When Visual Signals Mislead: A Mechanistic Study of Attribute Hallucination in Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Attribute hallucination—where vision-language models (VLMs) correctly identify an object but mischaracterize its properties—is prevalent yet mechanistically poorly understood. The dominant explanation, language-prior dominance, has motivated prior-suppression methods, but this explanation has not been directly tested at the attribute level. We present **VISOR** (Visual-Operational Remediation), a unified framework that couples null-image-based diagnosis with routed remediation. Its VSNR diagnostic compares a real-image logit margin with a null-image reference margin. Across the three model families and three attribute types, the real-image margin tracks the false-positive decision while the null-image reference carries little information. VISOR uses this diagnosis to separate two failure modes: low-margin but directionally correct signals in color/state attributes, and low-SNR or late-stage readout-misaligned signals in material attributes. The same diagnosis routes each query to the appropriate operator: calibration for threshold-placement errors, abstention for training-free low-SNR handling, or targeted visual adaptation for material failures that prior suppression cannot correct. On closed-vocabulary Yes/No attribute probes from VAW and GQA, VISOR attains the lowest false-positive rate across all nine model–attribute settings, and on material it does so without increasing the false-negative rate.

1 Introduction

Vision-language models (VLMs) (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Dai et al., 2023; Zhu et al., 2023) have achieved strong performance on object-level benchmarks (Liu et al., 2024d), yet they remain prone to *attribute hallucination*: correctly identifying an object while mischaracterizing its properties. A model may recognize a jacket yet assert it is *leather* rather than *denim*, or describe a surface as *paper* when it is *cardboard*. Unlike object-existence errors, attribute errors can pass standard object-level checks while corrupting decisions that depend on fine-grained visual properties.

The dominant explanation for VLM hallucination is **language prior dominance**: statistical co-occurrences from pre-training allow language-layer biases to override visual evidence (Leng et al., 2024). This hypothesis has motivated contrastive decoding (VCD; Leng et al., 2024), instruction-based correction (ICD; Wang et al., 2024), and preference optimization (Zhao et al., 2023)—all targeting the same causal pathway: suppress the language prior, reduce hallucination. Yet whether this pathway drives *attribute* hallucination has never been tested at the attribute level. What is missing is an operational way to determine, for a given query, which candidate source the error should be attributed to.

We address this gap with **VISOR**, a unified diagnosis-and-remediation framework for controlled yes/no attribute probing. VISOR first decomposes each prediction into a **visual logit difference**, $\delta_{\text{vis}} = \text{logit}(\text{yes}|\text{image}) - \text{logit}(\text{no}|\text{image})$, and a **null-image reference logit difference**, $\delta_{\text{null}} = \text{logit}(\text{yes}|\text{null}) - \text{logit}(\text{no}|\text{null})$, where the null image is a uniform blank. Because that blank is still encoded by the vision tower and passed through the cross-modal pathway, we treat δ_{null} as an operational null-image reference rather than a text-only language prior. Across 3,597 negative-ground-truth items on VAW (Pham et al., 2021), evaluated with three model families, the real-image

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

margin δ_{vis} tracks false positives while δ_{null} carries little information: the null-image reference alone reaches only 58–59% accuracy. The label is defined by the sign of δ_{vis}, so this is a consistency check rather than a causal test; it nonetheless provides the operational coordinate used for diagnosis and intervention below.

The same diagnostic coordinate also determines how VISOR intervenes. When the visual signal has the correct direction but an insufficient margin, VISOR shifts the decision boundary by a calibrated prior term. When the signal is low-SNR or misaligned in the final Yes/No logit space, VISOR either abstains without training or repairs the visual projection with a targeted adapter. The framework is therefore organized around one operational question: which diagnosed failure mode does this query exhibit, and which intervention matches it?

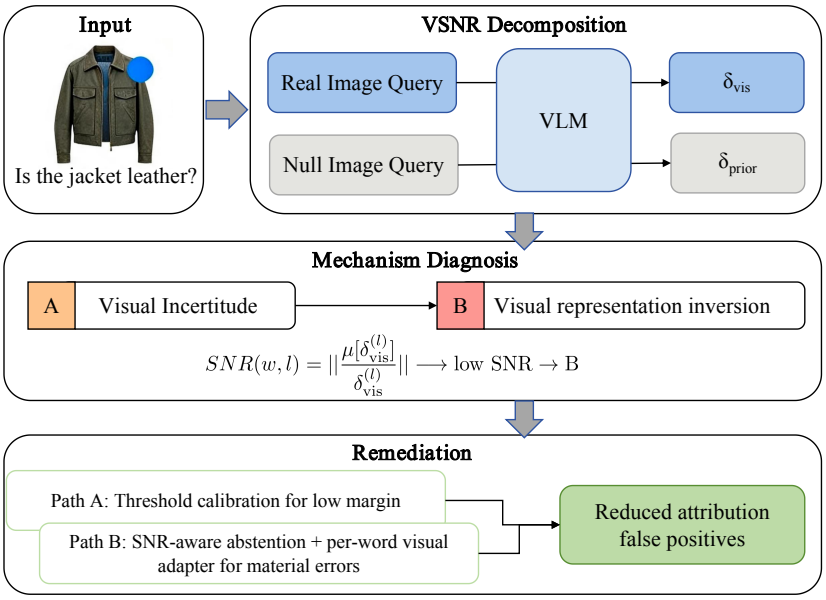

Figure 1: **Framework overview.** VISOR first uses VSNR to separate the real-image margin from a null-image reference margin, written δ_{prior} in the diagram. The same diagnostic signal supports mechanism analysis, layer-wise back-tracing, and routing to a mechanism-specific operator: Calib for low-margin visual signals (Path A), and Abstain or Adapt for material signals (Path B), where the material failure we analyse is a late-stage visual-to-logit readout misalignment.

Our contributions are:

1. **VISOR framework**: We introduce a unified framework that couples VSNR diagnosis with routed remediation, using the same visual-signal coordinate for hypothesis testing, layer-wise analysis, and operator selection.
2. **Diagnostic taxonomy**: We identify two failure modes: threshold-correctable low-margin errors in color/state, and late-stage visual-to-logit readout misalignment in material, supported by layer-wise and word-level evidence.
3. **Mechanism-specific remediation**: We evaluate VISOR as a routed system whose Calib, Abstain, and Adapt operators are selected by the diagnosed failure mode. Across nine model–attribute settings, VISOR attains the lowest false-positive rate; on material it does so without increasing the false-negative rate, whereas VCD and ICD do not reduce false positives at their default operating point.

2 Related Work

Hallucination evaluation. Object-level hallucination is commonly evaluated with POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), and related benchmarks summarized in recent surveys (Yin et al., 2023; Zhang et al., 2024; Bai et al., 2024; Liu et al., 2024b). HallusionBench (Liu et al.,

2024a), MME (Fu et al., 2023), and GQA (Hudson & Manning, 2019) include attribute-related questions, but they usually report aggregate accuracy rather than per-attribute false positives. VAW (Pham et al., 2021) provides dense attribute annotations; we repurpose it as a mechanistic yes/no diagnostic by separating false positives and false negatives for each attribute word.

Attribute recognition vs. attribute hallucination. ABO (Collins et al., 2022), OVAD (Bravo et al., 2023), and VASR (Bitton et al., 2023) evaluate attribute recognition with region proposals or task-specific models. Our setting instead asks why a full VLM asserts absent attributes during zero-shot inference, and connects per-word representation quality to false positive rates without privileged region annotations.

Internal visual-signal analysis. Recent work inspects internal visual processing rather than only output distributions. Jiang et al. (Jiang et al., 2025) use an attention lens to interpret, detect, and mitigate object hallucination in the middle layers of VLMs, and Li et al. (Li et al., 2025) track hidden visual information and steer decoding from internal activations. These works establish that layer-wise visual signals are both diagnostic and actionable. Our contribution is narrower and complementary: we formulate an attribute-level diagnostic over closed-vocabulary Yes/No queries, and we use that diagnostic to route each query to an operator.

Prior-suppression interventions. VCD (Leng et al., 2024), ICD (Wang et al., 2024), DAMRO (Gong et al., 2024), DeCo (Yao et al., 2024), and related contrastive decoding methods suppress outputs attributed to language or degraded-visual priors. Preference methods such as V-DPO (Xie et al., 2024), HA-DPO (Zhao et al., 2023), and RLHF-V (Yu et al., 2024) align responses toward visually faithful answers. These methods are well matched to failures where a correct visual signal is overridden; our analysis identifies material attribute cases where the visual signal itself is low-SNR or directionally misaligned, requiring different interventions.

3 VISOR: Diagnosis and Mechanism-Specific Remediation

VISOR couples a diagnostic module with mechanism-specific remediation (Figure 1). The diagnostic module, VSNR (Visual Signal-to-Noise Ratio), decomposes each attribute prediction into two scalar quantities extracted by paired real-image and null-image forward passes. These quantities define the decision variable, the prior estimate, and the coordinate used for layer-wise back-tracing. VISOR then routes a query to one of three operators according to the diagnosed failure mode: Calib for low-margin visual signals, Abstain for training-free low-SNR handling, and Adapt for targeted visual repair. The operators are therefore not separate methods; they are different actions taken by one diagnosis-driven framework.

Formulation. Let M be a VLM, I an image, and q_w a yes/no attribute query "Is the [object] [attribute w]?". We define the **visual logit difference**:

$$\delta_{\text{vis}}(M, I, w) = \text{logit}_M(\text{yes} \mid I, q_w) - \text{logit}_M(\text{no} \mid I, q_w) \tag{1}$$

For a negative sample (ground truth: No), hallucination occurs when $\delta_{\text{vis}} \geq 0$.

We additionally define the **null-image reference logit difference** under a blank image $I_\varnothing$ (uniform gray):

$$\delta_{\text{null}}(M, w) = \text{logit}_M(\text{yes} \mid I_\varnothing, q_w) - \text{logit}_M(\text{no} \mid I_\varnothing, q_w) \tag{2}$$

δ_{null} is the model's logit margin with a uniform-gray image input. It is a null-image reference, not an identified language-only prior: the image still passes through the visual encoder and multimodal pathway. We collect both quantities for every sample via paired forward passes. Alternative null-image checks test sensitivity to this reference choice, but do not establish that such an input removes visual processing.

Decision and layer-wise coordinates. Under deterministic yes/no decoding, the first answer token fixes the label, so $\delta_{\text{vis}} = 0$ is the decision boundary for negative samples. Empirically, $\text{sign}(\delta_{\text{vis}})$ matches the hallucination label across all 3,597 negative-ground-truth items and all three models. Because the same final norm and LM head can be applied to intermediate hidden states, we also define:

$$\begin{aligned} \delta_{\text{vis}}^{(l)} &= z_l(\text{yes}) - z_l(\text{no}), \\ z_l(t) &= \text{LMHead}(\text{Norm}(h_l^{[\text{last}]}))\big|_t. \end{aligned} \tag{3}$$

Applying the final normalization and language-model head to intermediate states gives a readout diagnostic rather than a trained layer-specific probe or causal intervention. We compute the descriptive per-layer ratio, $\mathrm{SNR}(w, l) = |\mu[\delta_{\text{vis}}^{(l)}]|/\sigma[\delta_{\text{vis}}^{(l)}]$, allowing us to trace when the decision-relevant visual signal becomes low-margin or misaligned.

Operational definitions. We use three terms operationally. At the word level, $\mathrm{SNR}(w) = |\mu[\delta_{\text{vis}}(w)]|/\sigma[\delta_{\text{vis}}(w)]$ is computed over the negative-ground-truth samples of word w, and this scalar is the VSNR routing statistic. A word is *directionally correct* when $\mu[\delta_{\text{vis}}(w)] < 0$, so its errors come from margin magnitude or variance rather than a wrong sign. A word is *low-margin* when it is directionally correct yet the zero threshold still falls inside the bulk of its δ_{vis} distribution, which keeps its FPR high. For material words, ***B1*** denotes a word whose FPR is mostly explained by annotation granularity, ***B2*** a word whose FPR is unchanged after those negatives are removed, and *mixed* the intermediate case; in Table 3 the split is unambiguous (steel $45.5\% \rightarrow 4.5\%$ versus rubber $44.4\% \rightarrow 44.4\%$), with ceramic in between ($32.5\% \rightarrow 27.5\%$).

Operational real/null margin comparison. We define the operational image-conditioned margin change:

$$\Delta_{\text{img}-\text{null}} = \delta_{\text{vis}} - \delta_{\text{null}} \tag{4}$$

This contrast measures how the logit margin changes when real content replaces a blank image. Because a blank image still passes through the vision tower, we read $\Delta_{\text{img}-\text{null}}$ as an operational margin change rather than as a likelihood term.

For an output-level contrastive correction $\delta'_{\text{vis}} = \delta_{\text{vis}} - \alpha\hat{\delta}_{\text{prior}}$, correcting a hallucinating sample with $\delta_{\text{vis}} > 0$ requires $\hat{\delta}_{\text{prior}} > \delta_{\text{vis}}/\alpha > 0$. The tested B2 material words instead have $\hat{\delta}_{\text{prior}} \leq 0$ across models; therefore output-level prior subtraction cannot flip the decision sign. The formal statement and proof are given in the supplementary material.

Unified routing. VISOR uses the same diagnostic coordinate for both analysis and intervention. For words whose visual signal is directionally correct but low-margin, VISOR applies a calibrated decision rule $\mathbf{1}[\delta_{\text{vis}} - \gamma\delta_{\text{null}} > 0]$. For low-SNR material words, the routed operator depends on deployment constraints: VISOR-Abstain returns uncertainty without training, whereas VISOR-Adapt loads a per-word visual LoRA adapter when definitive answers are required. Routing is per attribute word rather than per query, so every query for word w receives the same operator and VISOR-Adapt loads the adapter keyed by w. The offline pass uses only negative-ground-truth samples to estimate $\mu[\delta_{\text{vis}}(w)]$, $\sigma[\delta_{\text{vis}}(w)]$, and $\delta_{\text{null}}(w)$, and it is on this pass that γ for Calib and τ for Abstain are fixed. Inference therefore needs no ground-truth labels and no per-query statistic beyond the two logit margins. Calib, Abstain, and Adapt are not stitched together post hoc, but are the three actions exposed by the same VSNR diagnosis.

4 EXPERIMENTS

Experimental setup. We use the VAW (Visual Attributes in the Wild) dataset (Pham et al., 2021), which provides per-instance attribute annotations from Visual Genome across 620 attribute words. For each attribute type we construct a yes/no probing benchmark by issuing queries of the form *"Is the [object] [attribute]? Answer yes or no."* Negative samples are drawn from instances where the attribute appears in `negative_attributes`; positive samples from `positive_attributes`. We measure **False Positive Rate (FPR)**—the fraction of negative-ground-truth samples where the model answers "yes"—as the primary hallucination metric. Dataset statistics (Qwen2.5-VL-3B): color 2,992 samples (1,302 pos / 1,690 neg, 61 words); material 2,417 samples (969 pos / 1,448 neg, 45 words); state 803 samples (344 pos / 459 neg, 11 words). We evaluate Qwen2.5-VL-3B-Instruct (Bai et al., 2025), InternVL3.5-4B-Flash (Chen et al., 2024), and LLaVA-1.5-7B (Liu et al., 2024c). These three families span different visual encoder designs (Qwen-VL ViT, InternViT-300M, CLIP ViT-L/14 (Radford et al., 2021; Dosovitskiy et al., 2021)) and training recipes, providing architectural diversity for cross-model validation.

Logit extraction. Each (image, query) pair receives two forward passes: one with the real image to obtain δ_{vis}, one with a blank gray image ($128 \times 128 \times 3$, all channels $= 128$) to obtain δ_{null}. We

record logit(yes) and logit(no) as the maximum log-probability among surface forms {"Yes", "yes", " Yes", " yes"} and {"No", "no", " No", " no"} respectively from the final token position.

Metrics. Our yes/no probing benchmark follows the POPE (Li et al., 2023b) format: closed-vocabulary positive/negative queries. The splits retain the naturally occurring positive/negative counts reported above rather than enforcing equal class sizes. FPR is the primary metric because it isolates the hallucination direction. Table 1 reports the two directly measured quantities, FPR and Yes-Ratio, with the FNR they imply.

Table 1: **Method comparison on the VAW probing benchmark (Yes-Ratio / FPR / FNR, %).** The two directly measured quantities are FPR and Yes-Ratio; FNR follows from them under the stated class counts. Best FPR per column is in **bold**. VISOR denotes the routed framework: Calib is selected for color/state and Adapt for material, and VCD and ICD do not improve FPR in any column. The weighted full-material estimate reported later uses a separate protocol.

(a) Color attribute

	Qwen2.5-VL-3B			InternVL3.5-4B			LLaVA-1.5-7B		
Method	YR	FPR	FNR	YR	FPR	FNR	YR	FPR	FNR
Baseline	55.6	33.8	16.1	54.4	31.6	16.0	62.6	47.2	17.4
VCD	59.4	38.0	12.8	57.2	35.6	14.8	64.8	51.3	17.7
ICD	58.2	36.5	13.6	57.6	35.6	13.8	66.1	51.3	14.7
VISOR (ours)	43.7	**22.2**	28.4	37.8	**15.4**	33.1	46.1	**29.8**	32.7

(b) Material attribute

	Qwen2.5-VL-3B			InternVL3.5-4B			LLaVA-1.5-7B		
Method	YR	FPR	FNR	YR	FPR	FNR	YR	FPR	FNR
Baseline	37.9	15.7	28.9	42.3	19.5	23.6	45.7	25.8	24.6
VCD	43.4	15.7	15.2	44.2	19.5	18.9	48.5	25.8	17.6
ICD	46.0	15.7	8.7	47.7	19.5	10.2	42.9	25.8	31.5
VISOR (ours)	37.0	**14.2**	28.9	41.7	**17.8**	22.6	45.7	**24.5**	22.6

(c) State attribute

	Qwen2.5-VL-3B			InternVL3.5-4B			LLaVA-1.5-7B		
Method	YR	FPR	FNR	YR	FPR	FNR	YR	FPR	FNR
Baseline	46.6	25.9	25.8	43.5	25.7	32.7	50.1	32.5	26.4
VCD	46.0	25.7	26.9	42.7	28.1	37.8	47.5	37.0	38.5
ICD	46.0	25.7	26.9	42.7	28.1	37.8	47.5	37.0	38.5
VISOR (ours)	40.0	**19.0**	32.0	37.7	**19.6**	38.1	50.1	32.5	26.4

VISOR attains the lowest FPR in all nine settings. On material, the reduction is not bought with missed positives: FNR stays at 28.9% on Qwen and falls to 22.6% on both InternVL and LLaVA, whereas VCD and ICD leave material FPR unchanged at their default operating point. On color, Calib lowers FPR by 11–17 pp and raises FNR, the signature of moving a decision boundary.

Reproducibility. All inference experiments use greedy decoding with temperature$= 0$ and fixed random seed 42. LoRA fine-tuning uses seed 42 across all three model families. Experiments were conducted on a single NVIDIA A100 80GB GPU.

The null-image reference does not explain the errors. We assess whether the null-image reference (δ_{null}) or the real-image margin (δ_{vis}) explains attribute hallucination. We present a primary correlational analysis followed by an additional image-sensitivity analysis.

Correlational analysis (primary evidence). For the 3,597 negative-ground-truth items, we compute Spearman rank correlation between each logit quantity and the binary hallucination outcome (1 = model predicted "yes"):

The label is defined by the sign of δ_{vis}, so Table 2 is a consistency check rather than a causal test; the null-image reference alone reaches only 58–59% accuracy. Language bias is documented in general (McKenna et al., 2023), including in image-free settings (Asadi et al., 2025), but for these probes the decision is governed by the real-image margin, and δ_{null} is a reference point rather than the cause of the error.

Table 2: Decision-margin consistency. The negative-sample label is defined by the sign of δ_{vis}, so this is a consistency check on the readout; the null-image margin carries far less information.

Model	$\rho(\delta_{\text{vis}}, \text{halluc.})$	$\rho(\delta_{\text{null}}, \text{halluc.})$	Prior-only acc.
Qwen2.5-VL-3B	**0.755**	0.177	58.5%
InternVL3.5-4B	**0.760**	0.066	58.6%
LLaVA-1.5-7B	**0.835**	0.066	59.0%

Image-sensitivity analysis (additional evidence). We next test whether the real image can override a positive null-image prior. On the color negatives, we identify 278 *prior-positive conflict* cases where $\delta_{\text{null}} > 0$ but $\delta_{\text{vis}} < 0$:

Word-level prior diagnostics.

Null-image reference predicts color FPR but not material FPR. At the word level, δ_{null} is informative for color ($\rho = +0.420$, $p = 0.003$) but not for material ($\rho(\delta_{\text{null}}, \text{FPR}) = +0.034$, $p = 0.83$). This asymmetry is why Calib helps color/state but not material, where word-level FPR variance tracks the δ_{vis} distribution instead ($\rho = +0.752$, $p < 0.0001$).

5 Two Mechanisms of Attribute Hallucination

Mechanism A: Visual Incertitude. For **color** and **state** attributes, δ_{vis} points in the correct direction on average but with insufficient margin, which places the error at the decision boundary rather than in the sign of the visual signal. Compound-color diagnostics are provided in the supplementary material.

Mechanism B: Late-stage visual-to-logit misalignment. For **material** attributes, high FPR persists even for words with a near-neutral null-image reference. We separate annotation granularity noise from residual errors that annotation granularity cannot explain. Steel is mostly B1 label granularity: 18/20 hallucinating negatives are visually metallic, so its observed FPR drops from 45.5% to 4.5% after removing annotation-noise cases. Rubber, paper, and leather are B2 failures: observed and adjusted FPR are identical after the same filtering.

Table 3: Observed vs. adjusted FPR after annotation-noise removal. B1 words fall and B2 words remain unchanged. The residual B2 FPR is not explained by annotation granularity, but this does not by itself identify whether the cause is visual encoding or late readout.

Word	Observed FPR	True FPR	Type
steel	45.5%	4.5%	B1
rubber	44.4%	44.4%	B2
paper	41.9%	41.9%	B2
ceramic	32.5%	27.5%	mixed
leather	26.0%	26.0%	B2

Descriptive SNR–FPR association. We define $\text{SNR}(w) = |\bar{\delta}_{\text{vis}}(w)|/\sigma_{\delta_{\text{vis}}(w)}$. Across the 45 material words, this descriptive statistic is associated with FPR (Table 4). SNR and FPR share the same margin distribution, so this is a descriptive association between two views of that distribution.

Late-stage SNR collapse. Layer-wise probing applies the frozen LM head to intermediate hidden states, so it measures alignment with that Yes/No readout rather than the presence of material information in the hidden state itself. With that caveat, the probe localizes Mechanism B to the final decoder layers rather than to a complete absence of mid-layer visual evidence. For B2 words, SNR drops from about 35 near L17 to below 0.9 by L36, whereas high-SNR words maintain SNR around 2–4. This narrows the failure to late-stage visual-to-logit projection dynamics (Figure 2); cross-model trajectories, signed-signal plots, and case studies are in the supplementary material.

Table 4: Predictors of material FPR across the 45 material words in the benchmark. SNR achieves $\rho = -0.916$, outperforming either component alone.

Predictor	Spearman ρ with FPR	p-value
$\bar{\delta}_{\text{vis}}(w)$	-0.765	< 0.0001
$\sigma_{\delta_{\text{vis}}(w)}$	$+0.633$	< 0.0001
$\text{SNR}(w)$	-0.916	< 0.0001

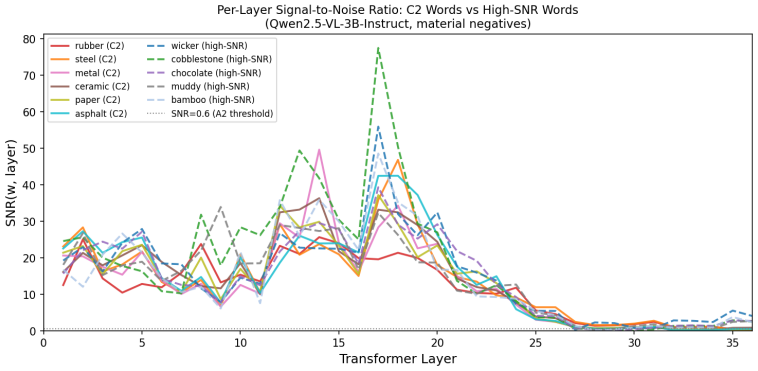

Figure 2: Per-layer SNR for low-SNR material words vs. high-SNR words across 36 decoder layers of Qwen2.5-VL-3B. The divergence occurs in L28–L36, localizing the degradation to late decoder layers.

6 VISOR Results and Mechanistic Validation

We next verify that VISOR's routed operators match the diagnosed mechanisms. Calib shifts the decision boundary for Mechanism A (Table 5), and gives 0 pp on Qwen and InternVL for material attributes, consistent with Mechanism B not being threshold correctable.

Table 5: VISOR Calib vs. VCD/ICD on color, state, and material across three models. Calib helps color/state where δ_{vis} is directionally correct but low-margin, and gives 0 pp on material for Qwen and InternVL; on LLaVA material it lowers FPR but raises FNR by 7.1 pp. This is the Calib-only branch and the reduction is an operating-point change; Table 1 reports the routed policy and its FNR cost.

Model	Attribute	Baseline	VISOR-Calib	VCD	ICD
Qwen2.5-VL-3B	color	33.8%	**22.2%** (-11.6 pp)	38.0% ($+4.2$ pp)	36.5% ($+2.7$ pp)
InternVL3.5-4B	color	31.6%	**15.4%** (-16.2 pp)	35.6% ($+4.0$ pp)	35.6% ($+4.0$ pp)
LLaVA-1.5-7B	color	47.2%	**29.8%** (-17.4 pp)	51.3% ($+4.1$ pp)	51.3% ($+4.1$ pp)
Qwen2.5-VL-3B	state	25.9%	**19.0%** (-6.9 pp)	25.7% (-0.2 pp)	25.7% (-0.2 pp)
InternVL3.5-4B	state	25.7%	**19.6%** (-6.1 pp)	28.1% ($+2.4$ pp)	28.1% ($+2.4$ pp)
LLaVA-1.5-7B	state	32.5%	32.5% (**0.0 pp**)	37.0% ($+4.6$ pp)	37.0% ($+4.6$ pp)
Qwen2.5-VL-3B	material	15.7%	15.7% (**0.0 pp**)	15.7% (0 pp)	15.7% (0 pp)
InternVL3.5-4B	material	19.5%	19.5% (**0.0 pp**)	19.5% (0 pp)	19.5% (0 pp)
LLaVA-1.5-7B	material	25.8%	**18.6%** (-7.2 pp)	25.8% (0 pp)	25.8% (0 pp)

For Mechanism B, output-level prior subtraction and language-layer tuning are insufficient because the discriminative material signal is weak or poorly projected. The training-free Abstain operator reduces Qwen material FPR from 15.7% to 10.2% at $\tau = 0.6$ with 14.2% abstention. The Adapt operator instead trains one visual LoRA adapter per high-FPR material word, avoiding the cross-contamination of a shared hard-negative LoRA. Across the six targets (Table 6) the mean FPR falls from 38.1% to 23.9% on Qwen, 44.7% to 32.8% on InternVL, and 34.2% to 26.8% on LLaVA, with MME regression within ± 1.25 pp; on Qwen the shared adapter worsens leather by 8 pp while Adapt lowers it by 12 pp. Steel's observed baseline is inflated by B1 annotation granularity (Table 3), so

the mechanism claim rests on the B2 words rubber, paper, and leather. Full probe results, per-word tables, and statistical tests are in the supplementary material.

Table 6: Per-word material FPR for Qwen2.5-VL-3B under baseline, shared hard-negative LoRA, and VISOR-Adapt per-word LoRA. VISOR-Adapt improves all six targets; steel is a B1 word (Table 3).

Word	Baseline	HN (shared)	VISOR-Adapt	95% CI
steel	45.5%	27.3%	**22.7%**	[10.3%, 35.1%]
rubber	44.4%	33.3%	**27.8%**	[14.2%, 41.4%]
paper	41.9%	38.7%	**19.4%**	[9.5%, 29.3%]
metal	38.0%	26.0%	32.0%	[19.2%, 44.8%]
leather	26.0%	34.0% ↑	**14.0%** ↓	[5.1%, 22.9%]
ceramic	32.5%	27.5%	**27.5%**	[15.8%, 39.2%]
Mean (6)	**38.1%**	31.1%	**23.9%**	—

The late-stage readout account predicts that VISOR-Adapt should raise final-layer material SNR rather than merely suppressing "yes" answers. Re-running the layer-wise probe after VISOR-Adapt training is consistent with this prediction: final-layer SNR increases for rubber (+0.307), steel (+0.407), metal (+0.177), ceramic (+0.246), and paper (+0.256), with the gains concentrated in L28–L36. SNR is computed from the same margin distribution that determines FPR, so this check is consistent with a readout-level change but does not by itself rule out a threshold-like shift.

Scope. VISOR targets structured attribute queries over a known closed vocabulary, and VISOR-Adapt covers the six highest-FPR material words (16.1% of material test samples). Free-form and open-vocabulary coverage is not evaluated here.

7 Discussion

The main implication is not that language statistics are irrelevant, but that attribute false positives should not be treated as a single prior-dominance failure. Methods that look similar at the output layer behave differently depending on the diagnosed failure mode: Calib helps when the visual signal has the right sign but an overly permissive threshold, whereas prior subtraction and language-layer tuning cannot help once the final visual logit direction is wrong.

The supplementary material summarises the negative controls: language-layer DPO leaves the visual-encoder gradient at zero, shared visual LoRA collapses to a blanket "No" policy, and hidden-state edits are unstable. The only partial recovery comes from a supervised L28 probe, hence VISOR-Adapt targets visual projections per word rather than applying a shared output-level correction.

The design transfers across architectures: material FNR is unchanged on Qwen (28.9%) and lower on InternVL (23.6%→22.6%) and LLaVA (24.6%→22.6%), so the FPR reduction does not come from a blanket shift toward answering no. Loading each Qwen adapter individually leaves MME accuracy within ± 1.25 pp of the 90.8% baseline ($t(5) = -0.53$, $p = 0.62$), the null-image reference correlates at $r > 0.90$ with black, white, and random-noise baselines in all three models, and the $\delta_{\text{vis}} \geq 0$ criterion yields zero violations on the GQA color subset.

Because the six adapted words cover 16.1% of the material negative samples, the targeted repair shifts the weighted full-material estimate for Qwen2.5-VL-3B from 15.7% to 13.5%.

Limitations

Our claims concern attribute hallucination as measured by verification queries and are diagnostic rather than causal: we characterise how a model's decision margin behaves under real and blank images and how it responds to intervention, but we do not identify the internal origin of the errors. The evidence is bounded by its instrumentation. The queries target a closed vocabulary of attributes over images from a single annotation family and name only the object, so the study does not separate reference ambiguity from attribute confusion, does not cover object-existence errors or free-form

generation, and does not by itself establish transfer to other domains or model families. The remedies we study inherit these bounds: adjusting the decision rule trades recall for fewer false positives, and adapting model parameters depends on per-target supervision over a small vocabulary, so general open-vocabulary mitigation remains unsolved. Finally, the reported rates come from one evaluation protocol with fixed hyperparameters and are best read as operating points that compare methods under identical conditions, rather than as absolute estimates of model behaviour.

8 CONCLUSION

We present VISOR, a unified framework that couples null-image-based VSNR diagnosis with routed remediation for attribute hallucination. Across controlled yes/no probing, false positives track the visual logit margin more closely than the null-image reference, motivating an operational account with two failure modes: low-margin color/state errors and low-SNR or late-stage readout-misaligned material errors. Calib then shifts the decision boundary when the visual direction is reliable, while Abstain and Adapt handle material cases where output-level prior suppression is insufficient. Within the validated scope of closed-vocabulary Yes/No probing on VAW and GQA, these results suggest that mitigation should start from a diagnosis of the decision margin rather than uniformly suppressing a presumed language prior.

AI USE STATEMENT

We used a generative AI assistant for language polishing and LaTeX formatting.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23716–23736, 2022.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2501.09905*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. VASR: Visual analogies of situation and relation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2745–2753, 2023.

María A. Bravo, Sudhanshu Mittal, Thomas Brox, and Björn Ommer. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7041–7051, 2023.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Ji, Jiangning Song, Muyi Sun, Xiying Tan, Lewei Yao, Haodong Duan, Wenbo Li, Yu Qiao, Jifeng Dai, and Conghui He. InternVL2.5: An open-source multimodal LLM for competitive performance with GPT-4o. *arXiv preprint arXiv:2412.05271*, 2024.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21094–21104, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 54846–54867, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. DAMRO: Dive into the attention mechanism of LVLM to reduce object hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Zhongqi Jiang, Junjie Chen, Bing Zhu, Tianze Luo, Yuxuan Shen, and Xin Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b.

Zhiyuan Li, Hangyu Shi, Yu Gao, Ding Liu, Zhen Wang, Yufan Chen, Tao Liu, Li Zhao, Haoran Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819, 2025.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024d.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2758–2774, 2023.

Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11741–11750, 2021.

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 (ICML)*, pp. 8748–8763, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4534–4545, 2018.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.