Where Reliability Lives in Vision-Language Models

A Mechanistic Study of Attention, Hidden States, and Causal Circuits

Logan Mann, Ajit Saravanan, Ishan Dave, Shikhar Shiromani, Saadullah Ismail, Yi Xia, Emily Huang

Algoverse AI Research · loganmann@ucsb.edu

ICML 2026 Mechanistic Interpretability Workshop

Accepted · Virtual Poster · Non-archival

Code & eval scripts
github.com/
itsloganmann
/VLM-Reliability-Probe

Vision-language models give confident, fluent answers that are wrong. To deploy them where errors matter, we need reliability signals that are both predictive and mechanistically meaningful. The intuitive guess is **visual attention**: if the model looks at the right region, its answer should be trustworthy. We call this the **Attention-Confidence Assumption**, and we test it directly across three model families.

THE QUESTION **Does the spatial structure of visual attention predict whether a VLM is right, or is correctness encoded later, in the residual stream?**

Method: VLM Reliability Probe

A cross-family pipeline that extracts, quantifies, localizes, and intervenes on internal signals tied to answer correctness, across three classes of evidence:

STRUCTURAL — attention-map entropy & cluster count

MECHANISTIC — logit-lens truth margin, hidden-state probes

BEHAVIORAL — self-consistency, token confidence

Three architectures, one harness:

LLaVA-1.5-7B **PaliGemma-3B** **Qwen2-VL-7B**

Evaluation: POPE (1,000), LLaVA-Bench (90), VQA v2, TextVQA. Correctness scored with correlation and AUROC, with sampled confidence intervals.

Where the signal actually lives

The strongest reliability signals emerge **later in the computation**, not in attention geometry:

R = 0.429 **Self-consistency** (majority-vote support over K=10 samples) is the strongest behavioral predictor of a correct answer.

AUROC > 0.95 **Hidden-state probes** give the best single-pass readout in our strongest settings.

Reliability is feature processing, not routing

Logit-lens analysis of the **truth margin** (correct minus top-incorrect logit) shows where reliability mechanically emerges.

82.1% of truth-margin growth at the peak comes from **MLP layers**, not token routing (attention).

Attention does not predict reliability

On the pooled 3,090-sample structural set, attention-structure metrics are **statistically indistinguishable from noise** ($p > 0.05$):

R = 0.001 cluster count vs. correctness
95% CI [-0.034, 0.036]

R = -0.012 spatial entropy vs. correctness
95% CI [-0.047, 0.024]

An XGBoost / Random-Forest ensemble on 11 attention-derived features reaches only **52–55%** accuracy (near chance); a supervised attention probe tops out at **AUROC 0.725**.

Architecture divergence

- **LLaVA** — late, fragile bottleneck (integration peaks near L31).
- **PaliGemma** — integrates early (peak L14).
- **Qwen2-VL** — cyclical, distributed integration.

"Symbolic Detachment" is an **architectural choice, not a universal law**.

...yet attention is causally necessary

Masking the top 30% of attended patches sharply degrades accuracy ($p < 0.001$): attention **enables feature extraction but does not encode uncertainty**.

LLaVA-1.5 accuracy drop **-8.2 pp**

PaliGemma accuracy drop **-11.3 pp**

Sparse, localized reliability circuits

L_1-regularized probes localize reliability to a handful of neurons. Causal ablation shows a **distributed circuit, not a single bottleneck**:

- Single-neuron ablation: **0.0 pp** change ($p = 1.00$).
- Ablating all top-5 jointly: **-2.0 pp** overall, **-8.3 pp** on object identification.

Takeaway

In current VLMs, reliability is better understood through **hidden-state geometry, layer-wise margin dynamics, and causal circuits** than through attention-map sharpness alone. Practical upshot: build reliability monitors on hidden states and self-consistency, not on how sharply the model attends.