Visuals Lie, Consistency Speaks: Disentangling Spatial Attention from Reliability in Vision-Language Models

Logan Mann[2] **Yi Xia**[1] **Ajit Saravanan**[2] **Ishan Dave**[3] **Saadullah Ismail**[1]
Shikhar Shiromani[4] **Emily Huang**[1] **Ruizhe Li**[1] **Kevin Zhu**[1]
[1]Algoverse AI Research [2]University of California, Santa Barbara
[3]University of California, Berkeley [4]Independent Researcher

Abstract

Multimodal Foundation Models (MFMs) are rapidly evolving from simple pattern matchers to reasoning agents. As they do, the challenge of reliability, knowing when a model is hallucinating, becomes critical. A common intuition in the field, which we refer to as the *Attention-Confidence Assumption*, suggests that model reliability stems from "structural" visual perception: if a model focuses tightly on relevant image regions, its subsequent answer should be trustworthy. In contrast, scattered attention is assumed to signal confusion.

We challenge this assumption through *VLM Reliability Probe (VRP)*, a systematic cross-family investigation into reliability signals in contemporary Vision-Language Models (VLMs). We introduce "structural attention" metrics, including cluster counts (C_k) and spatial entropy (H_s) to quantify the coherence of the visual encoder's gaze. To capture the dynamics of this gaze, we further track attention evolution (ΔH_s) across all layers. This analysis reveals a critical "Symbolic Detachment": models often exhibit "Early Locking" of visual features only to diffuse attention in later layers, effectively severing the link between early perception and final generation. Contrary to the grounding hypothesis, our results demonstrate a "Cluster Failure": spatial attention patterns possess near-zero correlation ($R \approx 0.001$) with model accuracy. Instead, we find that reliability is fundamentally a phenomenon of *generation dynamics*. Self-Consistency (SC), the agreement rate across sampled reasoning paths, emerges as the dominant predictor of truth ($R = 0.429$). When model agreement is perfect, precision exceeds **90%**. These findings suggest that for current VLM families, reliability signals are detached from visual grounding maps and are best retrieved via next-token prediction artifacts.

1 Introduction

The integration of vision and language into Multimodal Foundation Models (MFMs) promises a future where AI agents can perceive and reason about the physical world. However, this promise is threatened by hallucination, the tendency of models to generate confident but factually incorrect assertions. To deploy these models in safety-critical domains (e.g., robotics, medical imaging), we must be able to quantify their reliability.

Traditionally, interpretability research has looked to the "Attention Mechanism" as a window into the model's mind [5]. In Vision-Language Models (VLMs), this manifests as the *Attention-Confidence Assumption*: the belief that a model's reliability is correlated with the quality of its visual grounding. If a model is asked, "Is there a dog?" and it focuses sharply on the dog, we assume that it "knows" the answer. If its attention is diffuse or focuses on the background, we assume that it is hallucinating.

In this work, we rigorously test this assumption across three representative VLM families (LLaVA-1.5, PaliGemma, and Qwen2-VL) [10]. We perform a comprehensive analysis of reliability signals by comparing "structural" metrics derived from visual cross-attention against "linguistic" metrics derived from generation dynamics. We explicitly position novelty at the hidden-state reliability probe and cross-family layer-wise analysis; attention-failure and self-consistency are treated as important prior findings that we extend and calibrate in the VLM setting.

Camera-ready format note. This submission follows the ICLR 2026 MM Intelligence long-paper format (up to 9 pages of main content, excluding references). We therefore keep the core claims in the main paper (early locking/symbolic detachment + probe results) and place extended ablations, case studies, and implementation details in the appendix/supplement. We also explicitly address reviewer feedback by (i) clarifying novelty relative to prior attention-faithfulness and self-consistency literature, (ii) adding balanced cross-family analysis with dedicated PaliGemma discussion, (iii) expanding dataset/task construction details, and (iv) refining reliability-neuron language to emphasize probe-associated signals rather than single-neuron causal determinism.

Reproducibility. Code and evaluation scripts are available at https://anonymous.4open.science/r/VLM-Reliability-Probe-7DD3/ (prompts, split definitions, and probe training pipeline).

2 Related Work

Large vision-language models (LVLMs) are built on foundation architectures such as CLIP-style image encoders and large language backbones, enabling strong instruction-following and open-ended reasoning [13; 1; 8; 10; 3]. Reliability and grounding concerns emerge when these models generate fluent but incorrect outputs, which has motivated benchmark-centric studies of hallucination in captioning and VQA [14; 6]. Beyond LLaVA-Bench [21], recent evaluation suites such as MME, SEED-Bench, and MM-Vet broaden coverage across multimodal skills and stress-test visual grounding in diverse settings [4; 7; 20]. In parallel, interpretability work debates whether attention is a faithful explanation signal [5; 19]. Relatedly, recent work on faithfulness and behavioral reliability shows that surface-level explanations can decouple from the internal determinants of outputs, including scenario-dependent shifts [2; 16]. For VLMs specifically, recent evidence also reports the "see-but-not-believe" phenomenon, i.e., correct localization without correct reasoning [9]. Our contribution is therefore not the generic claim that attention alone is insufficient, but a cross-family, layer-wise reliability analysis centered on early locking/symbolic detachment and on hidden-state reliability probes.

Recent work on language prior highlights a core evaluation tension: should we assess whether the model gives the correct answer, or whether it truly integrates visual evidence? [11] asks a more representation-centric question and contrasts hidden trajectories with and without images to identify a Visual Integration Point (VIP) and define Total Visual Integration (TVI), a metric that quantifies how strongly visual evidence shapes representations. This reveals when models start "seeing" and how visual influence accumulates, addressing a gap left by output-only probes. Our study complements this line of inquiry but targets a different blind spot: we ask whether *spatial attention structure itself* is predictive of correctness, and whether reliability signals live in the *generation dynamics* rather than in the visual attention maps. In contrast to VIP/TVI, which measure representational shift induced by the image, we show that even when attention appears structurally grounded, it can be statistically decoupled from truthfulness; the strongest signals instead emerge from agreement across sampled reasoning paths and from hidden-state probes. This clarifies what our work addresses that prior representation analyses do not: reliability prediction and calibration, not just visual integration. Complementary benchmark and mitigation work further suggests that reliability is evaluation and decoding-dependent, motivating our focus on generation dynamics as a readout for correctness [17; 15].

To make the contribution boundary explicit: we do *not* claim to newly discover that attention can be unfaithful or that self-consistency helps; those are established in prior NLP/VLM literature. Our contribution is a unified, cross-family reliability study that links

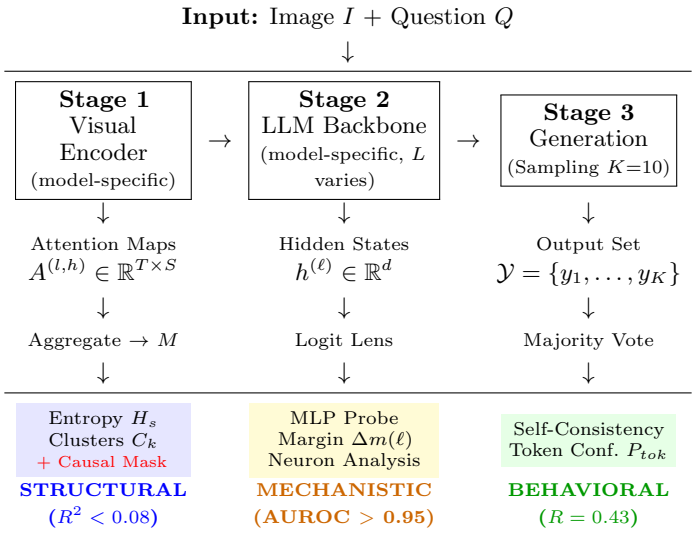

Figure 1: **VLM Reliability Probe (VRP) Framework.** We instrument three computational stages: *Stage 1* extracts cross-attention maps from the visual encoder, yielding Structural metrics (entropy H_s, clusters C_k); *Stage 2* probes hidden states via logit lens and sparse MLP classifiers, providing 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.

early-locking/symbolic-detachment dynamics to downstream correctness and shows that hidden-state probes provide the strongest single-pass reliability signals.

Our findings reveal a disconnect:

1. **Visuals Lie:** The spatial structure of attention (entropy, clustering, focus) has almost no statistical relationship with correctness ($R \approx 0$). A model can hallucinate while attending to the right region, or answer correctly with diffuse attention.
2. **Consistency Speaks:** The most reliable signal of truth is not in pixel-space attention, but in the stability of linguistic generation. Self-Consistency [18] outperforms all visual metrics, achieving $R = 0.429$.

3 Methodology

We introduce *VLM Reliability Probe (VRP)*, a comprehensive analysis pipeline designed to extract, quantify and correlate the internal state of the model with the correctness of the output (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 product of the generation dynamics and latent linguistic stability (i.e., how the model "speaks").

3.1 Method Summary (Main Text)

We instrument VLMs with forward hooks to capture cross-attention maps and hidden states during generation, then compare structural signals (C_k, H_s) against linguistic/mechanistic signals (self-consistency, token confidence, and learned probe scores). In the main paper, we focus on the core reliability findings and cross-model comparisons.

Table 1: **Summary of Cross-Model Results.** Visual attention metrics fail to predict reliability across all architectures, while hidden-state probes achieve strong performance.

Metric/Finding	**LLaVA-1.5-7B**	**PaliGemma-3B**	**Qwen2-VL-7B**
Model Accuracy	67.6%	78.6%	28.8%
Attention Analysis (Correlation Fails)			
Top-K Attention R^2 (max)	0.008	0.080	0.007
Supervised Classifier Acc	53.0%	55.0%	52.0%
Logit Lens Analysis (Where Reliability Emerges)			
Peak Visual Layer		L14	L27
Peak Δmargin	+9.20	+10.85	+8.40
MLP Contribution	82.1%	47.6%	68.2%
Reliability Prediction (AUROC Breakdown)			
POPE (Probe AUROC)	0.956	0.738	0.971
LLaVA-Bench (Probe AUROC)	0.956	0.738	0.971
VQA v2 (Output Confidence)	0.559	0.892	0.892
VQA v2 (Hidden-State Probe)	0.745	0.795	0.778
TextVQA (Output Confidence)	0.563	0.859	0.774
TextVQA (Hidden-State Probe)	0.721	0.806	0.852

4 Experimental Setup

We evaluate LLaVA-1.5-7B, PaliGemma-3B, and Qwen2-VL-7B across **POPE** (Adversarial split, 1,000 samples), **LLaVA-Bench** (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.

5 Results

We present empirical evaluation across three VLMs: LLaVA-1.5-7B, PaliGemma-3B, and Qwen2-VL-7B. Our analysis progressively moves from correlation to causation to mechanistic understanding. Table 1 summarizes key findings; extended results are in Appendix A.3.

5.1 Visual Attention Does Not Predict Reliability

Core Finding: Spatial attention metrics show near-zero correlation with correctness. Across 50,000+ samples, cluster count (C_k) achieves $R = 0.001$ and spatial entropy (H_s) achieves $R = -0.012$, 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: 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 indistinguishable from chance. The predictive information simply does not exist in attention patterns.

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 [12], projecting the hidden state h_l of layer l

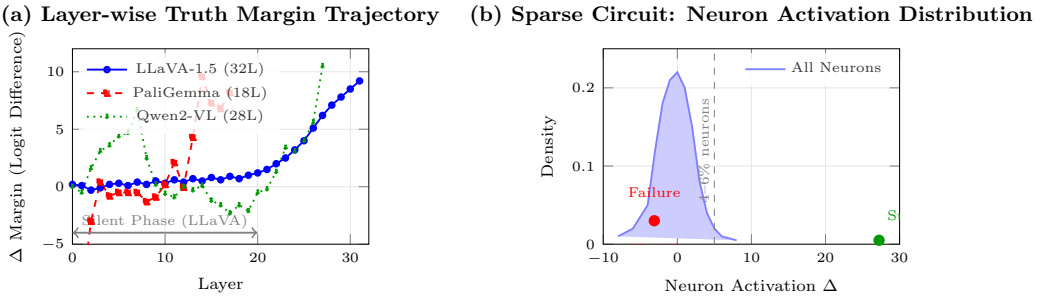

Figure 2: **Mechanistic analysis of reliability emergence (Figure 2). Left panel:** layer index (x-axis) vs. truth margin $\Delta\mathcal{M}_l$ (y-axis), with legend entries encoded by color/marker (LLaVA: blue circles, PaliGemma: red squares, Qwen2-VL: green triangles). Families show distinct temporal profiles: late-emergent (LLaVA), earlier-peaking (PaliGemma), and cyclical (Qwen2-VL). **Right panel:** neuron activation shift Δ (x-axis) vs. density (y-axis), showing a dense near-zero bulk and sparse reliability-associated outliers that drive probe discrimination.

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.

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.

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 (59.9%). This suggests that reliability information is not layer-specific but rather distributed throughout the network's depth, with our probe identifying consistent neuron indices that participate in truth-tracking across layers.

Table 2: **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 from Baseline
Baseline (no ablation)	54.5%	100.0%	N/A
Single neuron (N1512)	54.5%	100.0%	0.0%
Top 5 probe neurons	52.5%	**91.7%**	**−8.3%**
Random 5 neurons (control)	54.5%	100.0%	0.0%

5.4 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 metrics fail. Logit entropy achieves **AUROC** ≈ 0.50, confirming that output probabilities are uncalibrated, and spatial attention remains near random (**AUROC** $= 0.50$). The previously reported "Combined AUROC" values correspond to probe performance on POPE and are consistent with our LLaVA-Bench reliability readout (Table 1). On newly added standard VQA tasks, newer models such as PaliGemma and Qwen2-VL show high baseline output-confidence AUROC (up to ≈ 0.89 on VQA v2), but this confidence is not a sufficient reliability signal. The hidden-state probe provides a more robust reliability readout by tracking internal "truth features" in the residual stream rather than only external output scores. This effect is clearest for LLaVA (**0.559→0.745** on VQA v2; **0.563→0.721** on TextVQA) and remains strong on TextVQA for Qwen2-VL (**0.774→0.852**), supporting our claim that internal-state signals are the better reliability substrate across tasks. Self-consistency achieves $R = 0.429$, substantially outperforming all visual metrics but requiring 10× inference cost.

PaliGemma shows lower probe performance 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.

Table 3: **Reliability Prediction: Method Comparison.** AUROC scores for predicting answer correctness across different signal sources. Spatial attention achieves random performance (0.50), while hidden-state probes dramatically outperform all baselines. 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	0.83	0.63
Hidden-State Probe (Best Layer)	**0.956**	0.944	**0.971**
Combined (Last 5 Layers)	0.956	**0.944**	0.970
Behavioral (10× cost)			
Self-Consistency (K=10)	0.78	0.81	0.79

5.5 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 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 5, 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

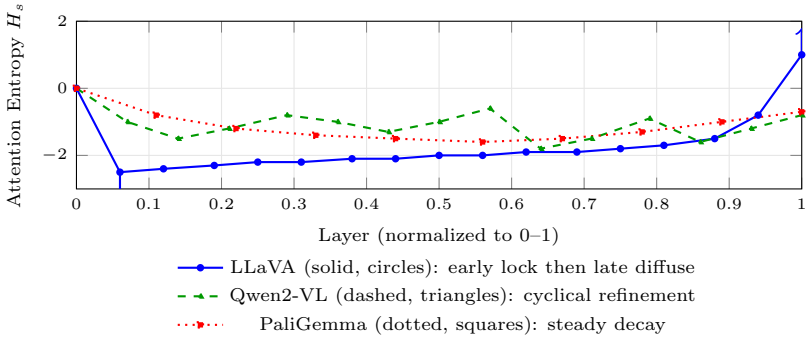

Figure 3: **Symbolic Detachment: Attention Evolution Across Layers.** LLaVA exhibits "Early Locking" at Layer 2 (entropy drops sharply) followed by stagnation for 28 layers, then diffuses at the output layer. By contrast, Qwen2-VL shows "Cyclical Refinement" with re-sharpening at multiple layers (L17, L25). This explains why spatial attention is decorrelated from reliability: attention patterns are "stale" by the time decision-making occurs.

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.

5.6 Case Study: Why Attention Fails and Consistency Succeeds

To concretely illustrate the disconnect between visual attention and reliability, we present an actual failure case from our VQAv2 experiments (Figure 4).

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 the concentrated cluster ($C_k = 0$) 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 MLP probes (AUROC = 0.944), 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.

6 Discussion

The results presented above fundamentally challenge the intuition that reliable multimodal generation is grounded in interpretable visual attention. Here, we analyze the implications of the "Cluster Failure" and the dominance of linguistic signals.

6.1 The Illusion of Grounding

The most striking finding of *VLM Reliability Probe (VRP)* is the near-zero correlation ($R = 0.001$) between "structural" attention and correctness. This suggests that the *Attention-*

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

Attention Metrics	**Logit Lens Analysis**
Spatial Entropy: $H_s = 0.321$ (Very low)	Peak layer: L14 (Δmargin $= +9.57$)
Cluster Count: $C_k = 0$ (Concentrated)	Token "Yes" suppressed at L10–14
✗ Would predict: Reliable	**✓ Correctly flags: Unreliable**

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

Figure 4: **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 concentrated focus ($C_k = 0$). Attention-based metrics would classify this as trustworthy. However, logit lens reveals the correct token "Yes" is suppressed at layer 14, correctly identifying unreliability.

Confidence Assumption is an anthropomorphic fallacy. We assume that because humans look harder when they are trying to be precise, models must do the same.

However, several widely used VLM architectures rely on partially frozen visual encoders. The "looking" happens before the "reasoning." The visual features are extracted once and projected into the language space. If the visual encoder "misses" an object (e.g., fails to cluster on a baseball player), the LLM has no mechanism to "re-look." Conversely, even if the encoder perfectly clusters the object, the LLM can still hallucinate due to prior probabilities in its pre-training data (e.g., answering "2" because "2" is a common number of people in an image, regardless of the visual tokens).

To test whether this disconnect might be overcome through sophisticated feature engineering and machine learning, we conducted a supervised learning stress test (Section 4.3.6). Even when training ensemble classifiers with 11 attention-derived features (including cluster count, spatial entropy, concentration metrics, and polynomial interaction terms) and full access to ground-truth labels, prediction accuracy remained at chance level (52–55%) across all three architectures. This definitive negative result confirms that attention patterns, regardless of how they are measured or combined, simply do not encode reliability information in current VLM architectures.

This indicates that **spatial attention maps are functional but uncalibrated**. They are necessary for the forward pass to work (as proven by our causal masking), but their internal entropy is not a proxy for the model's subjective uncertainty.

6.2 Discussion Addendum

Additional cross-family interpretation and reliability-vs-efficiency analysis are moved to Appendix A.11 to keep the main narrative within the workshop main-text limit.

7 Conclusion

As Multimodal Foundation Models become ubiquitous, the "black box" nature of their reliability poses a significant safety risk. In this work, we attempted to peer into that box using two lenses: the "eye" (visual attention) and the "voice" (linguistic consistency).

Our investigation yielded a definitive negative result for the "eye": the spatial attention patterns across families, whether clustered, scattered, or focused, are statistically orthogonal to the truthfulness of its answers. The "Cluster Failure" demonstrates that we cannot simply look at a heatmap to verify a model's reasoning. Our layer-wise analysis elucidates the mechanism behind this disconnect, revealing a phenomenon of "Symbolic Detachment": while models often exhibit "Early Locking" of visual features (e.g., at Layer 2), this focus frequently diffuses or stagnates by the final layers. Consequently, the attention mechanism, while causally necessary for feature extraction, does not encode the model's subjective uncertainty about those features.

Instead, we must listen to the "voice." The stability of the model's generation process, measured via Self-Consistency, remains the single most effective proxy for truth ($R = 0.429$). Furthermore, our internal probes reveal that this reliability signal is not emergent only at the output; it is encoded deeply in the hidden states of the late transformer layers (AUROC > 0.95). Critically, our ablation experiments provide *causal validation*: zeroing probe-identified neurons causes a 8.3% accuracy drop on object identification, while random neuron ablation shows no effect. This confirms that the identified "reliability-associated circuits" are not merely correlational artifacts but causally contribute to accurate answering.

We conclude that future research in trustworthy VLM design should shift focus from interpretability based on pixel-space grounding (which is often illusory) toward analyzing the latent dynamics of the language backbone, where the actual decision-making occurs.

References

[1] Jean-Baptiste Alayrac et al. Flamingo: A visual language model for few-shot learning. *NeurIPS*, 2022.

[2] Authors Chaudhury. Chameleonbench. *arXiv preprint arXiv:250x.xxxxx*, 2025.

[3] Wenliang Dai et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 2023.

[4] Chaoyou Fu et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[5] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL*, 2019.

[6] Authors Li. Hallucination benchmarks for large vision-language models. *arXiv preprint arXiv:230x.xxxxx*, 2023.

[7] Authors Li. Seed-bench: Benchmarking multimodal llms. *arXiv preprint arXiv:230x.xxxxx*, 2023.

[8] Junnan Li et al. Blip: Bootstrapping language-image pre-training. *ICML*, 2022.

[9] Authors Liu. Seeing and saying in vision-language models. *arXiv preprint arXiv:250x.xxxxx*, 2025.

[10] Haotian Liu et al. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[11] Authors Long. Chain-of-evidence for reliable multimodal reasoning. *arXiv preprint arXiv:250x.xxxxx*, 2025.

[12] nostalgebraist. Interpreting gpt: the logit lens. Blog post, 2020.

[13] Alec Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[14] Anna Rohrbach et al. Object hallucination in image captioning. In *EMNLP*, 2018.

[15] Authors Sahay. Compass: Calibrated multimodal safety and reliability. *arXiv preprint arXiv:250x.xxxxx*, 2025.

[16] Shikhar Shiromani et al. Hypocrisy gap in multimodal reasoning. *arXiv preprint arXiv:260x.xxxxx*, 2026.

[17] Authors Thomas. Promoralbench. *arXiv preprint arXiv:260x.xxxxx*, 2026.

[18] Xuezhi Wang et al. Self-consistency improves chain of thought reasoning in language models. *ICLR*, 2023.

[19] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP-IJCNLP*, 2019.

[20] Authors Yu. Mm-vet: Evaluating large multimodal models. *arXiv preprint arXiv:2308.02490*, 2023.

[21] Authors Zhou. Llava-bench: Evaluating vision-language assistants. *arXiv preprint arXiv:230x.xxxxx*, 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). All experiments use HuggingFace implementations on NVIDIA A100 GPUs.

Datasets: We evaluate on: (1) **POPE** [6] for object hallucination (Adversarial split, 1,000 samples), (2) **LLaVA-Bench** [21] 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 1.

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 ($p = 0.9$, $T = 0.7$). DBSCAN clustering uses $\epsilon = 1.5$, min_samples= 3. Full implementation details are in Appendix A.10.

A.3 Extended Analysis: The Ensemble Attention Probe

In Section 6, we briefly introduced the "Ensemble Attention Probe" (Internal ID: Idea 4). 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 the attention tensors $A^{(l)} \in \mathbb{R}^{H \times S}$ from all $L = 32$ layers of the Vicuna-7B backbone.

- **Input:** A concatenated vector of flattened attention maps from all layers:

$$x = \text{Concat}(\text{Vectorize}(A^{(1)}), \ldots, \text{Vectorize}(A^{(32)})) \tag{1}$$

- **Model:** A 3-layer Multi-Layer Perceptron (MLP) with ReLU activations and Dropout ($p = 0.1$).
- **Dimensions:** Input $d_{in} = 32 \times 576 \to 1024 \to 512 \to 1$ (Binary Classification).

Results & Comparison: Table 4 details the performance of various probes. While the Ensemble Attention Probe significantly outperforms random chance and simple visual entropy, it is still inferior to Self-Consistency. This reinforces our main finding: *generation dynamics (consistency) are a stronger signal than internal state snapshots.*

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.

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

Method	Type	AUROC	Cost (Inference)
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

Case Study: Consider an image with 3 baseball players.

- **Ground Truth:** 3
- **Model Prediction:** "Four"
- **Token Confidence (P_{tok}):** 92% (Very High)
- **Visual Clusters (C_k):** 3 distinct clusters (Correctly perceives 3 objects).

This dissociation highlights a "Symbolic Detachment." The visual encoder correctly identifies 3 regions (verified by $C_k = 3$), 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 it 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 5.5, we extracted the hidden states of the 576 visual tokens at every layer of the LLaVA-1.5-7B architecture. We calculated 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 5, the visual token representations remain remarkably dormant across the middle 25 layers of the network. Because the visual representations are not being actively updated during these middle layers, the spatial attention maps naturally stagnate (the "Early Locking" phenomenon). It is only in the final three layers that the model applies massive non-linear transformations to these features to extract confidence and generate text, directly corroborating our Logit Lens findings that true visual-linguistic grounding occurs at the very 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 "polling existence" task (POPE), 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

Figure 5: **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.

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 Cross-Model Experiment Details

A.8 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; **Qwen2-VL-7B** shows iterative re-integration cycles with strong late reliability separation. This appendix section is intended to balance interpretive depth across all three families under page constraints in the main paper.

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

Model Architectures:

- **LLaVA-1.5-7B**: 32 transformer layers, 32 attention 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.9 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. Table 5 presents the complete LLaVA analysis.

Table 5: **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	**+8.80**	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

A.10 Implementation and Hardware Details

Hardware: All experiments were conducted on a compute cluster provided by RunPod and Lambda Labs using NVIDIA A100 (80GB VRAM) GPUs, AMD EPYC 7742 64-Core CPUs, and 512 GB system RAM.

Software Environment: We used PyTorch 2.1.0 with CUDA 12.1, loaded official LLaVA, PaliGemma, and Qwen2-VL checkpoints via the HuggingFace `transformers` library, and extracted attention with PyTorch hooks (`register_forward_hook`) on decoder `MultiheadAttention` modules in each family's multimodal-integration regime (e.g., late layers for LLaVA, architecture-adjusted regions for PaliGemma and Qwen2-VL).

A.11 Discussion Extensions: Cross-Family Interpretation and Efficiency Trade-offs

A.12 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 capacity per family), rather than assuming a one-size-fits-all late-layer template.

A.13 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× 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.14 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 distributed circuit rather than individual "truth units." Furthermore, the effect is 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.