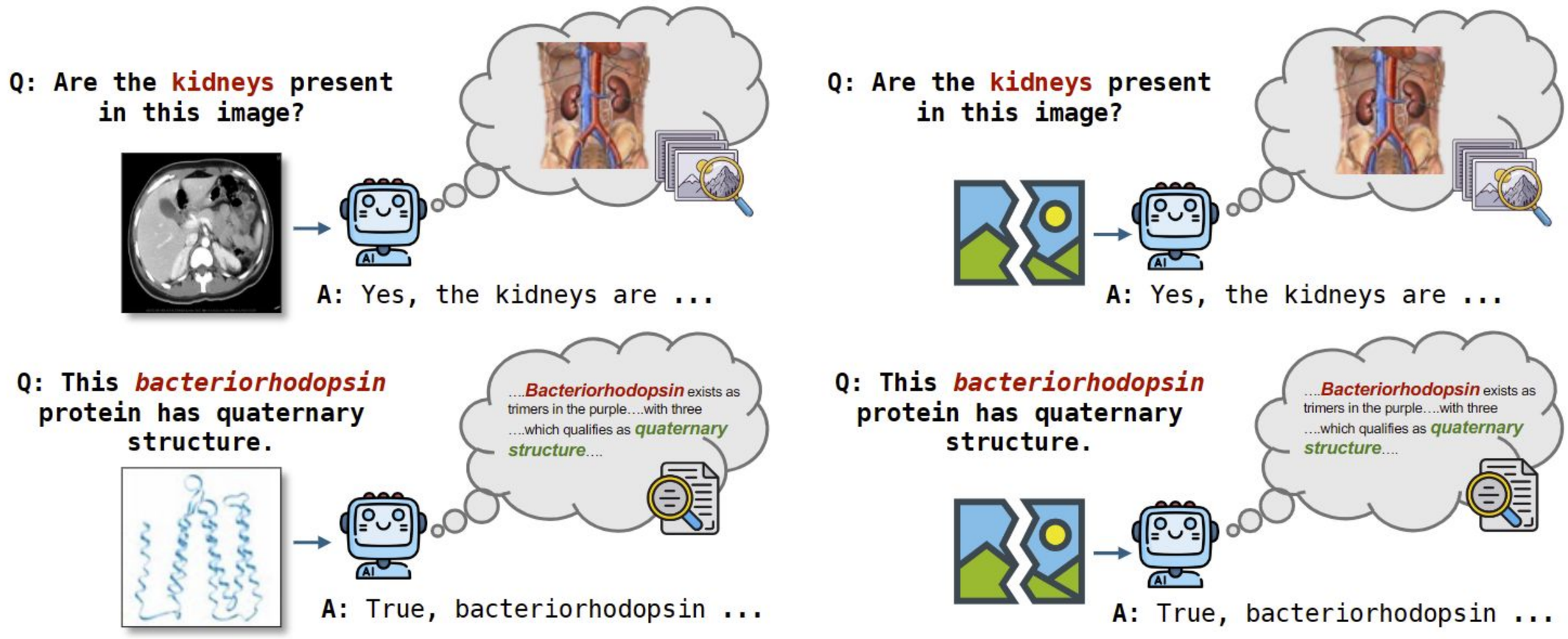
# VLMs Fake Visual Understanding Through Two Distinct Mechanisms




COLUMBIA UNIVERSITY

NYU

***Daniel Ben-Levi*** *Judah Goldfeder Weiliang Zhao Raz Lapid Amit LeVi*
*Allen G Roush Ravid Shwartz-Ziv Hod Lipson*

## Mirage Probes: How Vision Models Fake Visual Understanding

## 1 TL;DR

MAIN 3 TAKEAWAYS

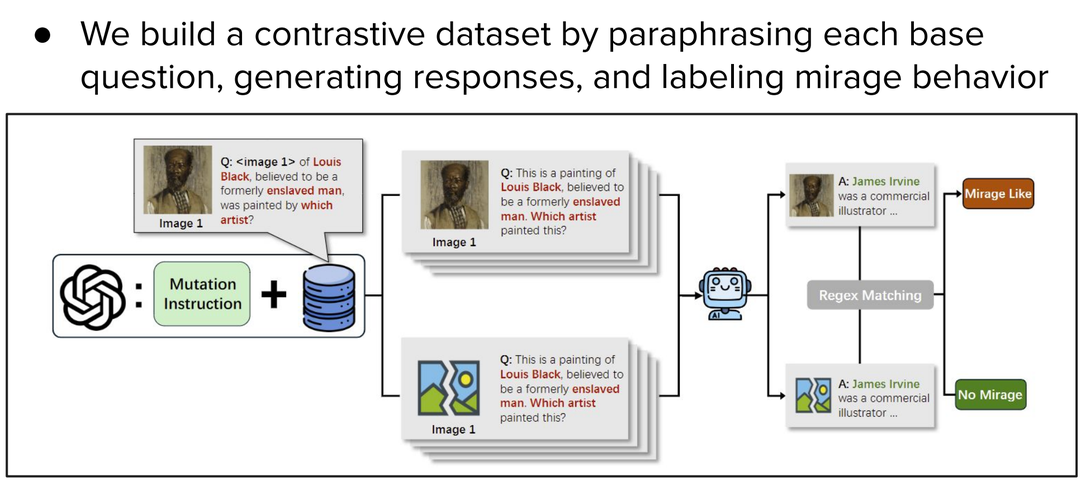
- VLMs provide **seemingly-grounded answers** to VQA queries **without incorporating** present **visual evidence**
- This image-present "mirage" behavior is **linearly decodable** from model latent space
- VLMs seem to exhibit **two distinct types of mirage behavior:** *spurious images* and *textual biases*

## 2 True Visual Grounding is Crucial

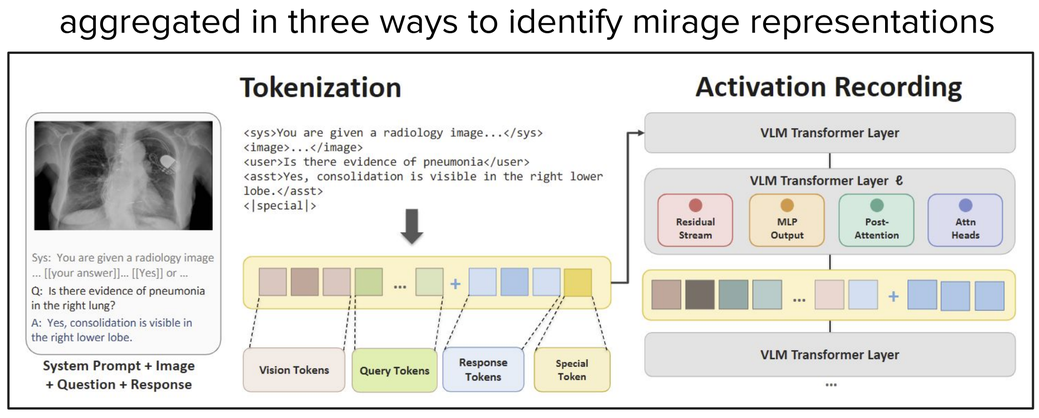
- VLMs increasingly used in real-world tasks like VQA, doc. analysis, scientific reasoning, and medical diagnostics
- Good high-stakes answers require true visual grounding
- Asadi et al. (2026) show VLMs can provide confident, image-descriptive reasoning even when no image is present
- Are these "mirage" mechanisms present even with images?

## 3 Contrastive Pairs and Probing

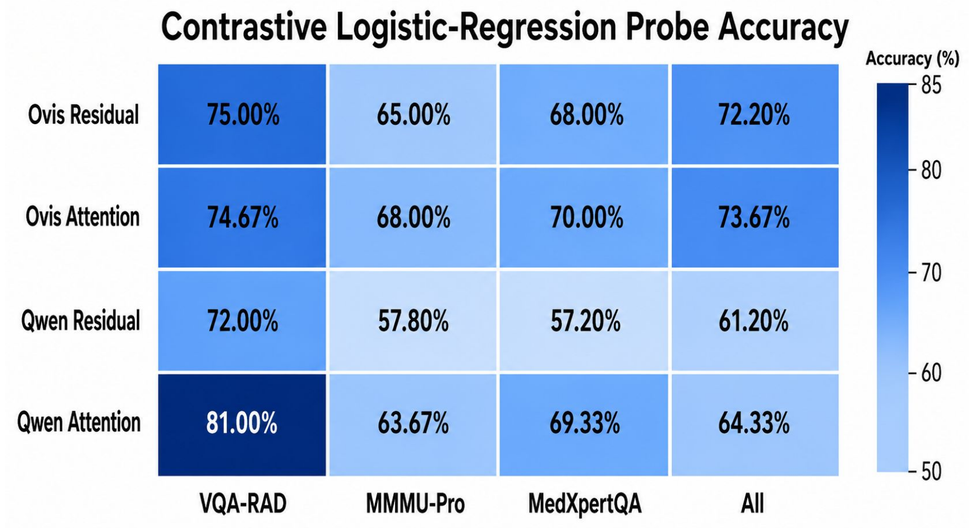
- We build a contrastive dataset by paraphrasing each base question, generating responses, and labeling mirage behavior




- We test four probe families on activations from four VLM sites aggregated in three ways to identify mirage representations




## 4 Image-Present Mirage: Two Mechanisms




- Mirage behavior is **linearly decodable** from **image-present** latent space. A Naive Bayes baseline **rules out surface lexical confounds**.

| VQA-RAD | | | MMMU-Pro | | |
|---|---|---|---|---|---|
| | + Mirage | - Mirage | | + Mirage | - Mirage |
| Image-Free | 60 | 3 | Image-Free | 836 | 18 |
| Image-Reliant | 2909 | 236 | Image-Reliant | 1139 | 1421 |

- Mirage behavior seems to be enabled by **two different underlying mechanisms**: *spurious images* and *textual biases*

| Dataset | Mirage Corr. | Mean PHI |
|---|---|---|
| VQA-RAD | -0.36 | -0.41 |
| MMMU-Pro | -0.23 | 1.20 |
| MedXpertQA | 0.14 | 0.30 |

## 5 Directions for Future Work

- Text-distribution cleaning targets textual-bias mirages but is unlikely to address spurious-image mirages. Novel interventions are needed to mitigate mirage behavior in VLMs.
- Novel methods for causal validation of our probe-identified directions and/or two mechanism hypothesis would be valuable for future mechanistic work in this area.

PAPER

CODE

*We sincerely thank the NSF AI Institute in Dynamic Systems (Grant #2112085) for funding our work.*



**db3651@columbia.edu**