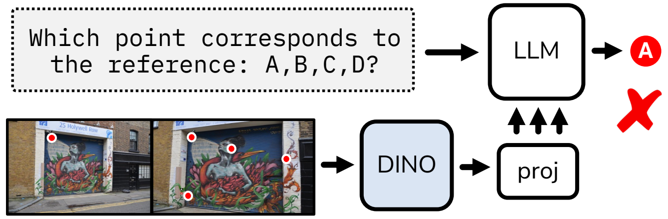
Hidden in plain sight: VLMs overlook their visual representations

Stephanie Fu Tyler Bonnen Devin Guillory Trevor Darrell

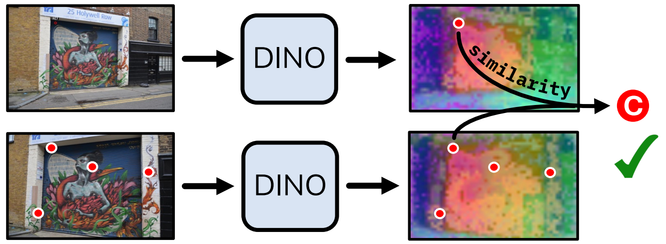
hidden-plain-sight.github.io

Berkeley
UNIVERSITY OF CALIFORNIA

VLMs often perform at chance-level on vision-centric tasks…



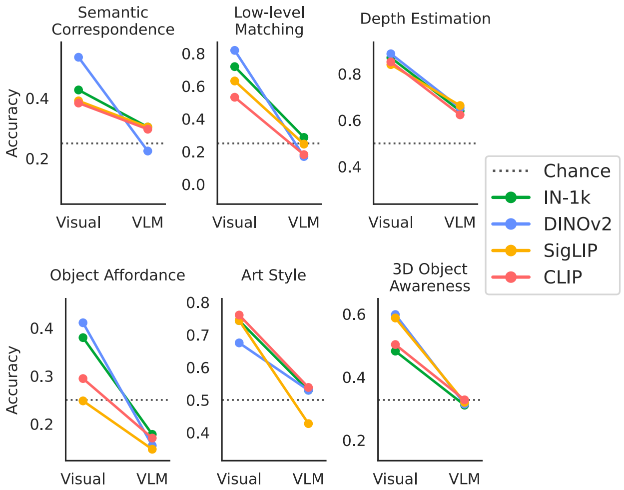
…even though their vision encoders have the right representations!



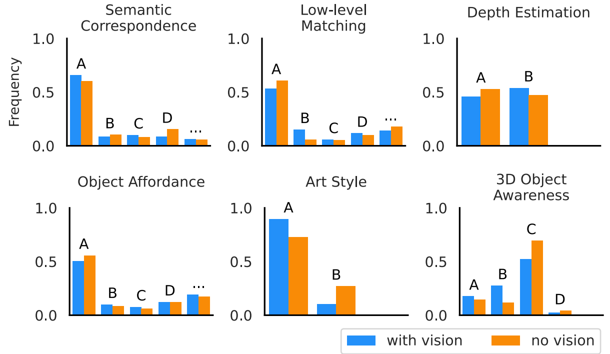
This performance drop is persistent across models and tasks.

Let's investigate this phenomenon!

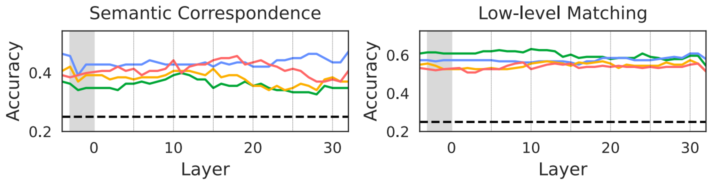
Our testbed: 6 vision-centric tasks



VLM multiple-choice answers also reflect blind baselines.

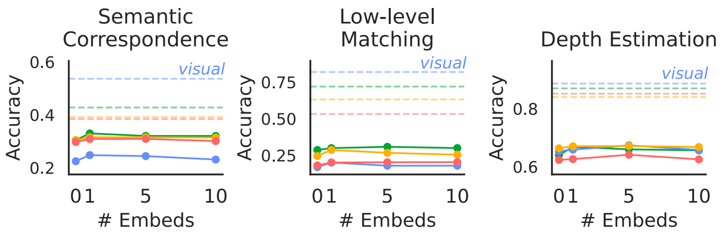


Hypothesis 1: Vision representations degrade throughout the VLM.



Not exactly. We probe vision representations at every layer and get similar accuracy as the vision model.

Hypothesis 2: The VLM is prompt-sensitive.

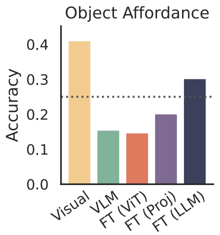


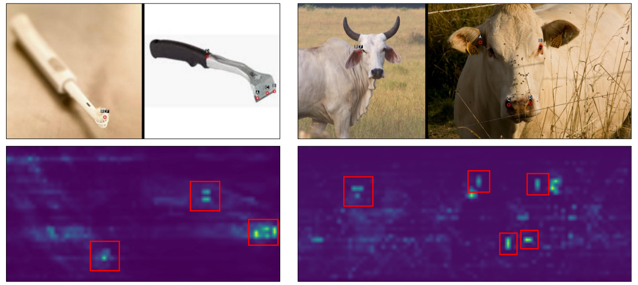
Prompt-tuning with prefix embeddings helps some, but **is not the answer.**

Hypothesis 3: The LLM underutilizes its vision representations.

We fine-tune each VLM module and find that the **LLM** has the most potential for:

- Closing the accuracy gap
- Mitigating language priors
- Improving attention over images





Difference between LLM-tuned and original attention maps

The vision representations in VLMs can be powerful, but are often **hidden in plain sight!**