Bigger Is Not Better

Inverse Scaling and Arbitration Failure in Counterfactual Visual Grounding

Fabian Grob[1,4] **Sanghwan Kim**[1,2,3] **Cordelia Schmid**[5,6] **Zeynep Akata**[1,2,3,4]

1 2 Munich Center for Machine Learning 3 HELMHOLTZ MUNICH 4 relAI 5 Google 6 Inria

Multimodal LLMs follow language priors over contradictory visual evidence, and scaling makes it worse, not better.

Background: Counterfactual Visual Grounding

Original Image Counterfactual Image

Q: How many legs this animal have?

GT: 5 **Biased:** 4

Q: How many points of the star are there on the logo of this car?

GT: 4 **Biased:** 3

Counterfactual visual grounding tests whether a model follows contradictory visual evidence, such as a horse with five legs, or falls back on a learned prior.

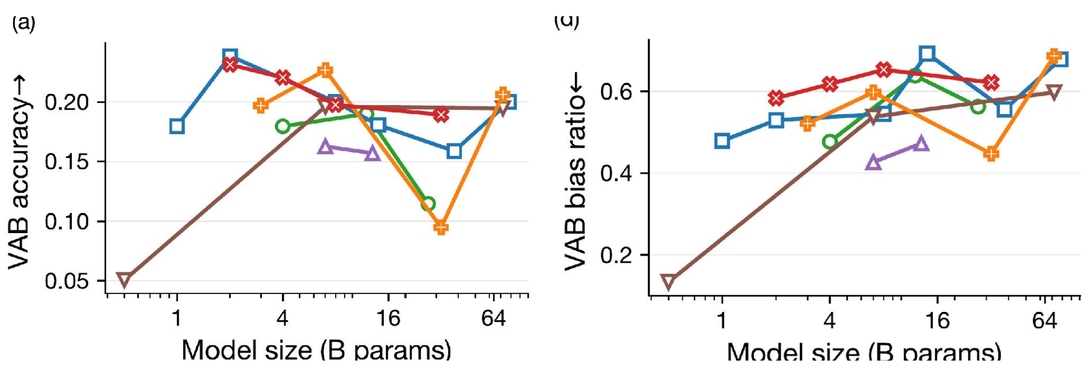
We ask three questions:

- **When** does the failure worsen?
- **Why** does it arise?
- **How** can we fix it?

The Problem, and Why It Happens

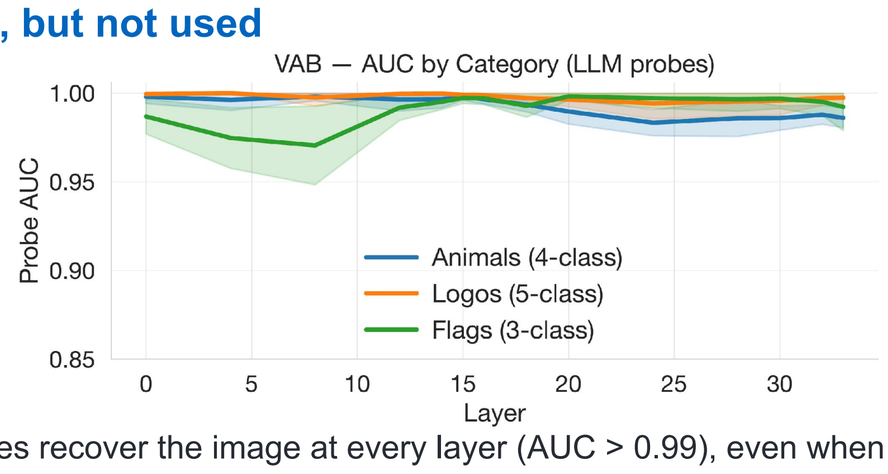
Scaling does not resolve the failure, and mechanistic analysis reveals why.

Scale makes it worse



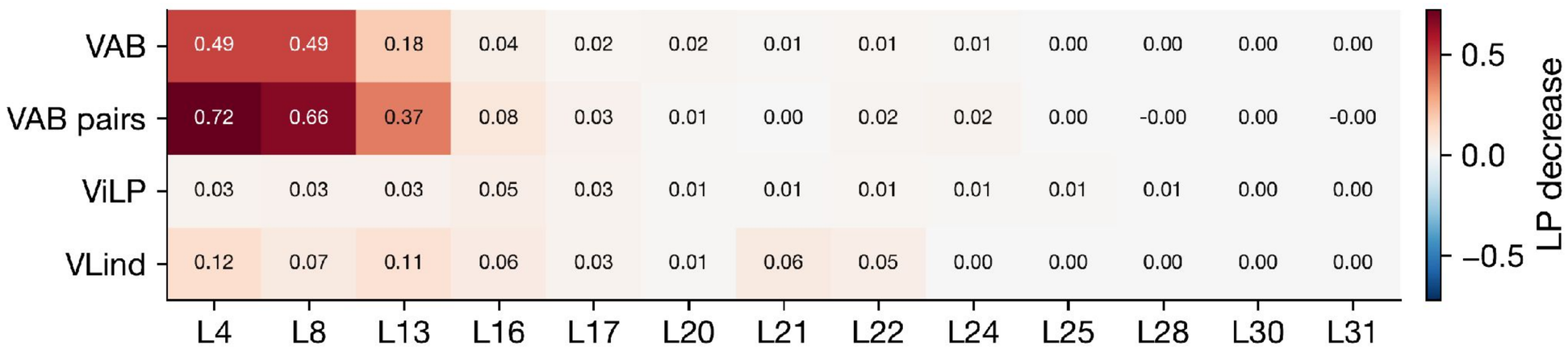
Accuracy stays flat while bias toward priors climbs with size (VAB; same trend on ViLP across 6 families).

Encoded, but not used



Linear probes recover the image at every layer (AUC > 0.99), even when the model answers wrong.

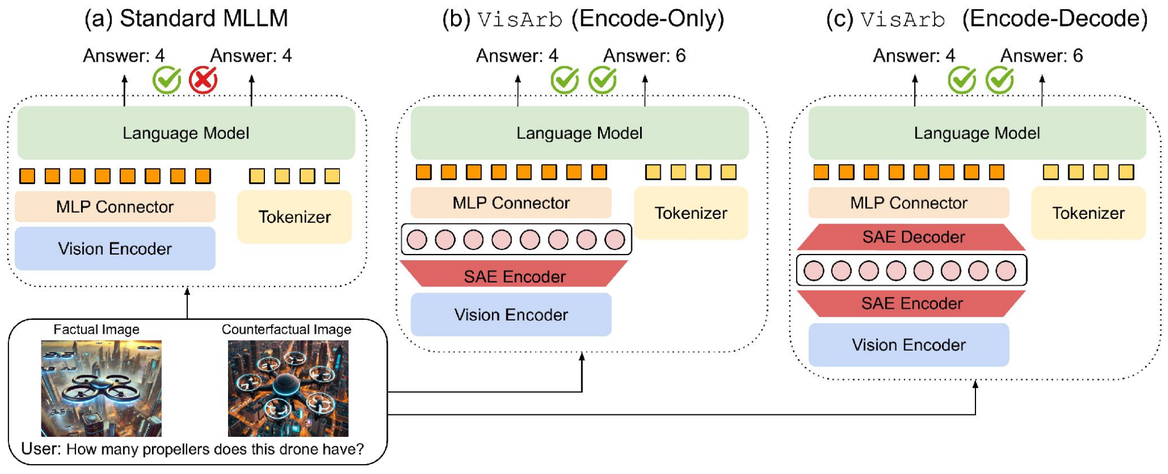
The prior commits early, before vision arrives



Concept-vector steering cuts language-prior reliance only in early layers 4 to 8; vision integrates at layers 14 to 19, too late to change the answer.

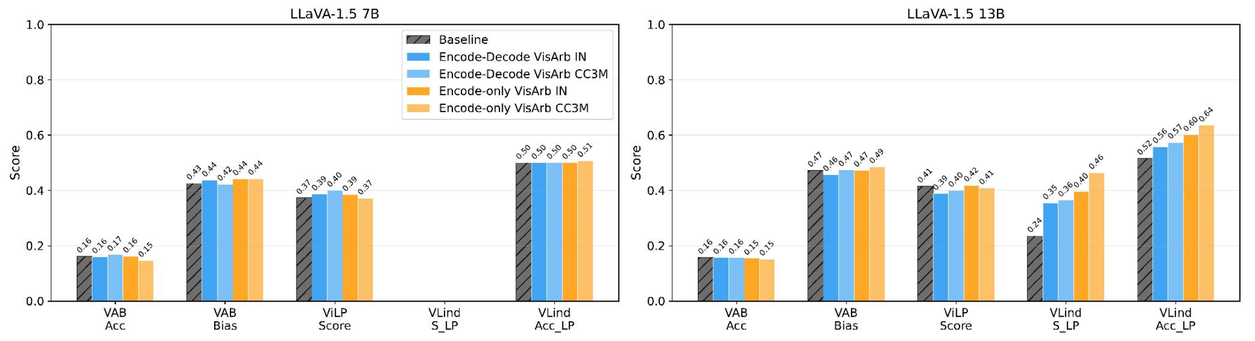
The Fix: VisArb, and Results

The Sparse Autoencoder bottleneck



A frozen SAE between the vision backbone and connector exposes concept-aligned visual evidence from the first LLM layer.

Results (LLaVA-1.5)



S_LP: language-prior override score · Acc_LP: accuracy under counterfactual conditions

+22 pp prior override **+12 pp** accuracy (VLind, 13B)

VAB, ViLP and standard VQA stay unchanged; the gain comes from improved conflict recognition (+32 pp).

Takeaways

- Larger models commit to language priors before they read the image.
- The failure is arbitration, not perception: visual evidence is encoded but unused.
- VisArb fixes it at the vision-to-language interface, with no loss on standard VQA.

Paper & code

Mechanistic Interpretability Workshop @ ICML 2026

fabian.grob@tum.de

DAAD Zuse Schools Konrad Zuse Schools of Excellence in Artificial Intelligence · Federal Ministry of Research, Technology and Space