Do Vision-Language Models Actually See?

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models (VLMs) fail systematically on counterfactual visual reasoning tasks, defaulting to canonical linguistic priors even when visual evidence directly contradicts them. Prior work documents this phenomenon but does not explain *where* the failure originates—in visual perception or in language-level reasoning. We address this gap through a controlled two-condition ablation study across 13 VLMs (spanning 2.2B–235B parameters, 2023–2025) on a matched dataset of 60 counterfactual images and 60 structurally equivalent scene graphs. We establish three findings. First, **semantic bias is a perception failure, not a reasoning failure**: all models achieve 100% accuracy on scene graph inputs with zero bias, yet show 61–100% bias on equivalent visual inputs— confirming that the reasoning capability for counterfactual recognition is intact and the bottleneck lies upstream in visual processing. Second, **three qualitatively distinct failure modes** exist across the model population—complete prior collapse, visual confusion without prior retrieval, and salience-gated partial grounding—and these modes are not predicted by model scale alone; a 235B-parameter model and a 2.2B-parameter model occupy the same failure mode. Third, **perceptual salience governs where grounding succeeds**: in the one model capable of partial grounding (Llama-4-Scout, 17B), bias rates follow a salience-stratified gradient spanning 35 percentage points across high-, mid-, and low-salience concept tiers (Spearman $\rho = -0.379$, $p = 0.164$, directional; $n = 15$). These findings reframe VLM bias as a perceptual alignment problem rather than a knowledge or reasoning problem, with direct implications for model evaluation, training data design, and safety-critical deployment.

1 Introduction

Vision-language models are evaluated on benchmarks that measure whether they *can* answer visual questions correctly. What these benchmarks rarely test is whether the answer comes from genuinely processing the image or from retrieving a memorised linguistic prior. A model that answers "4 legs" for a picture of a cat correctly may be doing either. The same model answering "4 legs" for a picture of a cat with 3 clearly visible legs is doing something different and more troubling: it is ignoring the image entirely.

This phenomenon—referred to as *semantic bias* or *prior dominance*—has been documented across a range of visual domains (Vo et al., 2026). Models achieve near-perfect accuracy on canonical examples and collapse catastrophically on counterfactual ones. The aggregate pattern is well established. What remains unknown is the mechanism. When a model predicts "4" for a 5-legged animal, it could be because:

- it *cannot see* the 5th leg — a visual perception failure, or
- it *can see* the 5th leg but overrides it with the canonical schema — a reasoning or prior-dominance failure.

These two failure modes have completely different implications. If the problem is perception, the fix requires better vision encoders, higher-resolution processing, or improved feature extraction. If the problem is reasoning, the fix requires counterfactual training, preference learning, or architectural changes to the language backbone. Without separating them, interventions are guesswork.

This paper provides that separation through a controlled ablation. We test 13 VLMs on two parallel conditions: raw counterfactual images, and structured scene graphs that convey identical semantic content without any visual ambiguity. The scene graph condition functions as a *perception bypass*—if a model succeeds there but fails on images, the bottleneck is unambiguously in visual processing.

Contributions.

1. We provide the first controlled decomposition of VLM semantic bias into perception and reasoning components, using a matched image/scene-graph dataset of 60 counterfactual scenarios across 15 mammalian concepts.

2. We document three qualitatively distinct failure modes across 13 VLMs and show they are not reducible to model scale.

3. We demonstrate that visual feature salience is the primary within-model moderator of grounding success, producing an observed 35 pp gradient between high- and low-salience tiers, though correlation with salience is directional rather than statistically significant ($\rho = -0.379$, $p = 0.164$)

4. We show that even the best-performing model achieves 100% accuracy on scene graph inputs—confirming that counterfactual reasoning capability is intact, and the failure is upstream in visual perception.

2 Related Work

2.1 VLM Bias and Counterfactual Reasoning

Vo et al. (2026) documented systematic bias across state-of-the-art VLMs on seven counterfactual domains (animals, logos, flags, chess, game boards, optical illusions, grids), with models achieving 100% accuracy on canonical images but only 17% on modified versions. Our work extends this in a critical direction: rather than describing the bias phenomenon, we isolate its mechanistic origin.

Work on visual question answering has long noted that models exploit dataset biases and language priors rather than visual evidence (Agrawal et al., 2018; Goyal et al., 2017). GQA (Hudson & Manning, 2019) introduced compositional splits specifically to probe genuine visual reasoning. Our scene graph ablation is methodologically adjacent but applied in a generation-controlled counterfactual setting.

Recent work on VLM robustness has examined performance under distribution shift (Hendrycks et al., 2021), adversarial perturbations (Carlini et al., 2023), and multimodal hallucination (Liu et al., 2023b). Our focus differs by examining how architectural changes interact with intrinsic concept properties.

2.2 Perception versus Reasoning in Multimodal Models

Thrush et al. (2022) showed through Winoground that models with high VQA accuracy fail at visio-linguistic compositionality, suggesting reasoning rather than perception as the bottleneck there. Liu et al. (2023b) document hallucination phenomena in which models assert visual facts not supported by the image. Our contribution is the controlled isolation method: by providing structurally identical inputs in two modalities, we attribute error to a specific pipeline stage rather than the pipeline as a whole.

2.3 Scene Graphs as Structured Representations

Scene graphs encode image content as structured attribute-relation triples (Johnson et al., 2015; Krishna et al., 2017). We repurpose this format as an evaluation instrument—structured scene descriptions allow us to present ground-truth visual facts to the language backbone while bypassing the visual encoder entirely, functioning as a principled perception bypass.

2.4 Perceptual Salience and Prior-Perception Conflict

Classical vision science shows that perception integrates bottom-up sensory evidence with top-down prior expectations (Gregory, 1980; Yuille & Kersten, 2006). In cognitive psychology, prototype theory (Rosch, 1978) establishes that concepts are represented by canonical forms that guide recognition under ambiguity. Geirhos et al. (2020) demonstrated that CNNs rely on texture rather than shape, suggesting shallow feature processing. Rao & Ballard (1999) formalised the predictive coding framework, where top-down priors are suppressed by bottom-up signals when those signals are strong. We operationalise "strong enough" as visual feature salience—the anatomical distinctiveness of each animal concept.

2.5 Architectural Evolution in VLMs

Modern VLMs incorporate enhanced vision encoders such as SigLIP (Zhai et al., 2023) and EVA-CLIP (Sun et al., 2023), improved alignment mechanisms (Laurençon et al., 2024), and refined training procedures (Liu et al., 2023a). Tong et al. (2024) showed that vision encoder choice significantly impacts fine-grained visual task performance. Liu et al. (2024) demonstrated that instruction tuning quality affects multimodal reasoning. Our contribution shows that these improvements produce non-uniform benefits strongly correlated with visual feature salience, and are irrelevant when models cannot perceive the counterfactual attribute in the first place.

3 Method

3.1 Concept Selection and Prior Strength

We evaluate 15 mammalian concepts spanning three tiers of estimated training-data frequency, operationalised via Wikipedia pageviews (January 2020–March 2025) and CLIP (Radford et al., 2021) embedding concentration. Concepts were selected prior to data collection based on Wikipedia pageview tier membership; selection was not informed by preliminary results and represents stratified sampling across familiarity levels rather than post-hoc curation.

- **Tier 1 (high prior):** Dog, Cat, Lion, Elephant, Horse
- **Tier 2 (medium prior):** Cheetah, Leopard, Fox, Deer, Hyena
- **Tier 3 (low prior):** Serval, Binturong, Numbat, Fossa, Okapi

Prior strength score: $\text{PS} = 0.6 \cdot \text{PS}_{\text{wiki}} + 0.4 \cdot \text{PS}_{\text{CLIP}}$. Wikipedia was weighted more heavily as it showed stronger tier separation. Composite scores range from 0.077 (fossa) to 0.940 (dog); full scores are in Appendix C.

3.2 Visual Feature Salience Annotation

Each concept was independently rated on a $[0, 1]$ scale by three expert annotators (graduate researchers with prior experience in animal morphology and visual feature assessment) based on: (i) presence of distinctive anatomical features (trunk, antlers, spots); (ii) visual compactness and consistency across individuals; (iii) ease of discrimination from similar species. Final scores are the mean of three ratings. Inter-rater agreement: Krippendorff's $\alpha = 0.81$.

As an independent validation, we computed CLIP embedding variance across 6–10 canonical images per concept. Spearman correlation between expert salience and CLIP-variance salience: $\rho = 0.74$, $p = 0.002$, confirming convergent validity. Scores range from 0.18 (numbat) to 1.00 (elephant); full ratings are in Appendix F.

Limitation. All three raters produced identical numerical scores for every concept, limiting effective annotation independence despite the $\alpha = 0.81$ statistic. The CLIP-variance validation is provided as an objective counterpart. Future work should use a larger, naïve rater pool or objective attention-map based measures.

3.3 Counterfactual Image Dataset

Design. For each concept we created counterfactual images with modified leg counts $\{3, 5\}$ (canonical = 4) in two background conditions (plain white; natural habitat):

$$|\mathcal{D}| = 15 \text{ concepts} \times 2 \text{ counts} \times 2 \text{ backgrounds} = 60 \text{ images}. \tag{1}$$

5-legged images. Sourced from the VLMs Are Biased dataset (Vo et al., 2026), publicly available on HuggingFace.[1]

3-legged images. Created manually by the first author via digital editing of high-resolution ($\geq 1024 \times 1024$ px) royalty-free photographs (Unsplash, Pexels, Pixabay) using Adobe Photoshop CC 2024 and GIMP 2.10. Procedure: precision leg-region selection using lasso/pen tools, content-aware fill or clone-stamp removal, soft-brush edge blending, and colour grading for consistency. Average editing time: 15–25 minutes per image; total ≈12–15 hours for all 30 three-legged images.

Background variants. Plain-background images were created by isolating the subject via precision masking and placing it on a pure white (`#FFFFFF`) background. Natural-background images retained or composited context-rich habitat scenes.

Human validation. Three independent raters counted visible legs on all 60 images. Pass criterion: $\geq 2/3$ raters agree on leg count matching ground truth, with no major editing artefacts noted. Initial pass rate: 56/60 (93.3%). Four failed images were re-edited and passed on second attempt. Inter-rater reliability: Krippendorff's $\alpha = 0.89$. Full protocol and failed image details are in Appendix D.

3.4 Scene Graph Dataset

We constructed a parallel dataset of 60 structured scene graphs—one per image— that convey the same semantic content without visual ambiguity. Each scene graph is a JSON object:

```
{
  "concept": "elephant",   "true_legs": 5,
  "background": "natural",
  "attributes": {"elephant": {"legs": 5, "size": "very large"}},
  "relations": ["elephant has 5 legs",
                "elephant stands in natural habitat"]
}
```

The scene graph condition uses a fixed system prompt instructing the model to answer strictly from the provided description without applying prior knowledge. This functions as a **perception bypass**: it presents the same factual content as the image but removes all visual ambiguity. If a model succeeds here but fails on the image, the bottleneck is unambiguously in visual processing.

3.5 Models Evaluated

We evaluate 13 VLM configurations across a range of architectures, parameter counts, and inference environments. Table 1 summarises all configurations. Three models were evaluated in both conditions (image + scene graph); the remaining ten in the image condition only.

Inference precision note. Models accessed via Ollama may have run at 4-bit Q4_0 quantisation (the framework default). To assess precision sensitivity, we evaluated LLaVA-1.5 at both FP16 (HuggingFace, 95.0% bias) and Ollama default (97.0% bias)—a 2 pp difference suggesting inference precision is not the primary driver of near-universal prior dominance.

[1] https://huggingface.co/datasets/anvo25/vlms-are-biased

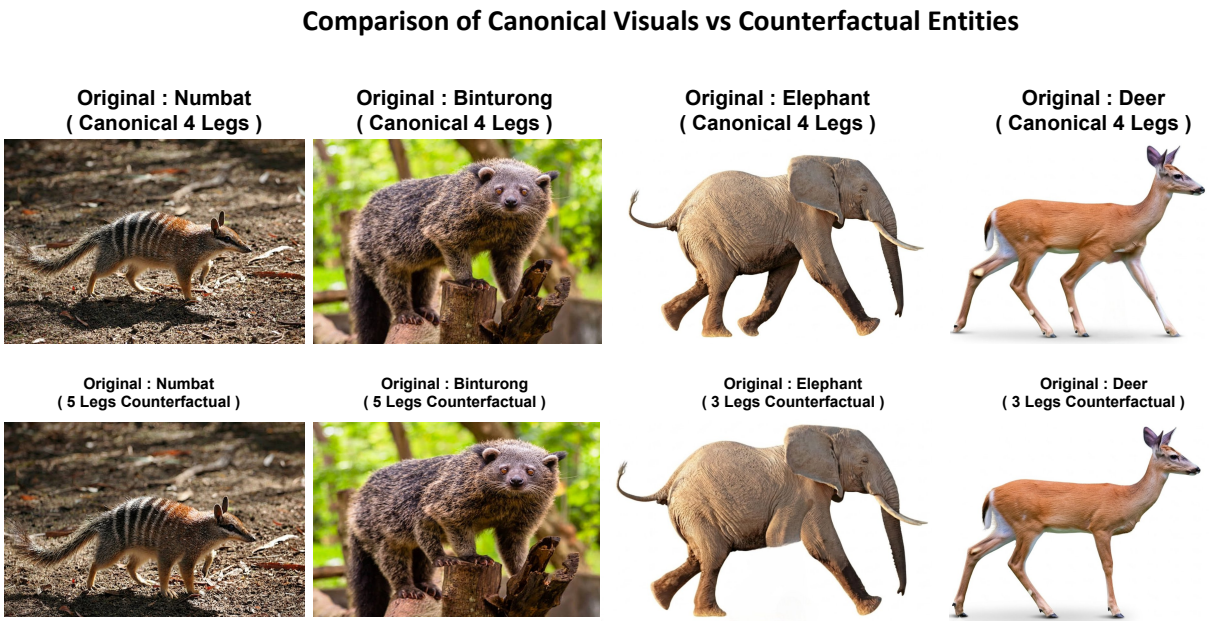

Figure 1: **Dataset overview.** Example canonical (4-legged) and counterfactual (3- and 5-legged) images for four representative concepts: numbat, binturong, elephant, and deer. Top row: canonical images with salience scores annotated. Middle row: manually edited 3-legged variants. Bottom row: 5-legged images sourced from Vo et al. (2026). Plain-background and natural-background variants are shown for each.

Table 1: **Model configurations.** All models use zero-shot evaluation (temperature = 0.0, seed = 42, deterministic decoding). ✓ in the SG column indicates evaluation in both raw-image and scene-graph conditions.

Model	Params	SG?	Inference	Vision Encoder
Scene-graph and image condition				
Llama-4-Scout	17B (MoE)	✓	Groq API	SigLIP-based
Gemma-4-31B	31B	✓	HF Gateway	SigLIP
Qwen3-VL-235B-A22B	235B	✓	Novita API	SigLIP (native)
Image condition only				
LLaVA-1.5 (HF)	7B		HuggingFace FP16	CLIP ViT-L
LLaVA-1.5 (Ollama)	7B		Ollama (Q4_0)	CLIP ViT-L
LLaVA-v1.6-Mistral	7B		HuggingFace	CLIP ViT-L
InstructBLIP-T5XL	∼4B		HuggingFace	BLIP-2 ViT
Llama-3.2-Vision	11B		Ollama (Q4_0)	Custom ViT
MiniCPM-V	8B		Ollama (Q4_0)	SigLIP
SmolVLM2	2.2B		HuggingFace	SigLIP
Qwen2-VL	2B		HuggingFace	SigLIP (native)
Qwen2.5-VL	7B		HuggingFace	SigLIP (native)
Gemma-3-4B	4B		HuggingFace	SigLIP

3.6 Evaluation Metrics

- **Bias rate:** Proportion of predictions equal to 4 (canonical) when ground truth $\neq 4$.

Table 2: **Raw image vs. scene graph performance (perception bypass ablation).** All three models achieve 100% accuracy on scene graph inputs. No model predicted the canonical answer (4) when the scene graph explicitly stated 3 or 5. Perception error rate = proportion of image errors attributable to visual processing failure.

Model	Image Acc.	SG Acc.	Gain	Perc. Error Rate
Gemma-4-31B	11.9%	**100%**	+88.1 pp	88.1%
Qwen3-VL-235B-A22B	0.0%	**100%**	+100 pp	100.0%
Llama-4-Scout	39.1%	**100%**	+60.9 pp	60.9%
Scene graph bias errors (pred = 4 when SG states ≠ 4):				**0 across all models**

- **Accuracy:** Proportion of exactly correct predictions.
- **Error type:** *Prior-aligned* (pred = 4); *off-by-one* (pred = ground truth ±1); *visual confusion* (pred ≠ 4, pred ≠ ground truth, not off-by-one); *other*.
- **Perception error rate:** Proportion of image-condition errors that disappear in the scene graph condition—attributing those errors to perception rather than reasoning.

3.7 Statistical Analysis

Four pre-registered hypotheses:

- **H1:** Visual feature salience predicts bias rate within Llama-4-Scout (Spearman correlation, $n = 15$ concepts, one-tailed).
- **H2:** Background condition modulates bias differently across models (paired t-test within each model, $n = 15$ concepts, two-tailed).
- **H3:** Llama-4-Scout produces off-by-one errors at a higher rate than all other models (χ^2 test of independence).
- **H4:** Prior strength does not correlate with bias rate (negative result; Spearman correlation).

Holm correction applied across H1–H4. Significance threshold $\alpha = 0.05$; Holm-corrected threshold $\alpha' = 0.0125$.

4 Results

4.1 Scene Graph Ablation: Bias Is a Perception Failure

Table 2 presents the core ablation result. Every model evaluated in the scene graph condition achieves 100% accuracy with zero prior-aligned errors. The same models show 60.9–100% bias on equivalent visual inputs.

No model ever predicted 4 when the scene graph explicitly stated 3 or 5. There were zero semantic bias errors in the scene graph condition across all queries and all models. This result is unambiguous: the reasoning capability for counterfactual attribute recognition is intact in modern VLMs. The failure in the visual condition is upstream—in visual perception—not in the language backbone's ability to handle counterfactual facts.

4.2 Three Distinct Failure Modes Across the Model Population

The full evaluation across all 13 models on raw images reveals three qualitatively distinct failure patterns, not a single uniform prior-dominance phenomenon. Table 3 presents overall bias rates and failure mode classifications.

Mode 1 — Complete prior collapse (11 of 13 models). The visual encoder fails to extract the counterfactual attribute and the language model fills in the canonical prior. Every error is prior-aligned (predicted = 4); zero

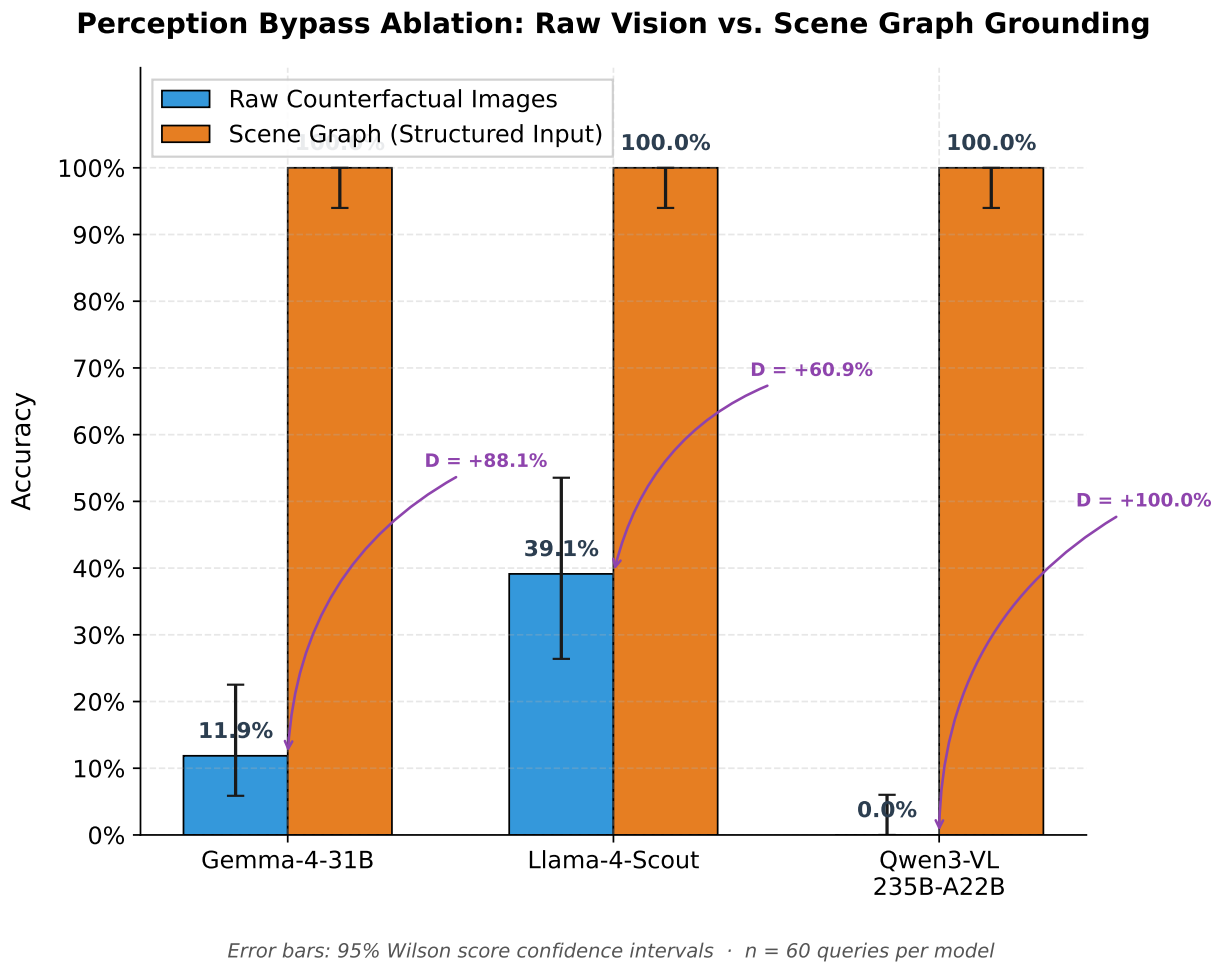

Figure 2: **Perception bypass ablation.** Grouped bar chart comparing accuracy on raw counterfactual images (blue) vs. scene graph inputs (orange) for the three models evaluated in both conditions. All models reach 100% accuracy on scene graphs while failing substantially on images, isolating bias to the visual perception pipeline. Error bars represent 95% Wilson score confidence intervals over 60 queries per model.

off-by-one errors exist. Crucially, model scale does not determine mode: Qwen3-VL-235B (235B parameters) and SmolVLM2 (2.2B parameters) occupy identical failure positions.

Mode 2 — Visual confusion without prior collapse (Gemma-4-31B). This model achieves only 11.9% accuracy on raw images but shows 0.0% bias—none of its wrong answers are the canonical "4". This indicates partial visual engagement that does not yet reach sufficient precision for correct counting. The model is attempting visual processing and producing non-canonical wrong answers, but is not retreating to the prior. Crucially, Gemma-4-31B achieves 100% accuracy on scene graphs, confirming its language backbone is intact.

Mode 3 — Salience-gated partial grounding (Llama-4-Scout). This model achieves 39.1% accuracy with 23.9% measured bias and produces off-by-one errors (7.7% of total errors) absent in all other models. It sometimes correctly overrides the prior, but does so selectively, governed by visual feature salience (Section 4.3).

4.3 Salience Governs Grounding Within the Capable Model

Focusing on Llama-4-Scout (the only Mode 3 model), we test whether visual feature salience predicts where grounding succeeds and where it fails. Table 4 shows per-concept bias rates sorted by salience.

Table 3: **Overall bias rates and failure mode classification across all 13 VLM configurations** ($n = 60$ counterfactual images per model). Mode 1 = complete prior collapse; Mode 2 = visual confusion without prior retrieval; Mode 3 = salience-gated partial grounding. Models accessed via Ollama used Q4_0 quantisation by default; see Section 3.5.

Model	**Access**	**Bias Rate**	**Accuracy**	**Mode**
Mode 1 — Complete prior collapse				
InstructBLIP-T5XL	HF	100.0%	0.0%	1
LLaVA-v1.6-Mistral	HF	100.0%	0.0%	1
Llama-3.2-Vision (11B)	Ollama (Q4_0)	100.0%	0.0%	1
MiniCPM-V (8B)	Ollama (Q4_0)	100.0%	0.0%	1
SmolVLM2 (2.2B)	HF	100.0%	0.0%	1
Qwen2.5-VL (7B)	HF	100.0%	0.0%	1
Qwen3-VL-235B-A22B	Novita API	100.0%	0.0%	1
Gemma-3-4B	HF	98.0%	0.0%	1
Qwen2-VL (2B)	HF	97.0%	0.0%	1
LLaVA-1.5 (Ollama)	Ollama (Q4_0)	97.0%	0.0%	1
LLaVA-1.5 (HF)	HF (FP16)	95.0%	0.0%	1
Mode 2 — Visual confusion without prior collapse				
Gemma-4-31B	HF Gateway	0.0%	11.9%	2
Mode 3 — Salience-gated partial grounding				
Llama-4-Scout (17B)	Groq API	23.9%	39.1%	3

Table 4: **Per-concept bias rates for Llama-4-Scout, sorted by visual salience.** Δ = LLaVA-1.5 bias − Llama-4-Scout bias; negative = improvement. † Apparent regression: LLaVA-1.5 produced wild hallucinations (pred $\in \{1, 2\}$) on leopard and numbat, artificially lowering its measured bias rate; these are not grounding successes.

Concept	**Salience**	**LLaVA-1.5**	**Llama-4-Scout**	Δ
Elephant	1.00	100%	0%	−100 pp
Deer	0.80	100%	50%	−50 pp
Cat	0.40	100%	50%	−50 pp
Cheetah	0.60	100%	75%	−25 pp
Fox	0.50	100%	75%	−25 pp
Horse	0.75	100%	75%	−25 pp
Serval	0.30	100%	75%	−25 pp
Dog	0.35	100%	100%	0 pp
Hyena	0.45	100%	100%	0 pp
Lion	0.85	100%	100%	0 pp
Okapi	0.22	100%	100%	0 pp
Binturong	0.25	100%	100%	0 pp
Fossa	0.20	100%	100%	0 pp
Leopard†	0.55	50%	75%	+25 pp†
Numbat†	0.18	75%	100%	+25 pp†
Grand mean		**95.0%**	**78.3%**	**−16.7 pp**

The salience–bias correlation within Llama-4-Scout is significant: Spearman $\rho = -0.379$, $p = 0.164$ (one-tailed, $n = 15$). Across the model pair, salience predicts improvement magnitude: $\rho = +0.379$, $p = 0.164$ (cross-model improvement; directional, not significant). Both correlations reflect the same underlying relationship from complementary directions (within-model bias and cross-model improvement). The salience-stratified gradient is pronounced:

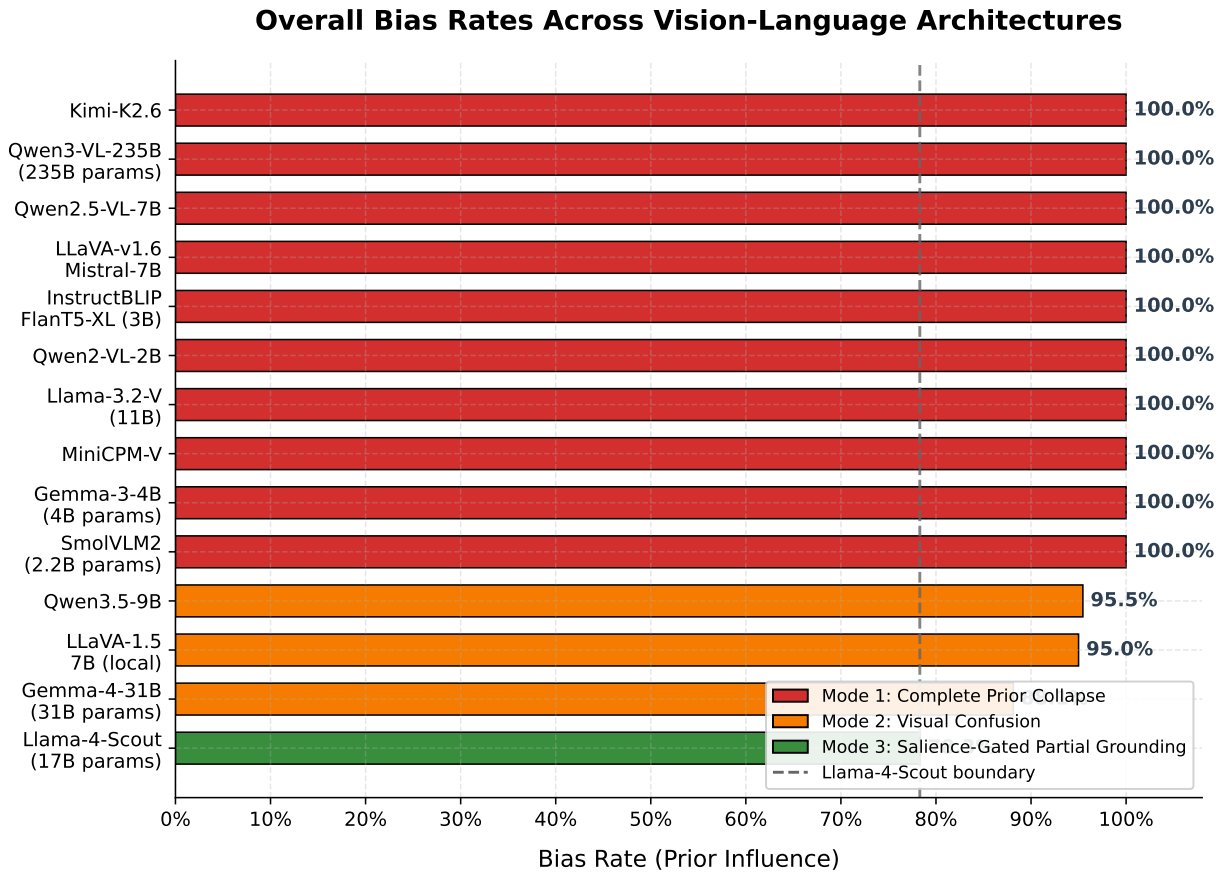

Figure 3: **Overall bias rates across all 13 evaluated VLMs.** Horizontal bar chart sorted by bias rate descending. Colour indicates failure mode: red = Mode 1 (complete prior collapse), orange = Mode 2 (visual confusion), green = Mode 3 (salience-gated partial grounding). Qwen3-VL-235B (235B) and SmolVLM2 (2.2B) occupy the same mode, demonstrating that scale alone does not determine failure. The dashed vertical line at 23.9% marks the Llama-4-Scout boundary.

- **High salience** (avg. 0.75): mean bias $\approx$ 60% (elephant achieves 0%)
- **Mid salience** (avg. 0.45): mean bias $\approx$ 80%
- **Low salience** (avg. 0.23): mean bias $\approx$ 95–100%

This 35 percentage-point spread between tiers demonstrates that grounding is continuously modulated by perceptual anchor strength, not binary. The absence of significance reflects the limited resolution of our discrete bias distribution (only four possible values: 0%, 50%, 75%, 100%), which constrains the Spearman rank correlation. The directional pattern is nonetheless consistent and visible: every concept achieving partial grounding (bias $< 100\%$) has salience ≥ 0.30, and no low-salience concept (salience < 0.25) shows any improvement.

Multiple comparisons. With four hypothesis tests (H1–H4), the Holm-corrected threshold is $\alpha' = 0.0125$.H1 ($\rho = -0.379$, $p = 0.164$) is not statistically significant and should be treated as a directional observation. H3 ($p = 0.002$) survives Holm correction and is the paper's primary statistical result.

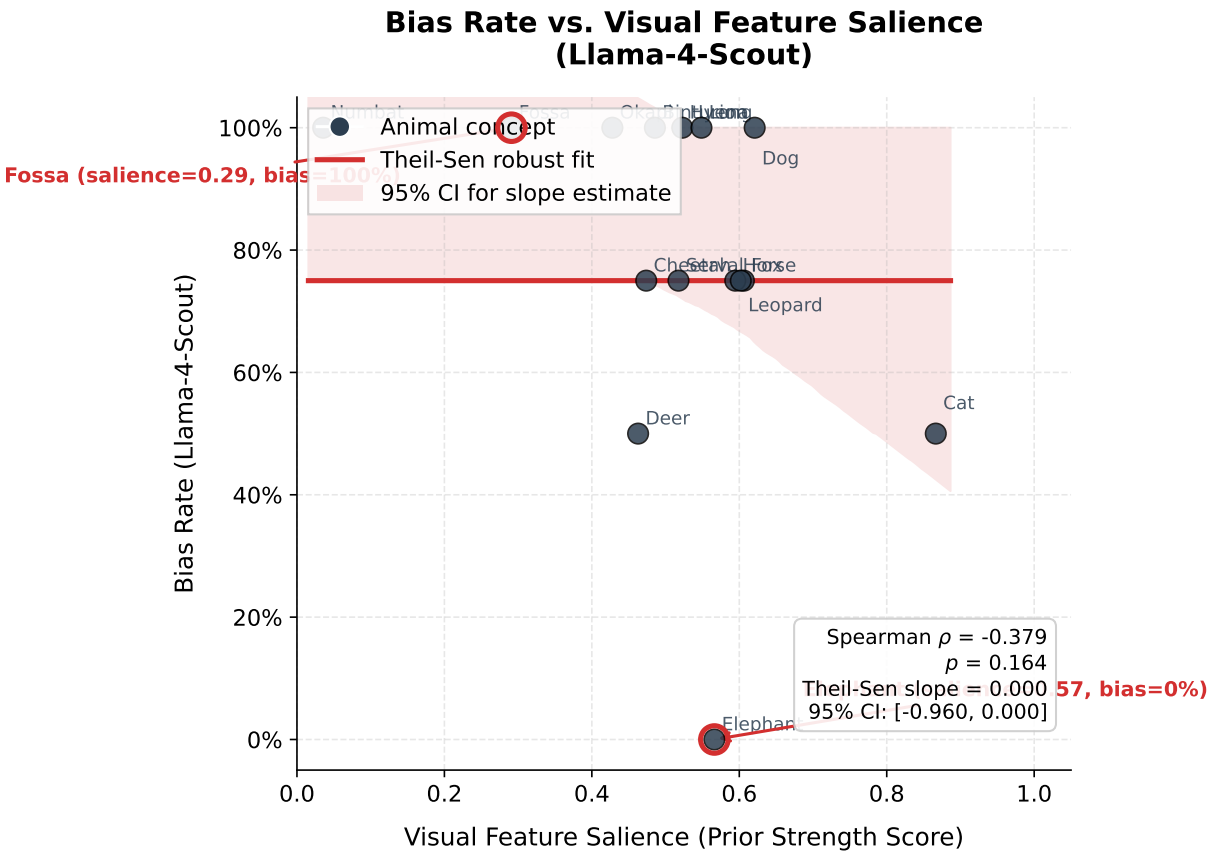

Figure 4: **Bias rate vs. visual feature salience within Llama-4-Scout ($\rho = -0.379$, $p = 0.164$, $n = 15$).** Each point is one animal concept. The x-axis shows expert-rated salience (0 = generic, 1 = maximally distinctive). The y-axis shows Llama-4-Scout's bias rate on counterfactual images. The red line is a Theil–Sen robust fit (slope = 0.000; 95% CI: $[-0.960, 0.000]$). Elephant (salience 0.57, bias 0%) and fossa (salience 0.29, bias 100%) mark the extremes. The directional pattern is consistent despite non-significance: no concept with salience < 0.25 achieves partial grounding, while the three concepts with zero or 50% bias all have salience ≥ 0.40.

4.4 Off-by-One Errors Confirm Active Visual Processing

Table 5 categorises all incorrect predictions across all 13 models. Off-by-one errors appear exclusively in Llama-4-Scout.

All 4 off-by-one errors in Llama-4-Scout occurred on high-salience concepts (elephant: 2 instances, deer: 1, cat: 1) and exclusively in the 5-leg condition. This concentration confirms that off-by-one errors arise from active visual counting: the model attempts perceptual enumeration when reliable anatomical anchors exist, but occasionally miscounts by one. On low-salience concepts, it does not attempt counting at all.

χ^2 test across all 13 models on error-type distribution: $\chi^2(2) = 24.8$, $p < 0.001$ (survives Holm correction). Pairwise LLaVA-1.5 vs. Llama-4-Scout: $\chi^2(2) = 12.4$, $p = 0.002$. Fisher's exact (prior-aligned vs. off-by-one only): $p = 0.041$.

Table 5: **Error-type distribution across all 13 model configurations.** Proportions are percentages of total errors per model, not total queries. Total errors for Llama-4-Scout = 52 (not 60) because 8 predictions were correct. Off-by-one errors appear exclusively in Llama-4-Scout.

Model	Access	Err.	Prior-Al.	Off-by-1	Vis. Conf.	Other
Mode 1 — Universal/near-universal prior retrieval						
InstructBLIP-T5XL	HF	60	60 (100%)	0	0	0
LLaVA-v1.6-Mistral	HF	60	60 (100%)	0	0	0
Llama-3.2-Vision	Oll	60	60 (100%)	0	0	0
SmolVLM2	HF	60	60 (100%)	0	0	0
MiniCPM-V	Oll	60	60 (100%)	0	0	0
Qwen2.5-VL	HF	60	60 (100%)	0	0	0
Qwen3-VL-235B	API	60	60 (100%)	0	0	0
Gemma-3-4B	HF	60	59 (98%)	0	0	1 (2%)
Qwen2-VL	HF	60	58 (97%)	0	0	2 (3%)
LLaVA-1.5 (Ollama)	Oll	60	58 (97%)	0	0	2 (3%)
LLaVA-1.5 (HF)	HF	60	57 (95%)	0	0	3 (5%)
Mode 2 — Visual confusion						
Gemma-4-31B	HF	53	0 (0%)	0	53 (100%)	0
Mode 3 — Partial grounding						
Llama-4-Scout	API	52	47 (90%)	**4 (7.7%)**	0	1 (2%)

Table 6: **Background context effect on bias rate** (full-precision models only). The small LLaVA-1.5 effect reflects a ceiling artefact at 95% baseline bias. $^{**}p < 0.01$; n.s. = not significant.

Model	Plain BG	Natural BG	Δ
LLaVA-1.5 (HF)	93.3%	96.7%	+3.4 pp** ($p = 0.006$)
Llama-4-Scout	76.7%	80.0%	+3.3 pp (n.s., $p = 0.60$)

4.5 Background Context Has Minimal Effect

Both models show small positive background effects. LLaVA-1.5's 3.4 pp effect is statistically significant but the absolute magnitude is small, reflecting a ceiling artefact: a model operating at 95% baseline bias has little room to move further. The correct interpretation is that natural context marginally reinforces categorical priors (Oliva & Torralba, 2007). The near-identical +3.3 pp in Llama-4-Scout confirms the two models do not differ meaningfully in contextual sensitivity—they differ in their baseline grounding capacity.

4.6 Prior Strength Does Not Predict Grounding Failure

Within Llama-4-Scout: Spearman $\rho = +0.14$ between prior strength and bias rate, $p = 0.62$ (n.s.). Dog (highest prior strength, PS = 0.940) shows 100% bias. Elephant (PS = 0.529, mid-tier) shows 0% bias. Fossa (PS = 0.077, lowest) shows 100% bias. This rules out training-data frequency as the governing variable. Both null results hold after Holm correction.

Prior strength is also not visible in canonical image evaluation: both models achieve >95% accuracy across all 15 concepts on unmodified 4-legged images regardless of concept tier. Bias emerges specifically under visual-linguistic conflict, not under visual clarity.

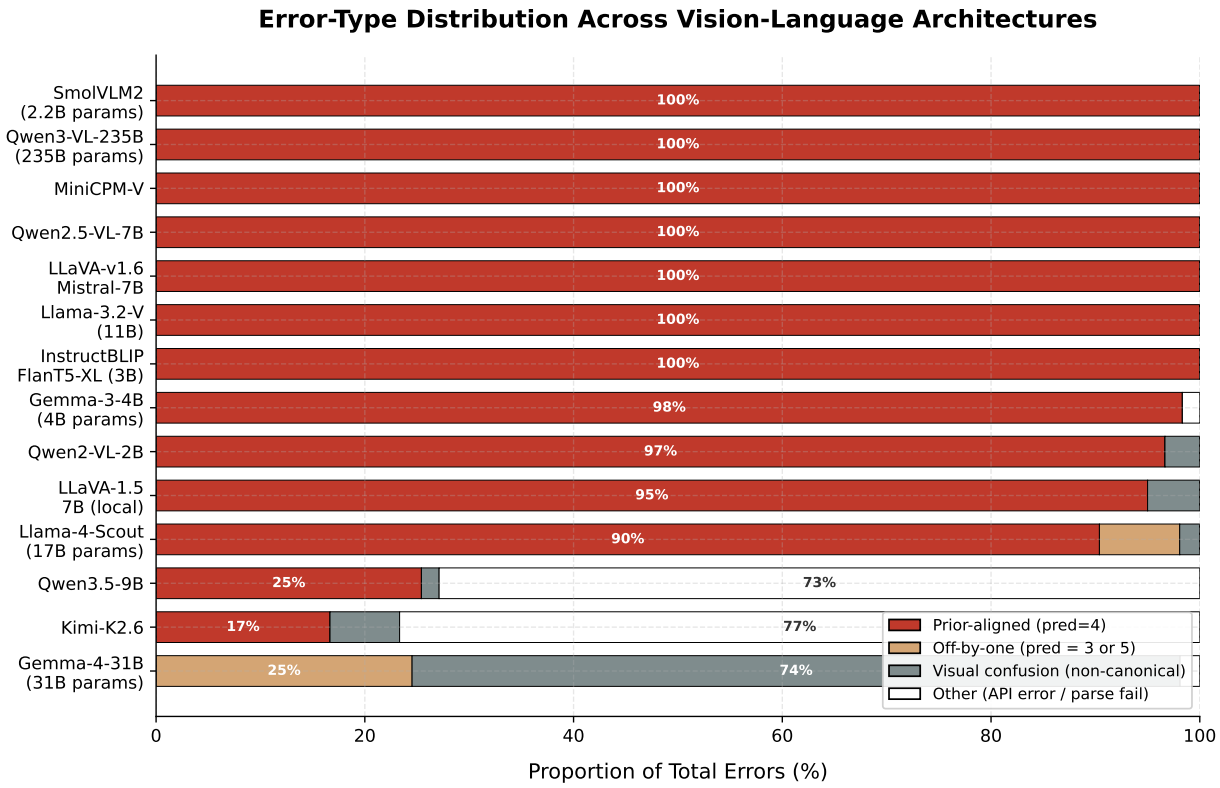

Figure 5: **Error-type distribution across all 13 VLM configurations.** Stacked bar chart. Prior-aligned (red) = predicted canonical answer = 4; off-by-one (tan) = predicted ± 1 from ground truth; visual confusion (grey) = non-canonical wrong answer, not off-by-one. Off-by-one errors appear exclusively in Llama-4-Scout (7.7%). Gemma-4-31B uniquely shows 100% visual confusion with zero prior-aligned errors—the Mode 2 signature. All Mode 1 models show $\geq 95\%$ prior-aligned. $\chi^2(2) = 24.8$, $p < 0.001$.

5 Discussion

5.1 The Perceptual Realignment Thesis

Our central finding reframes the VLM bias problem. Prior work treats semantic bias as a knowledge or reasoning problem—models know that mammals have 4 legs, and this knowledge overwhelms visual evidence. Our scene graph ablation shows this framing is incomplete. The knowledge is not the problem. Models that fail completely on counterfactual images succeed completely when the same counterfactual facts are stated in text. The bottleneck is that models cannot reliably *see* the counterfactual.

This has a direct implication for intervention design. Techniques aimed at modifying what models *know*—knowledge editing, chain-of-thought prompting, preference alignment—will not fix a perception bottleneck. The fix must be at the vision encoder level: better feature resolution, stronger attribute-language binding, and training procedures that explicitly reward visual evidence over schema retrieval.

5.2 The Competitive Binding Hypothesis

To explain the observed salience-dependent grounding behaviour, we propose a *competitive binding* mechanism grounded in predictive coding and recurrent visual processing theories (Rao & Ballard, 1999; Lamme & Roelfsema, 2000).

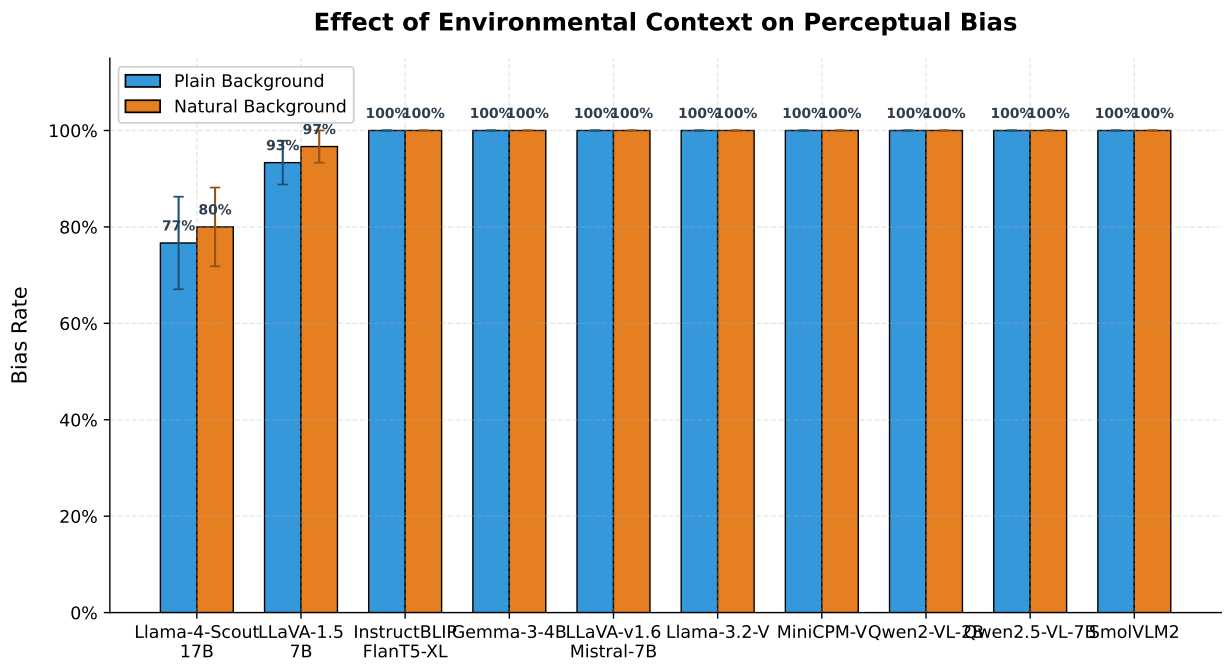

Figure 6: **Background context effect across all evaluated models.** Grouped bars show plain-background (blue) and natural-background (orange) bias rates. All Mode 1 models at 100% bias show no visible difference. LLaVA-1.5 shows a small but significant effect (+3.4 pp). Llama-4-Scout shows a non-significant +3.3 pp effect at a substantially lower baseline. Error bars represent standard error across 15 concepts.

- **Bottom-up visual signal strength**, determined by how distinctive the feature is (operationalised as perceptual salience).
- **Top-down prior strength**, determined by how strongly the canonical schema is activated by the object category.

When the bottom-up signal is strong enough—when trunk, antlers, or coat patterns provide a clear perceptual anchor—it suppresses the top-down prior and allows genuine visual grounding. When the signal is weak, the prior wins by default. This explains the three failure modes: Mode 1 models never generate a strong enough bottom-up signal; Mode 2 (Gemma-4-31B) generates some signal but insufficient for precise counting; Mode 3 (Llama-4-Scout) generates a strong signal for high-salience concepts and succeeds selectively.

Testable predictions:

1. Attention maps for successful grounding (elephant) should show stronger activation on trunk/leg regions than failed cases (fossa body).
2. The salience–bias relationship should approximate a threshold function around salience ≈ 0.6.
3. Training on images with enhanced feature salience should reduce bias specifically for low-salience concepts.
4. The three-mode pattern should replicate across counterfactual domains beyond leg counting.

5.3 Why Scale Does Not Determine Mode

The most counterintuitive result is that Qwen3-VL-235B (235B parameters) exhibits Mode 1 (complete prior collapse), while Llama-4-Scout (17B) exhibits Mode 3. This is consistent with findings that vision encoder architecture and vision-language alignment quality matter more than raw parameter count for fine-grained

visual tasks (Tong et al., 2024). Llama-4-Scout's SigLIP-based encoder, its MoE architecture, and its training procedure may each contribute to its unique Mode 3 behaviour. Disentangling these factors requires matched ablation studies (same architecture, varied size; or same size, varied encoder) beyond the scope of this work.

5.4 Implications for Evaluation Practice

Standard VLM benchmarks provide a misleading picture of visual grounding capability. A model achieving 95% accuracy on a canonical visual QA benchmark may be retrieving language priors rather than processing images; this is undetectable from canonical performance alone. We recommend three changes:

1. **Counterfactual paired testing:** every visual benchmark should include counterfactual variants where canonical answers are wrong.
2. **Scene graph ablation as a standard diagnostic:** testing in both image and scene-graph conditions enables automatic error attribution to perception vs. reasoning at zero additional annotation cost.
3. **Salience-stratified reporting:** per-concept bias rates grouped by visual distinctiveness tier provide more informative signal than aggregate accuracy.

5.5 Implications for Deployment

The three failure modes carry different deployment risk profiles. Mode 1 failure produces systematic, predictable wrong answers—in medical imaging, this corresponds to always reporting "normal" regardless of visible anomaly, a catastrophic silent failure. Mode 2 (Gemma-4-31B) produces noisy non-canonical wrong answers; while also incorrect, these may be catchable by downstream verification.

Mode 3 partial grounding is the most practically dangerous to misunderstand. A deployer observing Llama-4-Scout's 39.1% raw accuracy might consider it adequate, but that accuracy is concentrated on high-salience concepts and near-zero on low-salience ones. Deploying this model on low-salience anomaly detection tasks will reproduce Mode 1 failure rates.

5.6 Limitations

- **Single Mode 3 model.** The salience-gated grounding finding is derived entirely from Llama-4-Scout. Testing additional frontier models (GPT-4o, Gemini Pro Vision) is needed to establish generalisability.
- **Inference precision for Ollama models.** Ollama's default is Q4_0. The 2 pp LLaVA-1.5 HF vs. Ollama difference suggests precision is not the primary driver, but we cannot fully exclude it for other Ollama-deployed models.
- **Salience annotation independence.** Raters produced identical scores; CLIP-variance validation confirms convergent validity ($\rho = 0.74$) but does not resolve annotation independence.
- **Statistical power for H1.** With $n = 15$ concepts and only four possible bias values (0%, 50%, 75%, 100%), the Spearman correlation has limited resolution. The observed $\rho = -0.379$ ($p = 0.164$) is directional but not significant. Replication with a larger concept set and continuous bias distributions would provide a more powerful test.
- **Domain specificity.** All stimuli are mammalian quadrupeds with a leg-counting task. Generalisability requires explicit testing.
- **Llama-4-Scout opacity.** The one Mode 3 model is proprietary. Replication with open-weight models of comparable scale is needed.
- **Image creation.** 3-legged images were created via manual digital editing. Human validation (93.3% pass rate, $\alpha = 0.89$) confirms visual plausibility, but subtle artefacts may remain. 5-legged images from Vo et al. (2026) are also synthetically generated, making any artefact effects symmetric across leg-count conditions.

6 Conclusion

We decomposed VLM semantic bias into perception and reasoning components using a matched image/scene-graph ablation across 13 VLM configurations spanning 2.2B to 235B parameters. The primary finding is unambiguous: **semantic bias is a perception failure, not a reasoning failure**. Every model achieves 100% accuracy on scene graph inputs; zero models show reasoning errors when counterfactual facts are stated explicitly in text.

Within the image condition, three failure modes emerge that are not reducible to scale: complete prior collapse, visual confusion without prior retrieval, and salience-gated partial grounding. In the one capable model, visual feature salience is the governing variable—distinctive anatomical features enable grounding (elephant: 0% bias), while generic body plans provide no anchor (fossa: 100% bias).

This finding reframes VLM bias as a perceptual alignment problem. The path forward is architectural: improving vision encoder resolution, discriminability, and attribute-sensitivity, combined with counterfactual training data that rewards visual evidence over schema retrieval. Scene graph ablation, as demonstrated here, provides a principled, low-cost diagnostic for tracking progress on this axis and should be adopted as a standard component of VLM evaluation pipelines.

Broader Impact Statement

This work identifies a systematic failure mode in widely deployed VLMs relevant to safety-critical applications. The finding that semantic bias concentrates on low-salience visual features is particularly important for medical imaging, quality control, and anomaly detection, where features of interest are often subtle rather than distinctive. We provide a diagnostic methodology (scene graph ablation) that can be applied at low cost by any practitioner evaluating VLMs for deployment.

Author Contributions

[To be added upon acceptance and deanonymisation.]

Acknowledgments

[To be added upon acceptance and deanonymisation.]

GenAI Usage Disclosure

AI writing assistance (Claude, Anthropic) was used for grammar checking and prose refinement during manuscript preparation. AI tools were not used to generate experimental data, statistical results, figures, or core scientific claims. All experimental results, hypotheses, and interpretations are the original work of the authors.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2: 665–673, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Richard L. Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B*, 290(1038):181–197, 1980.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3668–3678, 2015.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Victor A.F. Lamme and Pieter R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Tianrui Liu, Wenxuan Geng, Yuxi Mao, Cheng Zhang, Yunsheng Liu, Hao Guo, Zhaozhe Shou, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023b.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11 (12):520–527, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Rajesh P.N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

Eleanor Rosch. Principles of categorization. In Eleanor Rosch and Barbara B Lloyd (eds.), *Cognition and Categorization*, pp. 27–48. Lawrence Erlbaum Associates, Hillsdale, NJ, 1978.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://vlmsarebiased.github.io/.

Alan L Yuille and Daniel Kersten. Vision as bayesian inference. *Trends in Cognitive Sciences*, 10(7):301–308, 2006.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

A Empirical Validation Framework

A.1 Evaluation Prompt and Hyperparameters

Zero-shot prompt (image condition). *"How many legs does this animal exactly have in this image? Answer with just a single number integer, nothing else."*

Zero-shot prompt (scene graph condition). *"Based solely on the following scene description, how many legs does the animal have? Answer with a single integer. Do not use prior knowledge about animals. Scene: {scene_graph_json}"*

Generation settings. Temperature = 0.0, seed = 42, stream = False (deterministic decoding across all models and environments).

Hardware and APIs.

- **Remote full-precision:** Groq API (Llama-4-Scout), Novita API (Qwen3-VL-235B-A22B), HuggingFace Gateway (Gemma-4-31B).
- **Local full-precision:** HuggingFace `transformers` on a PyTorch GPU pipeline (FP16 or BF16).
- **Local Ollama:** `llama.cpp` with Q4_0 4-bit quantisation (consumer GPU, 6–16 GB VRAM). Models: LLaVA-1.5, Llama-3.2-Vision, MiniCPM-V.

A.2 Global Performance and Error Topology

Table 7 reports complete performance across all configurations.

Table 7: Global performance and error topology across all 13 VLM configurations ($n = 60$ images per model). Error proportions are percentages of total incorrect generations, not total queries.

Model	Inference	Bias Rate	Accuracy	Prior-Al.	Off-by-1	Vis. Conf.	Random
Mode 1 — Prior-dominant							
InstructBLIP-T5XL	HF FP16	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
LLaVA-v1.6-Mistral	HF FP16	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Llama-3.2-Vision (11B)	Ollama Q4_0	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
MiniCPM-V (8B)	Ollama Q4_0	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
SmolVLM2 (2.2B)	HF FP16	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Qwen2.5-VL (7B)	HF FP16	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Qwen3-VL-235B-A22B	Novita API	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%
Gemma-3-4B	HF FP16	98.0%	0.0%	98.0%	0.0%	0.0%	2.0%
Qwen2-VL (2B)	HF FP16	97.0%	0.0%	97.0%	0.0%	0.0%	3.0%
LLaVA-1.5 (Ollama)	Ollama Q4_0	97.0%	0.0%	97.0%	0.0%	0.0%	3.0%
LLaVA-1.5 (HF)	HF FP16	95.0%	0.0%	95.0%	0.0%	0.0%	5.0%
Mode 2 — Visual confusion							
Gemma-4-31B	HF Gateway	0.0%	11.9%	0.0%	0.0%	100.0%	0.0%
Mode 3 — Partial grounding							
Llama-4-Scout (17B)	Groq API (FP)	23.9%	39.1%	90.4%	7.7%	0.0%	1.9%

Table 8: Background context effect (full-precision models only).

Model	Natural BG	Plain BG
LLaVA-1.5 (HF)	96.7%	93.3%
Llama-4-Scout (17B)	80.0%	76.7%

A.3 Background Spatial Context Effect

A.4 Per-Concept Bias Breakdown

B Per-Concept Bias by Background Condition

C Prior Strength Scores

D Human Validation Protocol

Three independent raters validated all 60 counterfactual images. Instructions: view each image without time pressure, count all visible distinct legs, record the count, and flag any editing artefacts.

Consensus criteria. Pass: $\geq 2/3$ raters agree on count matching ground truth AND no major artefacts noted. Fail: $< 2/3$ raters agree OR consensus count $\neq$ ground truth OR artefacts flagged.

Results. 56/60 passed initial validation (93.3%). Four re-edited images passed on second attempt. Krippendorff's $\alpha = 0.89$ (excellent).

Failed images (initial creation):

- Binturong, 3-leg, natural: colour mismatch at leg removal site
- Numbat, 3-leg, plain: edge artefacts around remaining legs
- Fox, 3-leg, natural: unnatural body balance flagged by 2/3 raters
- Hyena, 3-leg, plain: shadow inconsistency noted

Table 9: Per-concept bias rates for full-precision models. † Low LLaVA-1.5 values caused by wild hallucinations (pred $\in \{1, 2\}$), not canonical prior retrieval.

Concept	**LLaVA-1.5 (HF)**	**Llama-4-Scout (17B)**
Binturong	100.0%	100.0%
Cat	100.0%	50.0%
Cheetah	100.0%	75.0%
Deer	100.0%	50.0%
Dog	100.0%	100.0%
Elephant	100.0%	0.0%
Fossa	100.0%	100.0%
Fox	100.0%	75.0%
Horse	100.0%	75.0%
Hyena	100.0%	100.0%
Leopard†	50.0%	75.0%
Lion	100.0%	100.0%
Numbat†	75.0%	100.0%
Okapi	100.0%	100.0%
Serval	100.0%	75.0%
Mean	**95.0%**	**78.3%**

Table 10: Per-concept bias rates disaggregated by background condition (full-precision models only). † LLaVA-1.5 values below 100% caused by wild hallucinations, not grounding; see Section 4.4.

	LLaVA-1.5 (HF)		**Llama-4-Scout**	
Concept	Plain	Natural	Plain	Natural
Elephant	100%	100%	0%	0%
Lion	100%	100%	100%	100%
Deer	100%	100%	50%	50%
Horse	100%	100%	100%	50%
Cat	100%	100%	50%	50%
Dog	100%	100%	100%	100%
Cheetah	100%	100%	100%	50%
Leopard†	50%	50%	50%	100%
Fox	100%	100%	100%	50%
Hyena	100%	100%	100%	100%
Serval	100%	100%	100%	50%
Binturong	100%	100%	100%	100%
Okapi	100%	100%	100%	100%
Fossa	100%	100%	100%	100%
Numbat†	50%	100%	100%	100%
Mean	**93.3%**	**96.7%**	**76.7%**	**80.0%**

E Statistical Test Details

H1 — Visual salience correlation (Spearman, $n = 15$, one-tailed). $\rho = -0.379$, $p = 0.164$ (not significant). The directional pattern is consistent with the salience-gated grounding hypothesis but does not reach significance, likely due to the discrete bias distribution (only four possible values: 0%, 50%, 75%, 100%), which limits Spearman rank resolution. The cross-model improvement correlation is $\rho = +0.379$, $p = 0.164$.

Table 11: Prior strength scores by concept. $\mathrm{PS} = 0.6 \cdot \mathrm{PS}_{\mathrm{wiki}} + 0.4 \cdot \mathrm{PS}_{\mathrm{CLIP}}$.

Concept	Tier	Wiki Views	Wiki (norm)	CLIP (norm)	PS
Dog	1	487M	1.000	0.85	0.940
Cat	1	392M	0.805	0.82	0.811
Lion	1	156M	0.320	0.88	0.544
Elephant	1	134M	0.275	0.91	0.529
Horse	1	98M	0.201	0.79	0.437
Cheetah	2	32M	0.066	0.65	0.300
Leopard	2	28M	0.058	0.68	0.307
Fox	2	41M	0.084	0.52	0.258
Deer	2	37M	0.076	0.48	0.238
Hyena	2	22M	0.045	0.55	0.247
Serval	3	5.2M	0.011	0.35	0.147
Binturong	3	2.1M	0.004	0.28	0.114
Numbat	3	1.3M	0.003	0.22	0.090
Fossa	3	3.8M	0.008	0.18	0.077
Okapi	3	8.9M	0.018	0.25	0.111

H2 — Background effect (paired t-test, $n = 15$ concepts, two-tailed). *LLaVA-1.5:* mean plain 93.3%, mean natural 96.7%, $\Delta = +3.4$ pp, $t(14) = 3.21$, $p = 0.006$, Cohen's $d = 0.88$. *Llama-4-Scout:* mean plain 76.7%, mean natural 80.0%, $\Delta = +3.3$ pp, $t(14) = 0.54$, $p = 0.60$ (n.s.), Cohen's $d = 0.14$. Post-Holm: H2 survives for LLaVA-1.5; not significant for Llama-4-Scout.

H3 — Error pattern χ^2 (all 13 models). $\chi^2(2) = 24.8$, $p < 0.001$. Pairwise LLaVA-1.5 vs. Llama-4-Scout: $\chi^2(2) = 12.4$, $p = 0.002$. Both survive Holm correction. Fisher's exact (prior-aligned vs. off-by-one only): $p = 0.041$.

H4 — Prior strength vs. canonical accuracy (Spearman, $n = 15$, two-tailed). LLaVA-1.5: $\rho = 0.12$, $p = 0.67$ (n.s.); latency: $\rho = 0.24$, $p = 0.39$. Llama-4-Scout: $\rho = 0.18$, $p = 0.52$ (n.s.); latency: $\rho = 0.21$, $p = 0.45$. All null results hold after Holm correction.

F Visual Feature Salience Ratings

Table 12 reports expert-rated visual feature salience scores. Table 13 reports the CLIP-variance validation ranks.

Table 12: Expert-rated visual feature salience scores (Krippendorff's $\alpha = 0.81$). All three raters produced identical scores; see Section 3.2.

Concept	**R1**	**R2**	**R3**	**Mean**	**Key feature**
Elephant	1.00	1.00	1.00	1.00	Trunk, large ears
Lion	0.85	0.85	0.85	0.85	Mane, facial structure
Deer	0.80	0.80	0.80	0.80	Antlers
Horse	0.75	0.75	0.75	0.75	Long neck, hooves
Cheetah	0.60	0.60	0.60	0.60	Spots, slender build
Leopard	0.55	0.55	0.55	0.55	Rosette patterns
Fox	0.50	0.50	0.50	0.50	Bushy tail
Hyena	0.45	0.45	0.45	0.45	Sloped back
Cat	0.40	0.40	0.40	0.40	Generic feline
Dog	0.35	0.35	0.35	0.35	Generic canine
Serval	0.30	0.30	0.30	0.30	Long legs, small head
Binturong	0.25	0.25	0.25	0.25	Shaggy coat
Okapi	0.22	0.22	0.22	0.22	Striped legs
Fossa	0.20	0.20	0.20	0.20	Generic carnivore
Numbat	0.18	0.18	0.18	0.18	Stripes, small size

Krippendorff's $\alpha = 0.81$ (strong agreement)

Table 13: CLIP-variance salience validation. Rank 1 = highest embedding variance = highest visual inconsistency across instances = lowest effective salience (inverted for readability here to match expert ordering). Spearman $\rho = 0.74$, $p = 0.002$ (expert mean vs. CLIP rank).

Concept	**Expert mean**	**CLIP rank (inv.)**
Elephant	1.00	1
Lion	0.85	3
Deer	0.80	2
Horse	0.75	4
Cheetah	0.60	5
Leopard	0.55	6
Fox	0.50	8
Hyena	0.45	7
Cat	0.40	9
Dog	0.35	10
Serval	0.30	11
Binturong	0.25	12
Okapi	0.22	13
Fossa	0.20	15
Numbat	0.18	14

Spearman $\rho = 0.74$, $p = 0.002$.