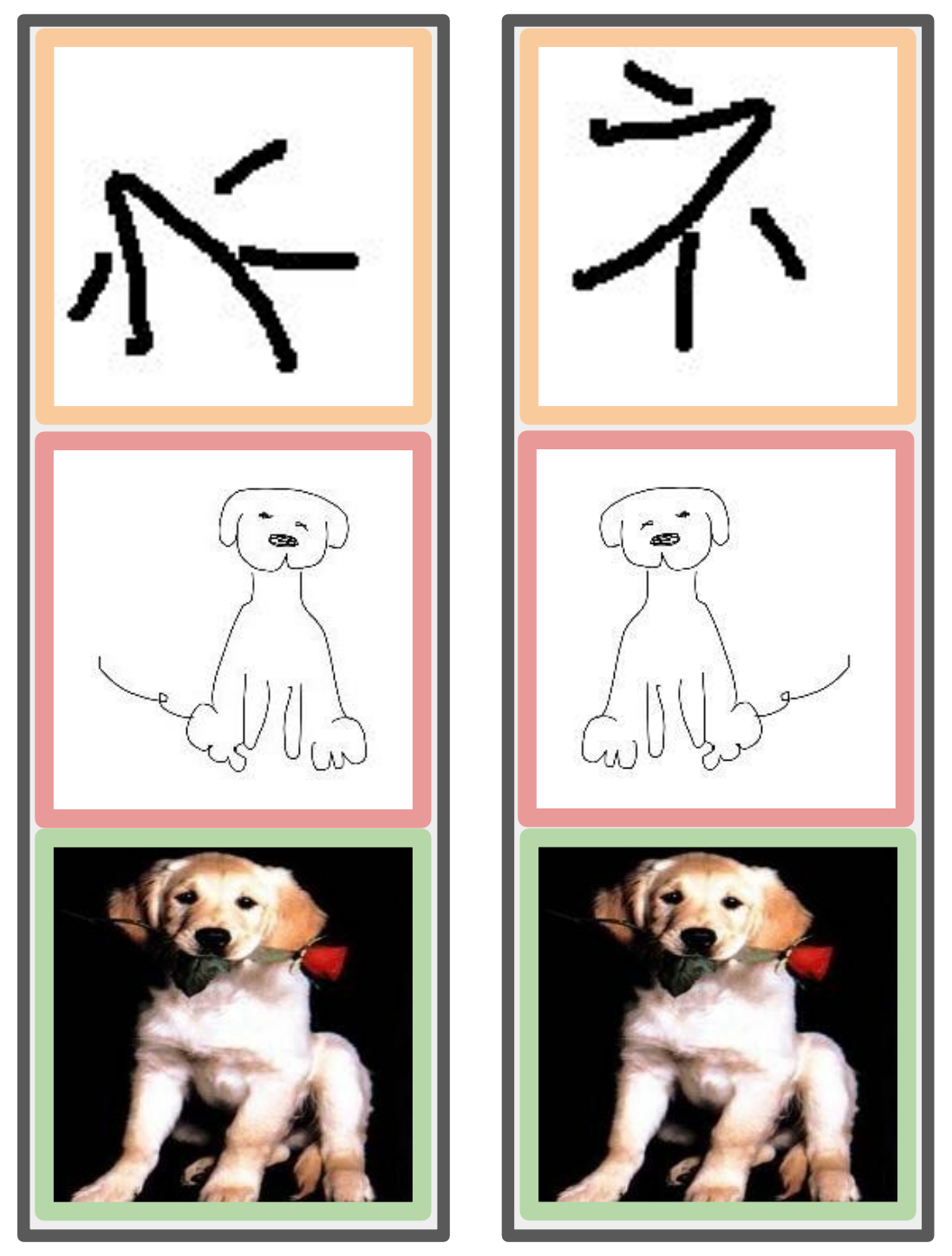
Key Idea

To gauge VLM visual understanding, probe for **transformation recognition** at scale across varying levels of semantics

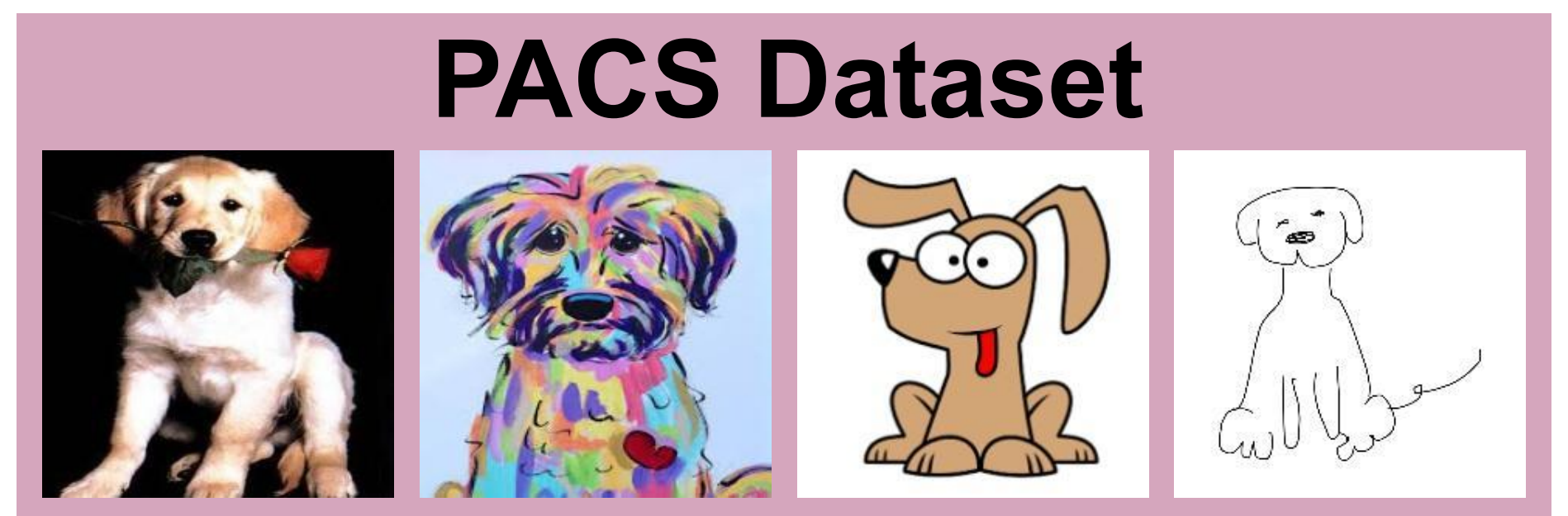
Rotation: "If I rotate the first image, can I get the second image?"

Reflection: "If I horizontally flip the first image, can I get the second image?"

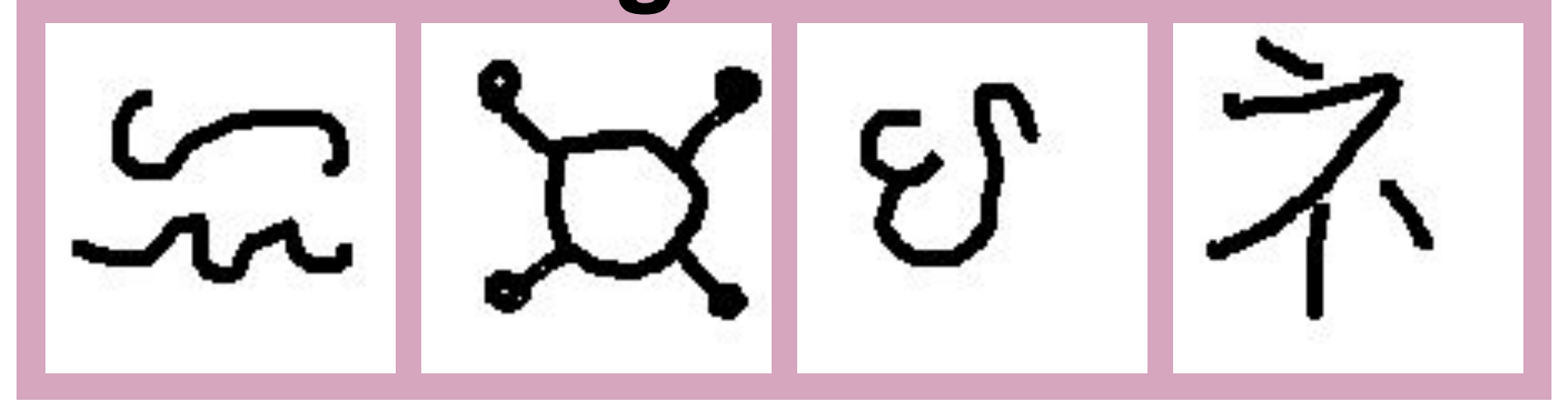
Identity: "Are these two images the same?"



PACS Dataset



Omniglot Dataset

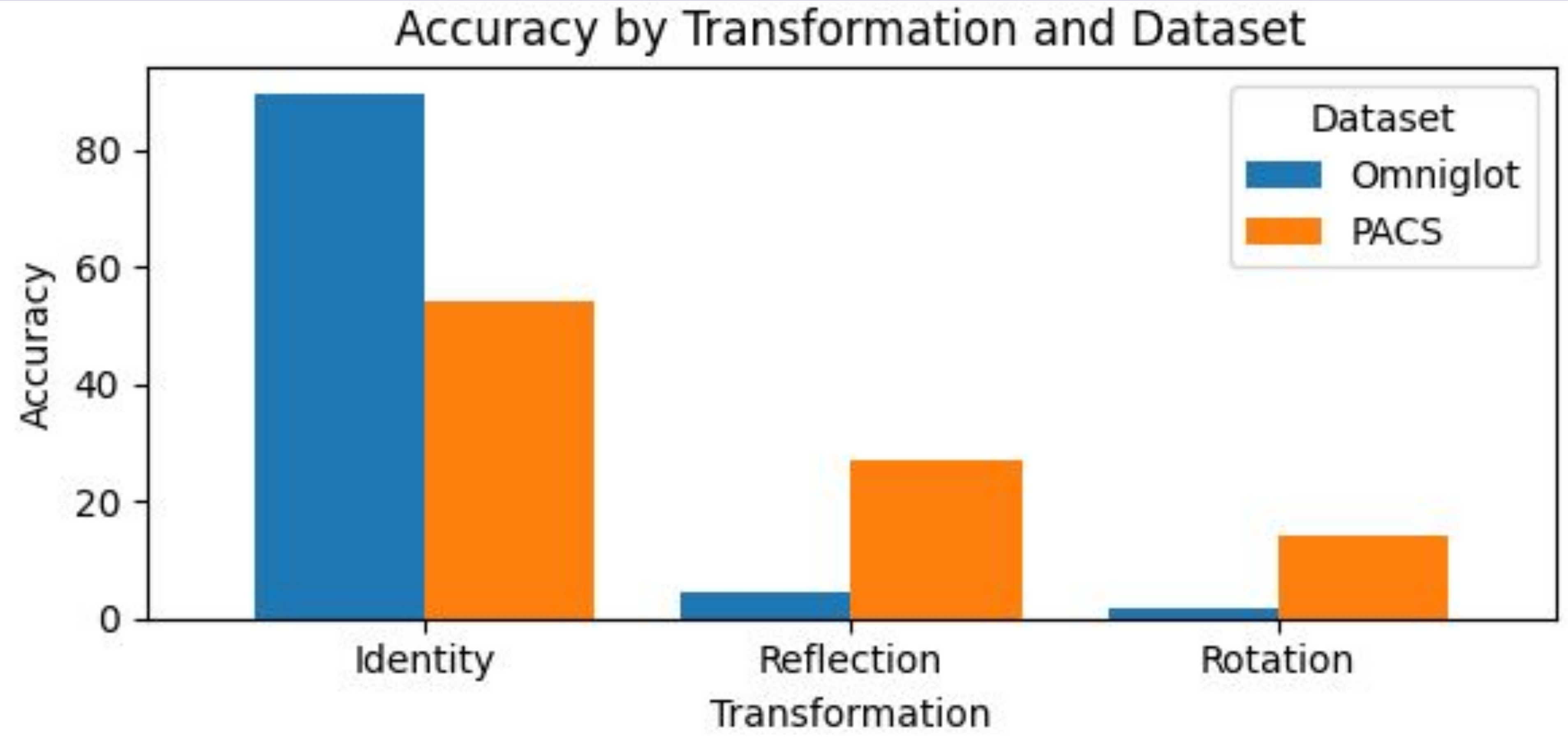


Yes or **No**

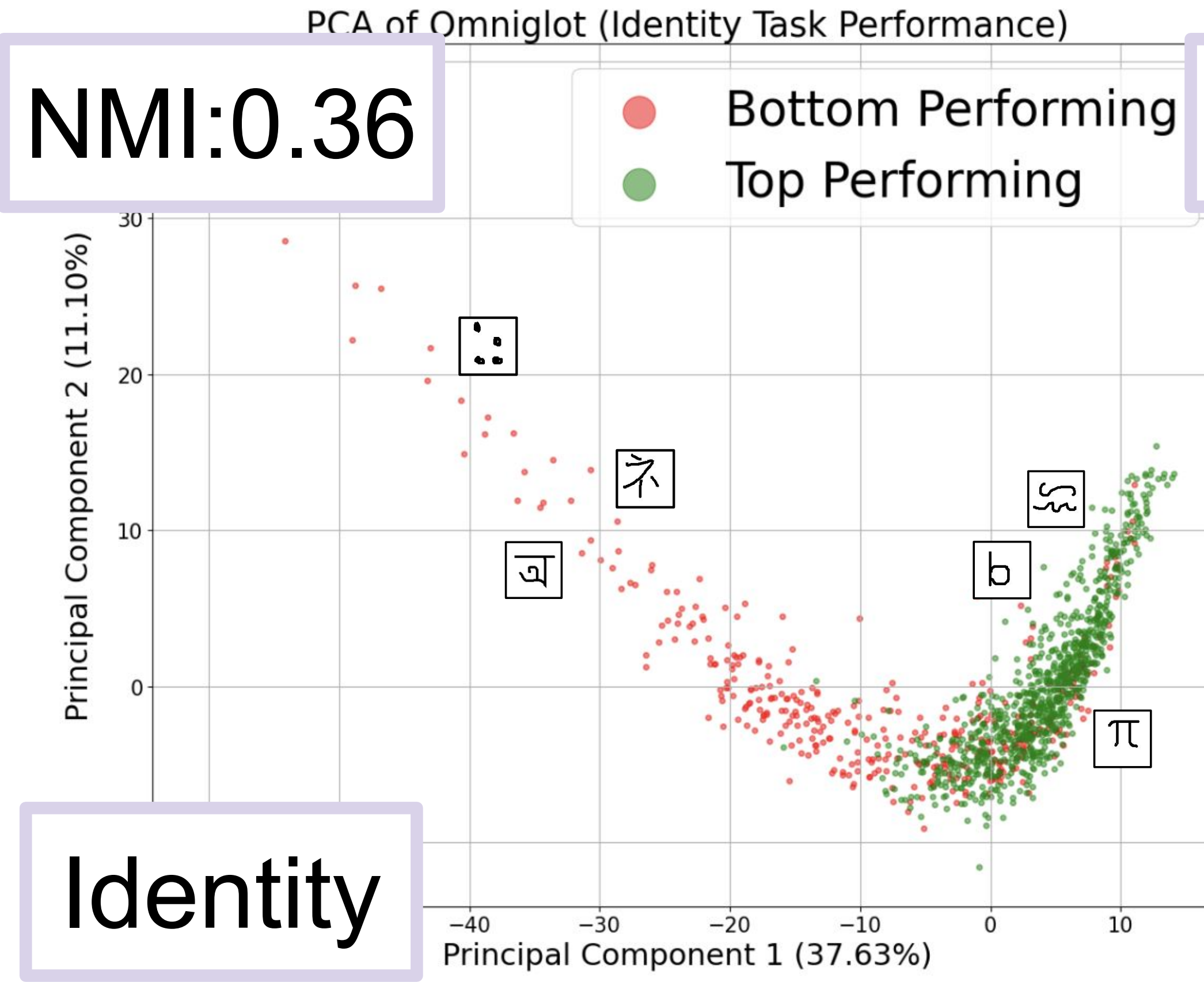
Semantic Richness

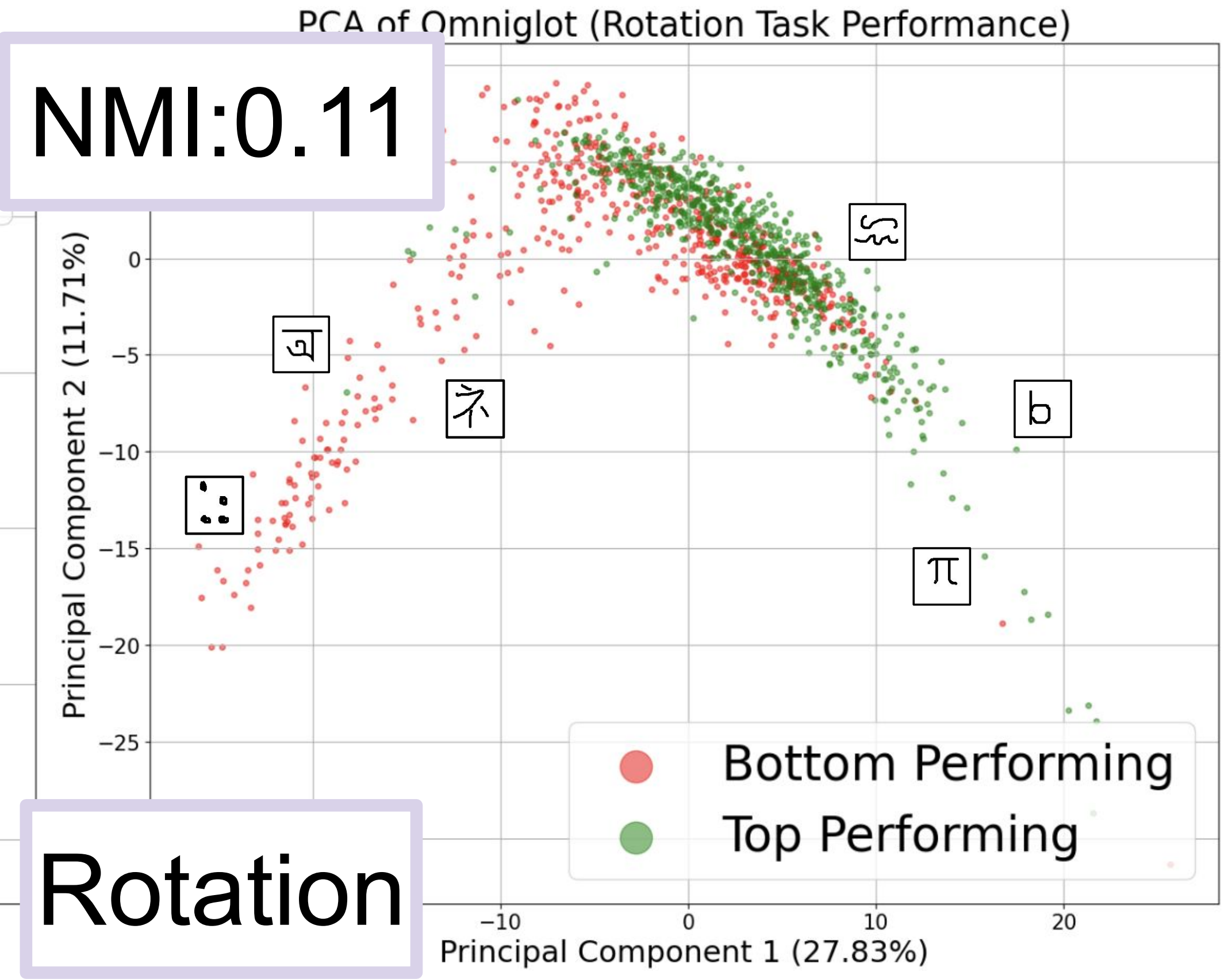
Results

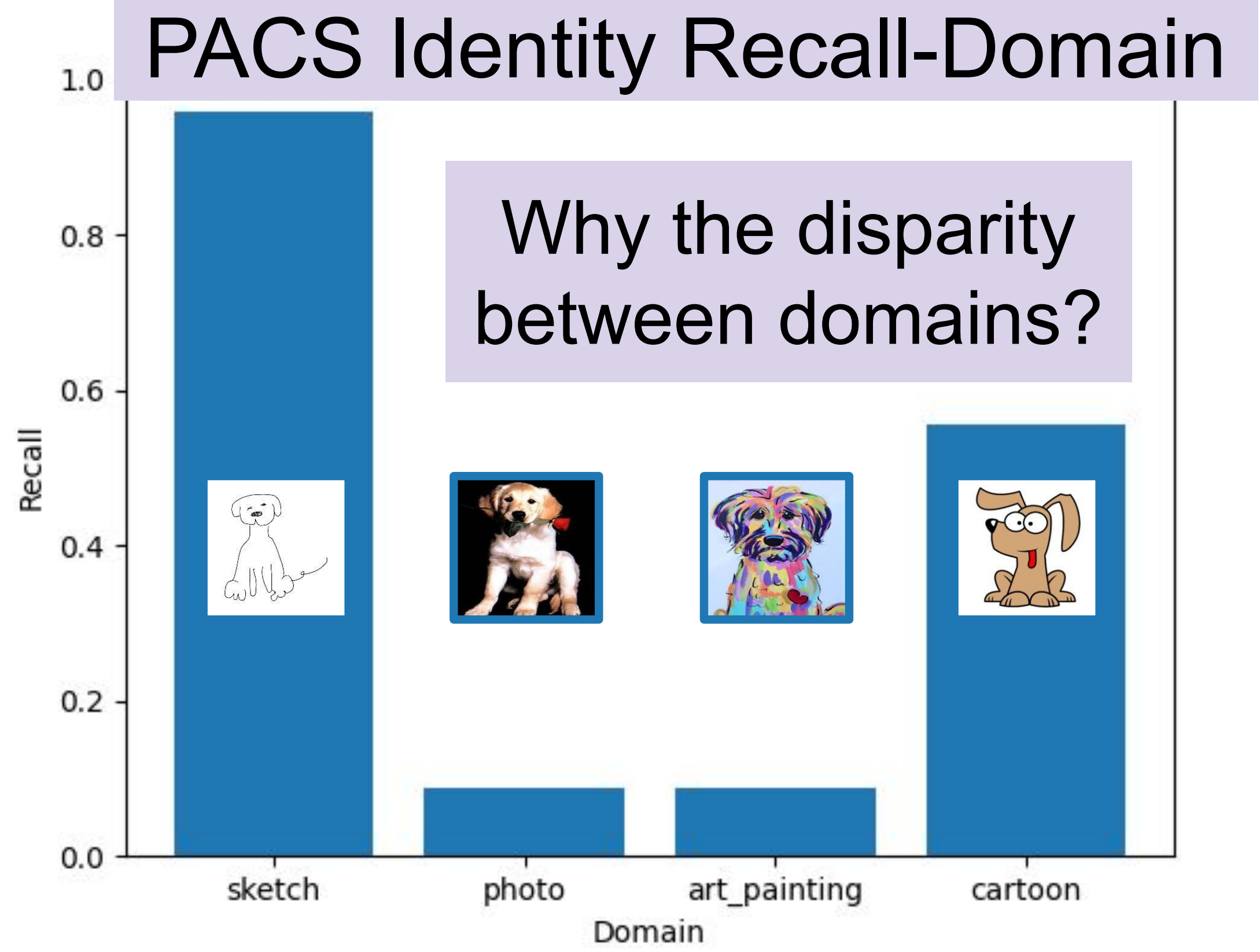
Qwen2.5-VL, a SOTA open-source VLM, often fails at identifying simple visual transformations (or lack thereof)



Qwen2.5-VL encodings of Omniglot characters form clusters based on performance at transformation identification tasks







Why the disparity between domains?

Key Findings:

1) Qwen2.5-VL fails at identifying simple image transformations
2) These failures vary by level of semantics
3) Qwen2.5VL encodes transformations of successes and failures differently