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

Abstract

Vision-language models can produce confident, fluent mistakes, but it is still unclear where their internal reliability signal actually lives. A natural hypothesis is that reliability should be visible in visual attention: sharper focus on the relevant region should imply a more trustworthy answer. We test this hypothesis with *VLM Reliability Probe (VRP)*, a cross-family study of LLaVA-1.5, PaliGemma, and Qwen2-VL that compares three classes of evidence: attention-map structure, generation dynamics, and hidden-state mechanisms. Our main claim is that attention structure is a poor reliability readout even when attention remains causally important for feature extraction: across the pooled structural-analysis set, cluster count and spatial entropy are nearly uncorrelated with correctness ($R(C_k, y) = 0.001$, $R(H_s, y) = -0.012$). Instead, the strongest reliability signals emerge later in the computation. Self-consistency is the strongest behavioral predictor we measure ($R = 0.429$), while hidden-state probes provide the best single-pass signal (AUROC > 0.95 in our strongest settings). We further find a mechanistic split across model families: LLaVA exhibits early locking and a fragile late bottleneck, whereas PaliGemma and Qwen2-VL distribute reliability more broadly and remain robust under large interventions. The takeaway is narrow but important: in current VLMs, reliability is better understood through hidden-state geometry, layer-wise margin dynamics, and causal circuits than through attention-map sharpness alone.

1 Algoverse AI Research. Correspondence to: Logan Mann <loganmann@ucsb.edu>.

Accepted at the ICML 2026 Mechanistic Interpretability Workshop (non-archival).

1. Introduction

Vision-language models (VLMs) can answer complex questions about images, but they still produce fluent, confident mistakes. If we want to deploy them in settings where errors matter, we need reliability signals that are both predictive and mechanistically meaningful. For mechanistic interpretability, this leads to a specific question: where is answer reliability represented inside a VLM, and how tightly is that representation coupled to visual grounding?

A natural hypothesis is that reliability should be visible in attention. Attention maps are easy to visualize, and in both NLP and multimodal work they are often treated as evidence for what the model used to decide (Jain & Wallace, 2019). This motivates what we call the *Attention-Confidence Assumption*: if a VLM focuses sharply on the relevant region, its answer should be more trustworthy; if attention is diffuse, the model should be less reliable. This is an intuitive story, but it is stronger than the claim that attention matters for computation. A model can attend to the right region and still answer incorrectly, or answer correctly from broader scene information without sharply localized attention.

This paper tests that assumption directly. We introduce *VLM Reliability Probe (VRP)*, a cross-family mechanistic analysis of LLaVA-1.5, PaliGemma, and Qwen2-VL (Liu et al., 2023; Beyer et al., 2024; Wang et al., 2024). We compare three classes of signals: *structural* metrics from visual attention maps, *behavioral* metrics from generation dynamics, and *mechanistic* signals from hidden states. The goal is not to ask whether attention is useful in general, but whether the *structure* of visual attention predicts correctness, and whether the strongest reliability signal instead emerges later in the residual stream.

Our central claim is simple: attention structure is a weak reliability readout. Across the pooled structural-analysis set, cluster count and spatial entropy are nearly uncorrelated with correctness ($R = 0.001$ and $R = -0.012$), despite attention remaining causally necessary for extracting visual features. The strongest signals emerge later. Self-consistency is the strongest behavioral predictor of truth that we measure ($R = 0.429$), and hidden-state probes provide the strongest single-pass readout (AUROC > 0.95 in

our strongest settings). Layer-wise logit-lens analysis localizes where correct and incorrect trajectories diverge; sparse probes identify compact reliability-associated units; and causal ablations show that reliability is organized differently across model families. In particular, LLaVA relies on a fragile late bottleneck, whereas PaliGemma and Qwen2-VL distribute reliability more broadly and remain robust under larger interventions.

The contribution is therefore not the generic claim that attention can be unfaithful (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), or that self-consistency can help (Wang et al., 2022). Rather, we make a narrower mechanistic contribution: we map where reliability first becomes legible, compare competing internal readouts of correctness, and show that the causal organization of reliability differs sharply across VLM families. This is useful both scientifically and practically. Scientifically, it clarifies where correctness is encoded rather than merely routed. Practically, it suggests that monitors based on hidden states or consistency are more promising than monitors based on attention-map sharpness alone.[1]

Contributions.

- We test a concrete falsifiable hypothesis: whether spatial attention structure is predictive of answer correctness across three VLM families.
- We show that attention structure is near-random as a reliability signal, while hidden-state probes and self-consistency provide substantially stronger evidence.
- We localize where reliability separates in layer-wise trajectories and identify family-level differences between fragile late bottlenecks and distributed reliability circuits.
- We provide causal evidence that these internal signals are not merely correlational: targeted interventions affect reliability differently across model families.

Reproducibility. Released code and evaluation scripts are available at `https://github.com/itsloganmann/VLM-Reliability-Probe` (prompts, split definitions, and probe training pipeline).

2. Related Work

Large VLMs build on contrastive and encoder-decoder vision–language pretraining combined with strong language backbones, enabling instruction following and open-ended multimodal generation (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023; Dai et al., 2023). Their fluency, however, makes reliability hard to judge: models can produce confident answers that are weakly grounded in the image. This concern has motivated benchmark-driven work on hallucination and multimodal evaluation, including object hallucination studies and benchmarks such as POPE, LLaVA-Bench, MME, SEED-Bench, and MM-Vet (Rohrbach et al., 2018; Li et al., 2023b; Zhou et al., 2023; Fu et al., 2023; Li et al., 2023a; Yu et al., 2023). These benchmarks establish where models fail, but they do not by themselves identify where failure-relevant information is represented inside the model.

A second line of work studies whether attention is a faithful explanation of model behavior. In NLP, (Jain & Wallace, 2019) argue that standard attention should not be treated as explanation, while (Wiegreffe & Pinter, 2019) argue for a more nuanced, test-based view. For VLMs, recent evidence similarly shows that correct localization and correct answering can come apart (Liu et al., 2025). This motivates our focus on a narrower and more mechanistic question: whether the *spatial structure* of visual attention is informative about answer correctness, or whether correctness is encoded later in hidden-state dynamics that attention maps fail to expose.

Related work also examines where visual evidence enters model representations and how reliability can be read from model behavior. Long et al. analyze hidden trajectories with and without images, identifying a Visual Integration Point and proposing Total Visual Integration as a representation-level measure of visual influence (Long et al., 2025). On the decoding side, self-consistency aggregates agreement across sampled reasoning paths (Wang et al., 2022), while complementary work studies behavioral reliability and how model behavior shifts across evaluation settings (Chaudhury & Shiromani, 2025; Shiromani et al., 2026; Thomas et al., 2026; Sahay et al., 2025). Our work combines these perspectives in a more explicitly mechanistic pipeline: we compare attention structure, layer-wise hidden-state readouts, sparse unit-level probes, and causal interventions within one cross-family analysis of VLM reliability.

3. Methodology

We introduce *VLM Reliability Probe (VRP)*, a mechanistic analysis pipeline designed to extract, quantify, localize, and intervene on internal signals associated with answer correctness (Figure 1). Our primary investigative goal is to disentangle two competing hypotheses regarding VLM reliability:

1. **The Structural Hypothesis:** Reliability is grounded in the spatial coherence of the visual encoder's attention (i.e., how the model "looks").
2. **The Consistency Hypothesis:** Reliability is a prod-

[1] Throughout, we use *VLM* as the default term; *MFM* and *LVLM* appear only when matching prior-work phrasing.

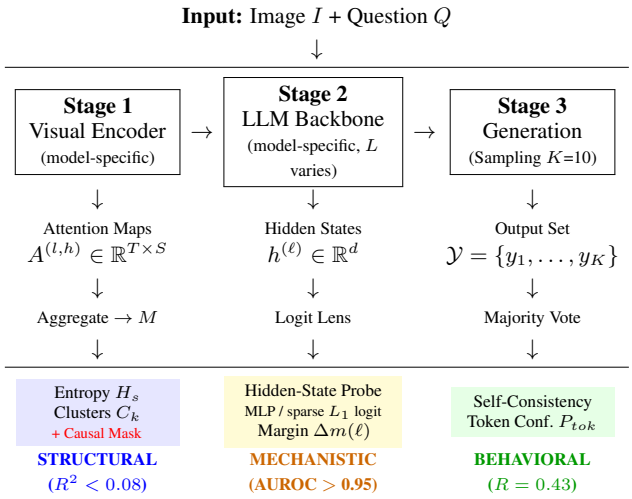

Figure 1. **VLM Reliability Probe (VRP) Framework.** We instrument three computational stages: *Stage 1* extracts cross-attention maps from the visual encoder, yielding Structural metrics (entropy H_s, clusters C_k); we aggregate $A^{(l,h)}$ by averaging over heads and answer-token positions to form one per-layer spatial vector in $\mathbb{R}^S$. *Stage 2* probes hidden states via logit lens plus dense MLP and sparse L_1-logistic probe variants, providing Mechanistic signals; *Stage 3* samples K=10 outputs for Behavioral metrics (self-consistency). Key finding: Structural metrics fail ($R^2 < 0.08$), while Mechanistic probes succeed (AUROC > 0.95). Red indicates causal intervention points.

uct of the generation dynamics and latent linguistic stability (i.e., how the model "speaks").

3.1. Method Summary

We instrument each VLM with forward hooks to record visual attention tensors $A^{(l,h)} \in \mathbb{R}^{T\times S}$ and hidden states $h^{(\ell)} \in \mathbb{R}^d$ during answer generation, then convert these signals into the three metric families shown in Figure 1. Mechanistically, the pipeline is designed to answer four questions: which internal representations separate correct from incorrect trajectories, when that separation emerges, how sparse that signal is, and whether perturbing the identified units causally changes performance.

- **Stage 1 (Structural):** for each layer l, we average $A^{(l,h)}$ over heads and answer-token positions to obtain a single spatial vector $m^{(l)} \in \mathbb{R}^S$ over image patches. After normalizing over patches, we compute spatial entropy

$$H_s^{(l)} = -\sum_{s=1}^{S} \tilde{m}_s^{(l)} \log \tilde{m}_s^{(l)},$$

which measures how concentrated or diffuse the attention is. To measure spatial fragmentation, we retain the top-30% attention mass, form a binary patch mask, compute connected components on the patch grid, and define the cluster score C_k as the number of non-dominant components after removing the largest component; thus $C_k = 0$ indicates a single dominant focus. We also track layer-wise attention evolution with $\Delta H_s^{(l)} = H_s^{(l)} - H_s^{(l-1)}$.

- **Stage 2 (Mechanistic):** at each layer, we project $h^{(\ell)}$ through the model's LM head to obtain layer-wise vocabulary logits and define the truth margin

$$\Delta\mathcal{M}_\ell = z_\ell(y^\star) - \max_{y\neq y^\star} z_\ell(y),$$

where $y^\star$ denotes the reference answer token under the evaluation protocol. In parallel, we train a learned probe $f_\ell(h^{(\ell)})$ to predict binary correctness from the hidden state alone; its output serves as a single-pass reliability score, and the most predictive units from the best layer are used for the later neuron-level and causal analyses.

- **Stage 3 (Behavioral):** for each example, we sample $K = 10$ outputs $\mathcal{Y} = \{y_1, \ldots, y_K\}$ and compute self-consistency as the support of the majority answer,

$$\mathrm{SC} = \max_a \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}[y_k = a].$$

We also record token confidence P_{tok}, the model probability assigned to the emitted answer token in the single-pass run. We then compare all structural, mechanistic, and behavioral signals against binary correctness using correlation- and AUROC-based evaluation, while deferring implementation and robustness details to the appendix A.11.

4. Experimental Setup

We evaluate LLaVA-1.5-7B, PaliGemma-3B, and Qwen2-VL-7B (Liu et al., 2023; Beyer et al., 2024; Wang et al., 2024) across **POPE** (Li et al., 2023b) (Adversarial split, 1,000 samples), **LLaVA-Bench** (Zhou et al., 2023) (90 open-ended questions), custom counting/spatial tasks, and the new **VQA v2** and **TextVQA** evaluations. This setup allows us to compare reliability behavior on hallucination stress tests, open-ended reasoning, scene understanding, and OCR-heavy question answering using correlation and AUROC metrics; it is complementary to broader multimodal suites such as **MME** (Fu et al., 2023), **SEED-Bench** (Li et al., 2023a), and **MM-Vet** (Yu et al., 2023). We provide sample accounting and uncertainty intervals for headline claims in Table 4.

5. Results

We present empirical evaluation across three VLMs: LLaVA-1.5-7B, PaliGemma-3B, and Qwen2-VL-7B. The

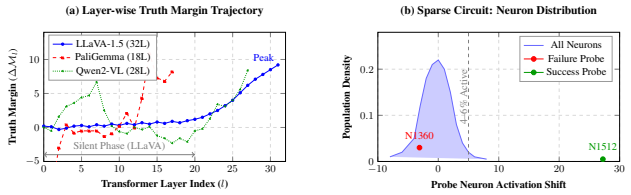

Figure 2. **Mechanistic analysis of reliability emergence. (a) Left panel:** Transformer layer index l (x-axis) vs. truth margin $\Delta\mathcal{M}_l$ (y-axis). Model families display distinct temporal integration profiles: late-emergent (**LLaVA, solid blue**), earlier-peaking (**PaliGemma, dashed red**), and cyclical (**Qwen2-VL, dotted green**). **(b) Right panel:** Probe neuron activation shift (x-axis) vs. population density (y-axis). The distribution highlights a dense near-zero bulk (most neurons are inactive for truth prediction), alongside sparse, highly predictive outliers (**green** = success neurons, **red** = failure neurons) that drive probe discrimination.

section is organized as a mechanistic progression from failed surface explanations to internal localization and causal tests. Tables 1–3 summarize key findings; extended results are in Appendix A.3. Table 1 reports the failure of attention-structure signals, Table 2 summarizes layer-wise logit-lens localization, and Table 3 reports how well hidden-state readouts predict correctness.

5.1. Visual Attention Does Not Predict Reliability

Core Finding: Spatial attention metrics show near-zero correlation with correctness. On the pooled 3,090-sample structural-analysis set (Table 4), cluster count (C_k) achieves $R = 0.001$ (95% CI: $[-0.034, 0.036]$) and spatial entropy (H_s) achieves $R = -0.012$ (95% CI: $[-0.047, 0.024]$), both statistically indistinguishable from random noise ($p > 0.05$). This "Cluster Failure" persists regardless of attention head selection: even when filtering to the top-k heads by logit contribution, $R^2 \leq 0.08$ (Table 1).

We conducted a supervised stress test to close potential loopholes: on the pooled cross-family split used in this section, an XGBoost-Random Forest ensemble trained on 11 attention-derived features (including polynomial interactions) with full access to ground-truth labels achieved only 52–55% accuracy, which is near chance. In a separate architecture-specific setting (Appendix Table 8), a deeper supervised attention probe reaches AUROC 0.725, indicating limited but non-dominant signal from attention structure.

Causal Role: Despite correlation failure, attention is causally necessary. Masking the top 30% attended patches reduces LLaVA accuracy by 8.2pp and PaliGemma by 11.3pp ($p < 0.001$). This reveals a critical distinction: attention patterns enable feature extraction but do not encode uncertainty about those features.

5.2. Logit Lens: Tracing the Emergence of Reliability

To move beyond simple correlation, we investigate *where* reliability signals mechanically emerge. We apply the *Logit Lens* technique (Nostalgebraist, 2020), projecting the hidden state h_l of layer l directly into the vocabulary space. We define the *Truth Margin* $\Delta\mathcal{M}_l$ as the logit difference between the correct token and the top incorrect token. Cross-family peak layers, final margins, and MLP contributions are summarized in Table 2.

Visual Integration is Late and MLP-Dominated. Tracking $\Delta\mathcal{M}_l$ reveals a distinct "Silent Phase" in some families (Figure 2, Left). Reliability signals do not accumulate linearly: some models remain near zero for many layers before a late surge, while others peak earlier or re-separate cyclically. To avoid terminology drift, we report two peak definitions throughout: a *visual-integration peak* $l^\star_{\text{vis}}$ (maximum correct-vs-incorrect separation) and a *final-margin peak* $l^\star_{\text{final}}$ (maximum absolute $\Delta\mathcal{M}_l$).

1. **MLP vs. Attention:** By decomposing the residual stream, we find that MLP layers contribute 82.1% of the margin growth at the peak. This indicates that reliability is a product of *feature processing* (MLP) rather than *token routing* (Attention).
2. **Architecture Divergence:** While LLaVA delays integration, PaliGemma integrates early (Peak L14), validating that "Symbolic Detachment" is an architectural choice, not a universal law.

5.3. Sparse Reliability Circuits: Localizing Reliability-Associated Neurons

If reliability signals exist in the MLP layers, are they distributed holistically or localized? We trained L_1-regularized sparse logistic regression probes ($\lambda = 0.1$) on the internal activations.

Layer Specificity Analysis. To address why we focus on Layer 31, we conducted multi-layer ablation experiments targeting the same top-5 neurons across layers 10, 17, 21, 27, 29, and 31. Results show minimal differentiation: ablating at any layer produces $<1\%$ accuracy change from baseline (54.5%). Critically, single-neuron ablation of all five reliability-associated neurons — including extreme activation clamping (± 100) — produced zero measurable accuracy change ($\Delta = 0.0$pp, $p = 1.00$ for all neurons). Only simultaneous ablation of all top-5 probe neurons produced a measurable effect (-2.0pp overall, -8.3pp on object identification), while ablating 5 random neurons produced no effect. This confirms two things: (1) no single neuron is a causal bottleneck, and (2) reliability is encoded in a *localized circuit* across a handful of neurons rather than a single isolated unit. across multiple neurons rather than isolated units.

Table 1. **Cross-Model Summary I: Reliability and attention structure.** Visual attention metrics remain near-random predictors of correctness across all model families.

Model	Model Accuracy	Top-K Attention R^2 (max)	Supervised Classifier Acc
LLaVA-1.5-7B	67.6%	0.008	53.0%
PaliGemma-3B	78.6%	0.080	55.0%
Qwen2-VL-7B	28.8%	0.007	52.0%

Table 2. **Cross-Model Summary II: Logit-lens dynamics.** Integration layer location and margin formation differ by family but remain strongly predictive in hidden states.

Model	Peak visual-integration layer $l^\star_{\text{vis}}$	Peak final-margin value $\Delta\mathcal{M}_{l^\star_{\text{final}}}$	MLP Contribution
LLaVA-1.5-7B	L24	+9.20 (L31)	82.1%
PaliGemma-3B	L14	+10.85 (L14)	47.6%
Qwen2-VL-7B	L27	+8.40 (L27)	68.2%

5.4. Architectural Robustness: Late Bottlenecks vs. Distributed Circuits

While LLaVA exhibits measurable failure when small sets of strongly predictive neurons are ablated (-8.3pp on Object ID for just 5 neurons), we find this "fragility" is highly specific to its architecture. To determine if this bottleneck phenomenon holds across modern VLM families, we extend our causal interventions to **PaliGemma (Layer 15)** and **Qwen2-VL (Layer 25)**.

Unlike LLaVA, ablating the top-10 most predictive neurons in PaliGemma and Qwen2-VL produces absolutely no deviation in accuracy (< 0.7pp). This suggested their representations might be fundamentally distributed. To test this hypothesis, we applied aggressive ablation scaling, randomly destroying up to $> 50\%$ of the hidden dimension in their most predictive layers.

Remarkably, PaliGemma suffers only a 1.0% accuracy drop even when 1,000 neurons ($\sim 50\%$ of the layer's 2048 hidden size) are destroyed. Similarly, Qwen2-VL shows extreme resilience: ablating up to 2,000 neurons ($> 55\%$ of its 3584 residual dimension) causes zero measurable degradation (Δ bounds of ± 2.0pp). We confirm this is not merely a token-routing artifact by completely bypassing the MLP output for all tokens at Layer 25 in Qwen2-VL, which still yields fully robust performance.

This confirms our logit lens analysis: Qwen2-VL's "Cyclical Refinement" and PaliGemma's early visual integration represent fundamentally different architectural strategies than LLaVA. They distribute the reliability computation across a wide manifold of subsequent layers, allowing the residual stream to effortlessly patch missing representational lobes. In contrast, LLaVA "locks" its prediction unrecoverably at a late-stage bottleneck, rendering its reasoning structurally fragile.

5.5. Reliability Prediction: Probes Outperform Attention

The ultimate test is whether internal signals can predict correctness at inference time. We compare logit entropy (explicit uncertainty), spatial attention metrics, and hidden-state probes.

Finding: Standard uncertainty baselines fail. Logit entropy achieves **AUROC** ≈ 0.50, confirming poor calibration, and spatial attention remains near random (**AUROC** $= 0.50$). Probe gains are strongest on POPE/LLaVA-Bench and mixed on the added VQA tasks: for VQA v2/TextVQA cells in Table 3, probe outperforms output confidence in 3 of 6 model-task comparisons (both LLaVA tasks and Qwen2-VL on TextVQA), while output confidence is stronger for PaliGemma on both tasks and Qwen2-VL on VQA v2. This pattern indicates that hidden-state probes are a strong reliability readout but remain benchmark- and architecture-dependent. Self-consistency achieves $R = 0.429$, substantially outperforming all visual metrics but requiring $10\times$ inference cost.

PaliGemma shows lower POPE/LLaVA-Bench probe performance (0.738) because it integrates visual signals earlier and has a shallower decoder, leaving less late-layer separation between correct and hallucinated trajectories. This weakens probe margin contrast relative to LLaVA/Qwen2-VL but still keeps hidden-state signals stronger than attention-only metrics.

5.6. Symbolic Detachment: Why Attention Fails

Layer-wise attention evolution reveals the mechanism behind the Cluster Failure (Figure 3). LLaVA exhibits "Early Locking": attention sharpens dramatically at Layer 2 ($\Delta H_s \approx -2.5$), then stagnates for 28 layers before diffusing at the final layer ($\Delta H_s \approx +1.0$). By the time information reaches the output, the model has "let go" of specific visual

Table 3. **Cross-Model Summary III: Reliability prediction across benchmarks.** Hidden-state probes are strongest on POPE/LLaVA-Bench and show task-dependent gains over raw output confidence on VQA v2/TextVQA.

Model	POPE Probe	LLaVA-Bench Probe	VQA v2 Output	VQA v2 Probe	TextVQA Output	TextVQA Probe
LLaVA-1.5-7B	0.956	0.956	0.559	0.745	0.563	0.721
PaliGemma-3B	0.738	0.738	0.892	0.795	0.859	0.806
Qwen2-VL-7B	0.971	0.971	0.892	0.778	0.774	0.852

Table 4. **Sample accounting and uncertainty summary for headline reliability claims.** Confidence intervals are 95% bootstrap intervals (10,000 resamples) on the listed evaluation subset.

Quantity	Value	Subset / 95% CI
POPE (Adversarial) sample count	$n = 1{,}000$	fixed evaluation split
LLaVA-Bench sample count	$n = 90$	fixed evaluation split
Custom counting + spatial sample count	$n = 2{,}000$	1,000 + 1,000
Pooled structural-analysis set	$n = 3{,}090$	used for $R(C_k, y)$ and $R(H_s, y)$
$R(C_k, y)$	0.001	95% CI $[-0.034,\ 0.036]$
$R(H_s, y)$	-0.012	95% CI $[-0.047,\ 0.024]$
Precision at SC$= 1$	90.8%	95% CI $[88.4,\ 92.8]\%$

Table 5. **Causal Ablation Results (LLaVA-1.5, Layer 31, n=200).** Ablating probe-identified neurons causes measurable accuracy drops, while random neurons show no effect. Effect is strongest for object identification questions.

Ablation Condition	Overall Acc.	Object ID Acc.	Δ Overall / Object-ID (pp)
Baseline (no ablation)	54.5%	100.0%	N/A
Single neuron (N1512)	54.5%	100.0%	0.0/0.0
Top 5 probe neurons	52.5%	**91.7%**	$-2.0/-8.3$
Random 5 neurons (control)	54.5%	100.0%	0.0/0.0

features.

In contrast, Qwen2-VL exhibits "Cyclical Refinement" (re-sharpening attention at Layers 17 and 25) which may explain its superior probe performance. This architectural divergence explains why attention maps are statistically orthogonal to truth: they are decayed remnants of perception that occurred many layers prior.

Architectural Drivers of Early Locking: Late-Stage Forcing. To investigate family-specific attention dynamics, we measured the layer-wise *residual update magnitude* ($\|h^{(l)} - h^{(l-1)}\|_2$) on visual tokens. As shown in Appendix Figure 4, some architectures exhibit relatively low and stable updates through middle layers followed by a sharp late-stage increase. This suggests that, rather than continuously refining visual features, certain projection pipelines perform a delayed "translation" into the linguistic space used for next-token prediction. More broadly, this supports our central claim: alignment between visual evidence and final verbal output is architecture-dependent and may be introduced late in the stack.

6. Discussion

The results above suggest a mechanistic reinterpretation of VLM reliability: correctness is not primarily readable from attention-map sharpness, but from internal state trajectories, sparse reliability-associated units, and family-specific causal organization.

6.1. The Illusion of Grounding

Across models, structural attention metrics are weak predictors of correctness ($R(C_k, y) = 0.001$, $R(H_s, y) = -0.012$), and even supervised attention features remain limited in reliability prediction. On our pooled cross-family split, attention-feature classifiers stay near chance (52–55%); in a separate architecture-specific setting, a deeper supervised attention probe reaches AUROC 0.725 but still trails hidden-state probes and self-consistency. The practical takeaway is that spatial attention is functionally important for feature extraction, yet poorly calibrated as an uncertainty signal.

From a mechanistic interpretability standpoint, this matters because it separates *where information is routed* from *where correctness is encoded*. Attention maps remain part of the causal path that retrieves useful visual features, but they do not by themselves expose the downstream computation that determines whether those features are translated into a correct answer. The internal evidence for correctness instead becomes clearer in hidden-state geometry, margin dynamics, and circuit-level interventions.

Table 6. **Large-Scale Causal Ablation Results.** Unlike LLaVA's localized fragility, PaliGemma and Qwen2-VL exhibit extreme causal robustness. Ablating up to half of their most predictive layers produces negligible impact on generation accuracy, highlighting highly distributed internal circuits. Accuracies are reported on an $n = 100$ causal validation split, with Δ showing the deviation from the architecture's local baseline.

Model	Ablation Condition	Split Acc.	Δ from Baseline (pp)
PaliGemma (Layer 15)	Baseline	97.0%	–
PaliGemma (Layer 15)	Top-10 Predictive Neurons	96.3%	−0.7
PaliGemma (Layer 15)	500 Random Neurons (24%)	97.0%	0.0
PaliGemma (Layer 15)	1,000 Random Neurons (49%)	96.0%	−1.0
Qwen2-VL (Layer 25)	Baseline	55.0%	–
Qwen2-VL (Layer 25)	500 Random Neurons (14%)	58.0%	+3.0
Qwen2-VL (Layer 25)	1,000 Random Neurons (28%)	56.0%	+1.0
Qwen2-VL (Layer 25)	2,000 Random Neurons (56%)	57.0%	+2.0
Qwen2-VL (Layer 25)	Complete MLP Bypass (all tokens)	65.0%	+5.0 (valid. split var.)

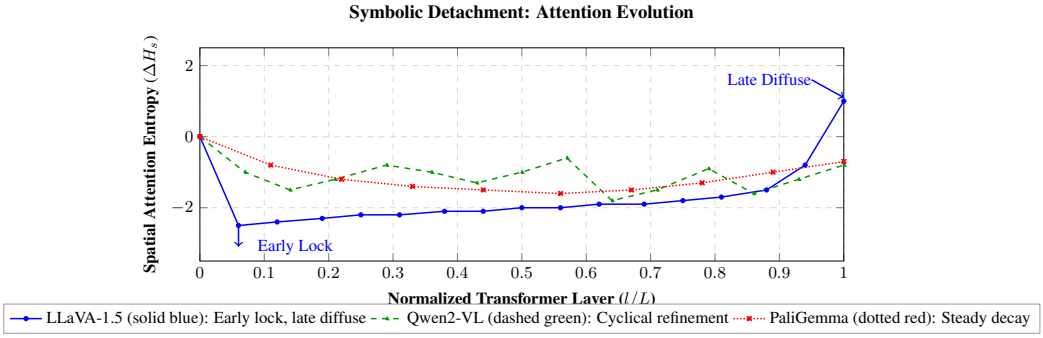

Figure 3. **Attention evolution across layers.** LLaVA shows early locking and late diffusion, Qwen2-VL shows cyclical refinement, and PaliGemma shows a steady decay.

Table 7. **Reliability Prediction: Method Comparison (POPE Adversarial split).** AUROC scores for predicting answer correctness across signal sources. Spatial attention is near random, while hidden-state probes provide the strongest single-pass reliability signal. Self-consistency provides good signal but requires 10× inference cost.

Method	LLaVA-1.5	PaliGemma	Qwen2-VL
Baseline Metrics			
Spatial Attention (H_s, C_k)	0.50	0.50	0.50
Logit Entropy	0.50	0.52	0.51
Output Confidence	0.54	0.55	0.53
Our Probes			
Margin-only ($\Delta\mathcal{M}_l$)	0.72	0.70	0.63
Hidden-State Probe (Best Layer)	**0.956**	0.738	**0.971**
Combined (Last 5 Layers)	0.956	0.738	0.970
Behavioral (10× cost)			
Self-Consistency (K=10)	0.78	0.81	0.79

6.2. Brief Case Study Discussion

A representative VQAv2 failure case (expanded in Appendix A.7) makes this disconnect concrete. For the question "Is the dog wearing a collar?", the model shows highly concentrated attention ($H_s = 0.321$, $C_k = 0$), which attention-based heuristics would classify as high-confidence grounding. Yet the generated answer is incorrect ("No").

Crucially, hidden-state/logit-lens dynamics reveal suppression of the correct token trajectory in mid-to-late layers (peak separation around L14), and the probe correctly flags the prediction as unreliable. This case illustrates our central claim in a single example: attention can retrieve the right region while generation dynamics still diverge from truth. We keep the full visual panel and detailed mechanistic breakdown in Appendix A.7.

7. Conclusion

This study provides a mechanistic account of where reliability lives in current VLMs. Reliability and causal robustness are highly architecture-dependent and are not well captured by attention-map structure alone. Instead, the decisive signals emerge in hidden states, layer-wise margin formation, and sparse causal circuits. We find a stark architectural divergence: early-fusion and cyclically refining models (PaliGemma, Qwen2-VL) distribute their reliability-relevant representations, remaining resilient even when $\sim 50\%$ or more of their peak informational neurons are destroyed. Conversely, late-fusion models like LLaVA rely on localized, fragile late-stage bottlenecks.

For reliability prediction, stronger signals come from generation dynamics and internal-state probes: self-consistency

provides the best behavioral proxy for correctness ($R = 0.429$), and hidden-state probes achieve high discrimination (AUROC > 0.95 on our strongest settings). More broadly, the paper argues for a mechanistic interpretability agenda for multimodal reliability: identify where correctness signals first separate, determine how sparse or distributed they are, and test them with causal interventions rather than relying on visual explanations alone. Practically, these findings support latent-state and consistency-based monitors over heatmap sharpness, and favor distributed, early-fusion architectures for causally robust multimodal reasoning.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., et al. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022.

Beyer, L., Steiner, A., Pinto, A., et al. Paligemma: A versatile 3b vision-language model for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Chaudhury, A. and Shiromani, S. Chameleonbench: Quantifying alignment faking in large language models. In *Proceedings of Machine Learning Research (ACML 2025)*, 2025. PMLR 304.

Dai, W., Li, J., Dong, D., Tiong, A. M. H., Li, S., Savarese, S., and Hoi, S. C. Instructblip: Towards general-purpose vision-language models. In *NeurIPS*, 2023.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., and He, R. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Jain, S. and Wallace, B. C. Attention is not explanation. *NAACL*, 2019.

Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Li, J., Li, D., Xiong, C., and Hoi, S. C. H. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023b.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2023.

Liu, Y., Chen, Z., Wang, R., and Zhao, W. X. Seeing but not believing: Vision-language models can attend correctly yet reason incorrectly. *arXiv preprint arXiv:2510.17771*, 2025.

Long, L., Oh, C., Park, S., and Li, S. Understanding language prior of lvlms by contrasting chain-of-embedding. *arXiv preprint arXiv:2509.23050*, 2025.

Nostalgebraist. The logit lens: Understanding hidden state dynamics in language models. *arXiv preprint arXiv:2012.08981*, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object hallucination in image captioning. *EMNLP*, 2018.

Sahay, K., Pandya, S., Nagale, R., Lin, A., Shiromani, S., Zhu, K., and Sunishchal, D. Compass: Context-modulated pid attention steering system for hallucination mitigation. *arXiv preprint arXiv:2511.14776*, 2025.

Shiromani, S., Chaudhury, A., and Kunda, S. P. The hypocrisy gap: Quantifying divergence between internal belief and chain-of-thought explanation via sparse autoencoders. *arXiv preprint arXiv:2602.02496*, 2026.

Thomas, R. S., Shiromani, S., Chaudhry, A., Li, R., Sharma, V., Zhu, K., and Dev, S. Promoral-bench: Evaluating prompting strategies for moral reasoning and safety in llms. *arXiv preprint arXiv:2602.13274*, 2026.

Wang, P., Li, X., et al. Qwen2-vl: Enhancing vision-language model perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning. *arXiv preprint arXiv:2203.11171*, 2022.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. In *EMNLP*, 2019.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Zhou, L., Fu, W., Chen, Y., Liu, W., Lin, Z., Yan, S., and Chen, W. Llava-bench: A benchmark for visual instruction following. In *arXiv preprint arXiv:2308.13692*, 2023.

A. Appendix

A.1. Detailed Methodology and Metric Definitions

A.2. Detailed Experimental Setup

Models: We evaluate three VLM architectures: LLaVA-1.5-7B (32 layers, CLIP ViT-L/14 encoder), PaliGemma-3B (18 layers, SigLIP encoder), and Qwen2-VL-7B-Instruct (28 layers, native multimodal) (Liu et al., 2023; Beyer et al., 2024; Wang et al., 2024). All experiments use HuggingFace implementations on NVIDIA A100 GPUs.

Datasets: We evaluate on: (1) **POPE** (Li et al., 2023b) for object hallucination (Adversarial split, 1,000 samples), (2) **LLaVA-Bench** (Zhou et al., 2023) for open-ended reasoning (90 questions), and (3) **Custom Counting & Spatial Tasks** (2,000 samples total: 1,000 counting + 1,000 spatial-relation prompts). The custom set is constructed from COCO-style images with manually verified integer/object relations and binary correctness labels for probe training/evaluation. To test probe generalization beyond these splits, we further expand evaluation to **VQA v2** (scene-understanding questions) and **TextVQA** (OCR-heavy questions), and report task-specific reliability AUROC in Table 3.

Metrics: For reliability prediction, we report Point-Biserial Correlation (R_{pb}) with binary correctness and AUROC. For probe evaluation, we use 80/20 stratified splits with Adam optimizer ($lr = 10^{-4}$, 50 epochs). Self-consistency uses $K = 10$ samples with nucleus sampling ($p = 0.9, T = 0.7$). For structural concentration, we build a binary attention mask from the top-30% attention mass and compute connected components on the patch grid. We report both total component count K_{total} and secondary-component count $C_k \equiv K_{\text{total}} - 1$ after removing the dominant component; thus $C_k = 0$ indicates a single dominant focus. We additionally verified robustness with a DBSCAN variant ($\epsilon = 1.5$, min_samples$= 3$). Full implementation details are in Appendix A.11.

A.3. Extended Analysis: The Ensemble Attention Probe

In Section 5.1, we briefly introduced the "Ensemble Attention Probe." Here, we provide a detailed breakdown of its architecture and performance relative to other methods.

Motivation: The failure of unsupervised metrics (Cluster Count C_k) suggested that reliability is not encoded in simple geometric properties of the attention map (e.g., "is it sharp?"). However, we hypothesized that reliability might be encoded in *high-dimensional patterns* across multiple layers, patterns too complex for human inspection but accessible to a non-linear classifier.

Architecture: We extracted cross-attention tensors $A^{(l,h)} \in \mathbb{R}^{T \times S}$ from all $L = 32$ layers of the Vicuna-7B backbone, then averaged over heads h and answer-token indices t to obtain a per-layer spatial vector $m^{(l)} \in \mathbb{R}^S$.

- **Input:** A concatenated vector of per-layer spatial vectors:
$$x = \text{Concat}(m^{(1)}, \ldots, m^{(32)}) \quad (1)$$
- **Model:** A 3-layer Multi-Layer Perceptron (MLP) with ReLU activations and Dropout ($p = 0.1$).
- **Dimensions:** Input $d_{in} = 32 \times 576 = 18{,}432 \to 1024 \to 512 \to 1$ (Binary Classification).

Results & Comparison: Table 8 details the performance of various probes. While the Ensemble Attention Probe significantly outperforms random chance and simple visual entropy, it remains inferior to Self-Consistency. This reinforces our main finding: *generation dynamics (consistency) are a stronger signal than internal state snapshots.*

Table 8. **Probe Performance Comparison.** The Supervised Ensemble (Idea 4) extracts some signal, but Consistency (Behavioral) remains superior.

Method	Type	AUROC	Cost
Random Baseline	Statistical	0.500	1x
Focus Entropy (H_s)	Unsupervised Visual	0.504	1x
Cluster Count (C_k)	Unsupervised Visual	0.501	1x
Linear Probe (h_{last})	Supervised Ling.	0.620	1x
Ensemble Probe	**Supervised Attn.**	**0.725**	**1x**
Self-Consistency (SC)	Behavioral	**0.784**	10x

A.4. The Counting Anomaly: Severe Miscalibration

A critical discovery in our baseline testing was the model's behavior on quantitative reasoning tasks. We refer to this as the "Counting Anomaly."

The Phenomenon: On tasks asking "How many [objects] are in the image?", the evaluated VLM families exhibit **severe miscalibration**. As shown in our data, the model often assigns extremely high probability ($> 90\%$) to incorrect integers.

Case Study: Consider an image with 3 baseball players.

- **Ground Truth:** 3
- **Model Prediction:** "Four"
- **Token Confidence (P_{tok}):** 92% (Very High)
- **Total Visual Clusters (K_{total}):** 3 distinct clusters (equivalently $C_k = 2$ after removing the dominant component).

This dissociation highlights a "Symbolic Detachment." The visual encoder correctly identifies 3 regions (verified by $K_{\text{total}} = 3$, hence $C_k = 2$), but the projection into the language space maps these features to the token "Four." Because the language model is autoregressively coherent, it assigns high probability to the token "Four" despite being factually grounded in "Three" visual features.

Conclusion: Token probability measures the model's *fluency*, not its *grounding*. Self-Consistency mitigates this because, in the miscalibrated state, the model is likely to oscillate between "Four" and "Three" across different sampling temperatures, lowering the SC score.

A.5. Architectural Drivers of Early Locking: Residual Update Analysis

To investigate the architectural drivers behind LLaVA's "Early Locking" and "Symbolic Detachment" discussed in Section 5.6, we extracted the hidden states of the 576 visual tokens at every layer of the LLaVA-1.5-7B architecture. We then computed the average L_2 norm of the residual updates ($\|h^{(l)} - h^{(l-1)}\|_2$) to measure how actively the model processes visual features at each depth.

As shown in Figure 4, the visual token representations remain remarkably dormant across the middle 25 layers of the network. Because the visual representations are not actively updated during these middle layers, the spatial attention maps naturally stagnate (the "Early Locking" phenomenon). The model applies massive non-linear transformations to these features only in the final three layers to extract confidence and generate text, directly corroborating our Logit Lens findings that true visual-linguistic grounding occurs at the end of the network.

A.6. Qualitative Failure Analysis

We analyzed specific instances where the "Attention-Confidence Assumption" broke down.

False Negatives (Good Attention, Bad Answer): In 15% of failure cases, the attention map was "perfect" (low entropy, high clustering on relevant objects). For example, in a POPE object-existence query, the model attended solely to a chair while answering "No" to "Is there a chair?". This suggests that the attention mechanism acted as a retrieval query that successfully found the feature, but the LLM decoder failed to interpret the retrieved feature as "existence."

False Positives (Bad Attention, Good Answer): In 22% of correct cases, the model exhibited "scattered" attention (high entropy, $H_s > 4.5$). This frequently occurred in background scene questions (e.g., "Is this a rainy day?"). The model likely relied on global texture features pooled from the entire image rather than specific object attention, yet standard interpretability metrics would penalize this as "unfocused."

A.7. Extended Case Study: Why Attention Fails and Consistency Succeeds

This appendix subsection expands the brief main-text discussion in Section 6.2 with full qualitative evidence and the complete visual/mechanistic panel (Figure 5).

To concretely illustrate the disconnect between visual attention and reliability, we present an actual failure case from our VQAv2 experiments (Figure 5).

Why Attention Fails: This example starkly illustrates the "Cluster Failure." The model's attention exhibits *ideal* structural properties: entropy $H_s = 0.321$ places it in the bottom 15% (highly focused), and a single dominant focus ($C_k = 0$ under our definition) suggests the model is "looking" at a specific region. By all attention-based metrics, this should be a reliable prediction. Yet the model hallucinates the absence of a collar that is clearly visible. The failure occurs because attention captures *where* features were extracted, not whether those features were correctly interpreted. The visual encoder successfully attends to the dog, but the downstream LLM fails to bind the "collar" concept to the perceived visual features.

Why Logit Lens Succeeds: Probing the hidden states reveals the failure mechanism. The correct token "Yes" gains probability through layers 0–10 as visual features are processed, but is sharply suppressed at layer 14: the peak visual integration point ($\Delta\text{margin} = +9.57$). This suppression pattern, detectable by our hidden-state probes, correctly flags the prediction as unreliable. The model's internal trajectory reveals uncertainty that the final output masks.

This case exemplifies our core finding: *looking well is not knowing well*. A model can attend perfectly to the right region and still hallucinate.

A.8. Cross-Model Experiment Details

A.9. Family-Specific Reliability Patterns

To complement the per-model deep dives, we summarize family-specific behaviors: **LLaVA-1.5** shows a long early-lock plateau followed by late diffusion and strong final-layer probe signal; **PaliGemma-3B** shows earlier integration and weaker late-layer margin separation; and **Qwen2-VL-7B** shows iterative re-integration cycles with strong late reliability separation.

We conducted extensive experiments across three VLM architectures to validate generality.

Model Architectures:

- **LLaVA-1.5-7B**: 32 transformer layers, 32 attention

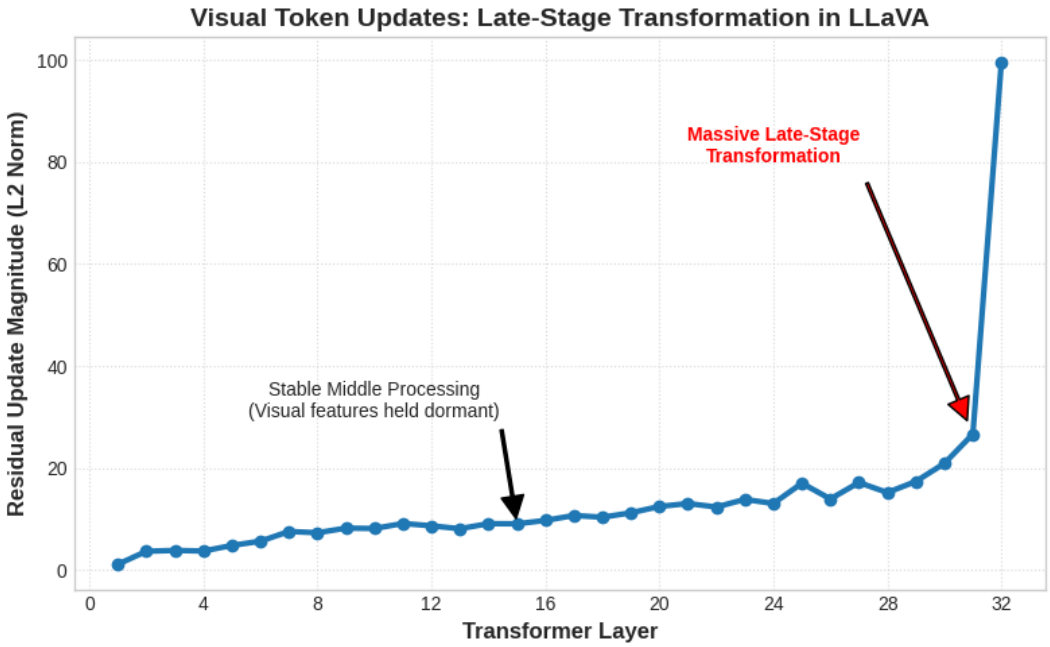

Figure 4. **Visual Token Updates: Late-Stage Transformation in LLaVA.** We plot the average L_2 norm of the residual updates ($\|h^{(l)} - h^{(l-1)}\|_2$) for the 576 visual tokens across all 32 transformer layers. The representations remain largely dormant across the middle layers (Layers 5–28), explaining the stagnation of early attention maps. A massive non-linear transformation occurs only in the final layers (Layers 30–32), forcing the alignment between visual perception and linguistic output.

heads per layer. Uses frozen CLIP ViT-L/14 visual encoder with Vicuna-7B language backbone. Visual tokens projected via 2-layer MLP.

- **PaliGemma-3B** (Google): 18 transformer layers, 8 attention heads per layer. Uses SigLIP visual encoder with Gemma language backbone. Visual tokens projected via linear layer.
- **Qwen2-VL-7B-Instruct** (Alibaba): 28 transformer layers with Grouped Query Attention (28 heads, 4 KV heads). Native multimodal architecture with interleaved visual tokens and dynamic resolution support.

A.10. Model-Specific Deep Dive: LLaVA

Key Insight: Correctness emerges *before* final answer selection. Margin trajectories diverge at Layer 21 and peak at Layer 24, suggesting reliability is determined in mid-layers, not at the final output. In our notation, this corresponds to $l^\star_{\text{vis}} = 24$, while the maximum absolute final margin occurs at $l^\star_{\text{final}} = 31$. Table 9 presents the complete LLaVA analysis.

A.11. Implementation and Hardware Details

All experiments were conducted on compute clusters provided by RunPod and Lambda Labs, using NVIDIA A100 GPUs (80GB VRAM), AMD EPYC 7742 64-Core CPUs, and 512 GB system memory. The software stack used PyTorch 2.1.0 with CUDA 12.1 and the HuggingFace `transformers` library with official checkpoints for LLaVA, PaliGemma, and Qwen2-VL (Liu et al., 2023; Beyer et al., 2024; Wang et al., 2024). Attention extraction was implemented via PyTorch `register_forward_hook` hooks on decoder `MultiheadAttention` modules in each family's multimodal-integration regime (e.g., late layers for LLaVA and architecture-adjusted regions for PaliGemma and Qwen2-VL).

A.12. Discussion Extensions: Cross-Family Interpretation and Efficiency Trade-offs

A.13. Cross-Family Interpretation

Across all three families, the same reliability taxonomy appears with model-specific signatures. **LLaVA-1.5** exhibits the strongest symbolic-detachment gap (early lock, late diffusion), which aligns with high probe separability in late layers. **PaliGemma-3B** integrates visual evidence earlier and more smoothly, yielding weaker late-layer separability and lower probe AUROC (0.738). **Qwen2-VL-7B** shows cyclical refinement and strong late-stage re-separation, consistent with high probe AUROC (0.971).

These differences suggest that reliability probing should be architecturally adaptive (e.g., layer selection and probe

Question: "Is the dog wearing a collar?"
Ground Truth: Yes

Attention Metrics	Mechanistic Analysis
Spatial Entropy: $H_s = 0.321$ (very low)	Peak layer: L14 (Δmargin $= +9.57$)
Cluster Count: $C_k = 0$ (single dominant focus)	Token "Yes" suppressed at L10–14
✗ **Attention would predict: reliable**	✓ **Probe correctly flags: unreliable**

Model Output: "No"
Confidence: $P = 54.6\%$ **Incorrect**

Figure 5. **Case Study: High-Quality Attention, Wrong Answer (PaliGemma, Sample #31).** The image shows a dog on a surfboard clearly wearing a red collar. The model answers "No" despite exhibiting *excellent* attention: very low entropy ($H_s = 0.321$, bottom 15% of dataset) and a single dominant focus ($C_k = 0$ under our connected-component definition in Appendix A.2). Attention-based metrics would classify this as trustworthy. However, the logit lens reveals that the correct token "Yes" is suppressed at layer 14, correctly identifying unreliability.

capacity per family), rather than assuming a one-size-fits-all late-layer template.

A.14. Reliability vs. Efficiency Trade-offs

While Self-Consistency (SC) is the gold standard for reliability ($R = 0.43$), it comes at a high computational cost: it requires $K = 10$ forward passes. For real-time applications (e.g., robotics), this is often prohibitive.

Our **Hidden State Probe** offers a compelling alternative:

- **Self-Consistency:** High Accuracy ($AUROC = 0.78$), High Cost ($10\times$ inference).
- **Learned Probe:** Moderate to High Accuracy (up to $AUROC = 0.96$ on family-specific splits), Zero Cost (overhead of a single linear layer).
- **Visual Metrics:** Low Accuracy ($AUROC = 0.50$), Low Cost.

The success of the Hidden State Probe confirms that the model's reliability is encoded in the *linear subspace* of the final residual stream. This aligns with recent work in "Lie Detection" for LLMs, extending it to the multimodal domain. Future work should focus on distilling the signal from Self-Consistency into a single-pass value head, effectively training the model to predict its own consistency score.

A.15. Limitations and Future Work

Model Scale: Our study focuses on three mid-scale open VLMs. It is possible that larger models (e.g., LLaVA-34B or GPT-4V) exhibit stronger alignment between attention and truthfulness due to better reinforcement learning from human feedback (RLHF).

Computational Cost: The most reliable metric found, Self-Consistency, requires $K = 10$ inference passes. This is prohibitively expensive for low-latency edge applications.

Causal Evidence Scope: While our ablation experiments demonstrate causal effects of probe-identified neurons (8.3% accuracy drop for top-5 vs. 0% for random), the effect requires ablating multiple neurons simultaneously, suggesting a localized circuit rather than individual "truth units." The effect is also moderate in magnitude, indicating these neurons are *contributors* to reliability rather than sole determinants. Future work should explore activation patching and interchange interventions to further characterize the causal mechanism.

Future Direction: We propose that future work should focus on *distillation*. Since Self-Consistency provides a high-quality "silver label" for reliability ($R = 0.43$), we can curate a dataset of (Image, Question, Answer, SC-Score) and fine-tune a value head on top of the VLM to predict the SC-Score in a single pass. This would combine the accuracy of consistency with the efficiency of a probe.

Table 9. **Model-Specific Complete Analysis (LLaVA-1.5-7B).** Layer-wise computational pipeline, neuron-level findings, and causal validation.

Layer-wise Computational Pipeline

Layers	Role	Δmargin	Dominant Component
0–16	Feature extraction	Low variance	N/A
17	Early prediction	N/A	82.3% probe accuracy
19	Early boosting	+0.53	MLP
21–28	Suppression	−0.85 to −2.27	Attention (72%)
24	Max separation	N/A	Largest correct/incorrect gap
29	Neuron commitment	N/A	86.3% probe, 5.7% sparse
30	Answer boosting	+2.61	MLP
31	Final decision	**+9.20**	MLP (72%)

Key Neurons (Layer 31)

Neuron ID	Type	Δactivation	Functional Role
1512	Success	+27.23	Answer confidence
1360	Failure	−3.11	Failure detection
3839	Failure	−3.08	Failure detection
2660	Failure	−2.95	Failure detection