R^2-VLM: DETECTING AND CORRECTING VISION-LANGUAGE HALLUCINATION VIA ACTIVE REPRESENTATIONAL PERTURBATION AND RESIDUAL SPECTRAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) often confidently generate text that contradicts visual input, a failure known as *perception hallucination*. While existing mitigation methods passively analyze unperturbed forward passes, we propose **R^2-VLM**: a training-free, single-pass detect-and-correct framework that evaluates structural stability under *active* representational perturbation. Our detection module subjects a frozen VLM's hidden states to structured latent perturbations across visual, linguistic, and cross-modal subspaces. This yields a calibrated hallucination risk score and reveals that joint cross-modal reliability, rather than unimodal signals, is the primary indicator of hallucination. To correct these failures, our **Residual Spectral Grounding (RSG)** module restores lost visual evidence by additively injecting a per-sample spectral decomposition of the visual representation into the decoder's hidden states. We provide theoretical guarantees that the risk score is bounded in $[0, 1]$, adaptive thresholds converge, and RSG strictly improves representational energy stability . Empirically, across Qwen2.5-VL (3B/7B) and LLaVA-1.5 (7B) on ColorBench, POPE, and CableDetection, our risk score strongly correlates with ground-truth correctness (point-biserial r up to 0.54, Hedges' $|g|$ up to 1.50, $p < 0.001$). Furthermore, RSG nearly doubles cross-modal representational alignment ($0.35 \rightarrow 0.66$) and achieves a relative average accuracy improvement of $+2.4\%$ on colour-attribute tasks and $+1.84\%$ on CableDetection, successfully avoiding the performance degradation seen in competing subtractive baselines.

1 INTRODUCTION

Large Vision-Language Models (VLMs) such as LLaVA (Liu et al., 2023) and Qwen-VL (Bai et al., 2023; Team, 2025; Wang et al., 2024) demonstrate strong zero-shot generalisation, yet exhibit a persistent failure mode, *perception hallucination*: fluent, confident text unsupported by the visual evidence (Liang et al., 2025). This is a critical barrier to deployment in safety-sensitive domains (medical imaging, autonomous driving, industrial inspection).

The core problem. Hallucination arises from an asymmetry in visual versus linguistic grounding: VLMs are trained on far more text than image-text pairs, so the language backbone develops statistical priors that can override genuine visual evidence (Sarkar et al., 2025; Xu et al., 2026). A model that hallucinates does not *see* incorrectly—it *ignores* what it sees.

Limitations of existing approaches. Mitigation strategies fall into five categories: (i) *training-time* augmentation with negative examples (Li et al., 2024); (ii) *decoding-time* methods (VCD (Leng et al., 2024), OPERA (Huang et al., 2024)); (iii) *post-hoc* verification against external knowledge bases (Yin et al., 2024); (iv) *domain restriction* of a foundation model's VQA scope (Dutta et al., 2025); and, most relevant here, (v) *internal-representation probing*, which reads a hallucination signal from hidden states via a learned graph over the model's computational trace (Samuel et al., 2026), a Variational-Information-Bottleneck probe (Zhang et al., 2026). These confirm hallucination *is* legible in internal geometry—but every one is *passive*: each observes a single, unperturbed forward pass and

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.

trains a classifier on it or computes a static statistic over it. None asks whether the representation is *structurally stable*, nor localizes *where* or *when* grounding gives way to spurious correlation.

Our approach. We treat hallucination as *representational fragility*, diagnosed by *actively* stress-testing the representation rather than passively reading it. A genuinely grounded model should be *resilient* to targeted latent perturbations; a hallucinating model, relying on spurious correlations, will be *fragile* (Wani et al., 2026)—collapsing under the same perturbations. Because our perturbations act independently on visual, linguistic, and joint subspaces across a range of severities, the resulting stability profile *localizes* failure to a specific modality and stress threshold, turning detection into a diagnostic. This motivates R^2-VLM, which contributes:

(1) An active perturbation framework: thirteen operators stress-testing visual, linguistic, and joint cross-modal subspaces of the final-layer hidden state, in contrast to the passive probing used by all prior internal-representation detectors.

(2) Where-and-when collapse localization: a multi-factor stability score $S_i \in [0, 1]$, aggregated per modality and per severity, reveals *which* subspace collapses first and *at what magnitude*—showing joint cross-modal reliability, not either unimodal signal, dominates hallucination in three of four settings (Section 6.1).

(3) Adaptive dynamic flagging that recalibrates risk thresholds via a sliding window, stratifying predictions into LOW/MEDIUM/HIGH tiers under distribution shift.

(4) Residual Spectral Grounding (RSG): a training-free correction that *adds*, rather than suppresses, a per-sample visual decomposition into decoder hidden states—the additive counterpart to concurrent subtractive corrections, which we show degrade fine-grained attribute grounding by nearly half (Section 6.2).

2 Related Work

Characterizing hallucination. CHAIR (Rohrbach et al., 2018) first formalized object-level hallucination; POPE (Li et al., 2023) reframed it as polling-based binary classification; MMVP (Tong et al., 2024) traced a root cause to *CLIP-blind pairs* in the vision encoder. These define *where* and *why* hallucination arises at the interface, but diagnose it from outputs or static embeddings, not decoder-internal geometry.

Representation- and uncertainty-based detection. SelfCheckGPT (Manakul et al., 2023) and Semantic Entropy (Farquhar et al., 2024) detect hallucination via inconsistency across sampled generations—effective but costly at inference. Single-pass alternatives are *passive*: **UniProbe** (Samuel et al., 2026) learns a graph (GNN+ViT+GRU) over the computational trace; **HALP** (Kogilathota et al., 2026) trains linear probes; INSIDE (Chen et al., 2024) computes an EigenScore over sampled hidden-state covariance; Representation Engineering (Zou et al., 2023) reads out population-level activation directions. Every method here requires multi-sample generation, supervised training, or a static statistic over one forward pass—none actively tests structural stability under stress, which R^2-VLM does without training or sampling overhead.

Perturbation-based robustness. Visual and Textual Intervention (VTI) (Liu et al., 2024) pre-computes perturbation directions to *steer* representations at test time; we instead use perturbation to *measure* stability directly, without a steering step. This follows the adversarial-robustness literature, where reliability is read off perturbed internal features (Feinman et al., 2017; Ma et al., 2018). Unlike these single-signal diagnostics, we fuse three complementary stability signals across four severities per operator, enabling *where*-and-*when* localization rather than a single anomaly score.

Frequency-domain correction direction. Vision encoders disproportionately exploit high-frequency content largely imperceptible to humans (Wang et al., 2020). Concurrent work corrects this *subtractively*: MFP (Li et al., 2025) suppresses redundant frequency features during training; MFCD (Liu et al., 2025) contrasts against frequency-ablated images; SPIN (Sarkar et al., 2025) suppresses low-image-focus attention heads. Our **Residual Spectral Grounding** takes the opposite, *additive* stance, analogous to residual/skip connections (He et al., 2015): it adds a corrective signal

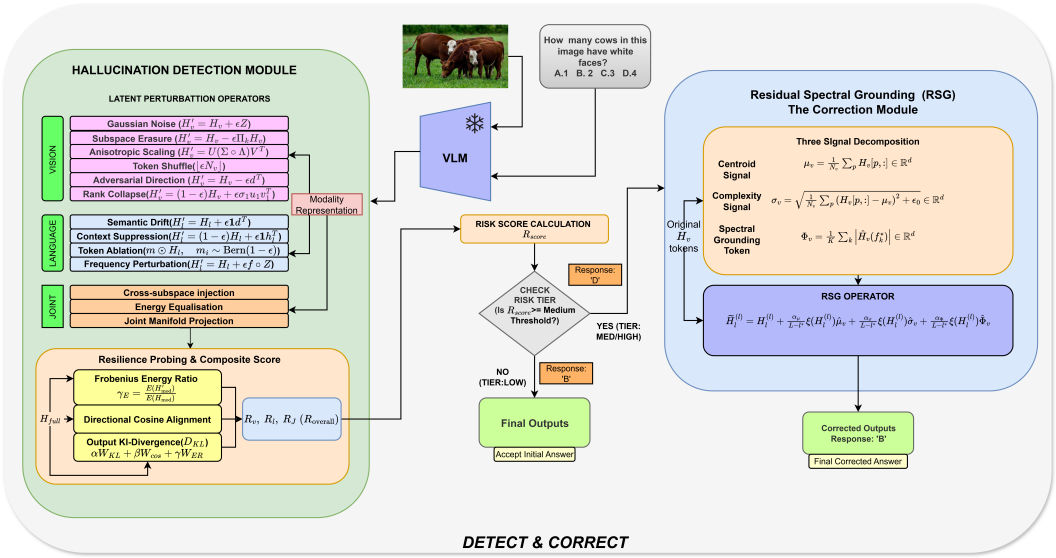


Figure 1: Overview of the **R^2-VLM** detect-and-correct framework. The Hallucination Detection Module actively stress-tests the representation to generate a composite risk score. High-risk predictions are subsequently routed to the Residual Spectral Grounding (RSG) module, which additively injects visual evidence into the decoder's hidden states to restore grounding.

atop the existing decoder state rather than deleting a computational path. As Section 6.2 shows, this is precisely why RSG *improves* fine-grained grounding where SPIN *degrades* it by nearly half.

Positioning. No prior work combines *active*, perturbation-based localization of *where* and *when* representational collapse occurs with an *additive*, residual-style correction—the space R^2-VLM occupies.

3 Background and Problem Formulation

A VLM $\mathcal{M}$ processes an image-instruction pair (I, T) through a visual encoder, modality projection, and L-layer autoregressive decoder, producing a final hidden-state matrix $H_{\text{full}} \in \mathbb{R}^{N\times d}$, $N{=}N_v{+}N_l$.

Definition 3.1 (Representational Resilience)**.** *Let $\mathcal{P} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$ be a perturbation and P, P' the output distributions before/after applying $\mathcal{P}$ to H_{full}. $\mathcal{M}$ is $(\mathcal{P}, \delta)$*-resilient *if $D_{\text{KL}}(P\|P') \le \delta$ and $\|\mathcal{P}(H) - H\|_F \,/\, \|H\|_F \le \eta$.*

Assumption 3.1 (Fragility-Hallucination Correspondence)**.** *A prediction is hallucinated iff $\mathcal{M}$ is not $(\mathcal{P}, \delta)$-resilient for a sufficiently rich family $\{\mathcal{P}_m\}_{m=1}^{M}$ targeting the visual, linguistic, and joint cross-modal subspaces of H_{full}.*

Testing this assumption requires *actively intervening* on H rather than passively observing it—the distinguishing feature of our approach relative to prior internal-geometry detectors (validated empirically in Section 6).

4 Detection Methodology: R^2-VLM

4.1 Setup

H_{full} partitions into visual/textual token subsets $H_v{=}H_{\text{full}}[\mathcal{I}_v] \in \mathbb{R}^{N_v\times d}$, $H_l{=}H_{\text{full}}[\mathcal{I}_l] \in \mathbb{R}^{N_l\times d}$ (for Qwen2.5-VL, via `VISION_START/END` delimiters). The language head projects the last-token state to logits z, and option probabilities follow temperature-scaled softmax $P(\theta) = \exp(z_\theta/\tau)/\sum_{j\in\Theta}\exp(z_j/\tau)$.

4.2 Structured Latent Perturbation Operators

Thirteen operators, in three modality groups, are each applied at four severities $\mathcal{E} = \{0.1, 0.3, 0.5, 1.0\}$ (52 attacks/sample). Further severities are subjected to computational availability and latency. Table 1 summarizes them; full equations are in Appendix B.

Table 1: Perturbation operator summary (full equations in Appendix B).

Group	Operators	Property tested
Vision (V1–V6)	Noise, subspace erasure, anisotropic scaling, token shuffle, adversarial direction, rank collapse	Spectral/positional robustness, vision-to-answer pathway
Language (L1–L4)	Semantic drift, context suppression, token ablation, frequency perturbation	Contextual/token-level robustness
Joint (J1–J3)	Subspace swap, energy equalisation, joint manifold projection	Cross-modal fusion robustness

As an example, vision *subspace erasure* (V2) projects $H_v{=}U\Sigma V^\top$ onto the complement of its top-k singular directions, $\mathcal{P}_{V2}(H_v; k, \varepsilon) = H_v - \varepsilon\, U_k(U_k^\top H_v)$, forcing reliance on residual, lower-energy visual features.

4.3 Stability Score and Where-When Localization

Each attack yields H', scored by three signals: Frobenius energy ratio γ_E (energy preservation), directional centroid alignment $S_{\cos}$ (mean-direction preservation), and output divergence D_{KL} (decision-level shift). These fuse into

$$S_i = \alpha \exp(-D_{\mathrm{KL}}) + \beta \max(0, S_{\cos}) + \gamma \exp(-|\ln \gamma_E|), \quad \alpha{=}0.40, \beta{=}0.30, \gamma{=}0.30, \tag{1}$$

fixed prior to all experiments so S_i is dataset-agnostic.

Proposition 4.1 (Score Boundedness). *$S_i \in [0, 1]$ for all attack results. (Proof in Appendix A.)*

Per-modality reliability $R_m = \frac{1}{|\mathcal{A}_m|} \sum_{i \in \mathcal{A}_m} S_i$ for $m \in \{v, l, J\}$, and $R_{\text{overall}} = (R_v{+}R_l{+}R_J)/3$, support *localization*: comparing R_v, R_l, R_J reveals *where* a representation is fragile (overwhelmingly the joint subspace, not either modality alone; Section 6.1), while sweeping $\varepsilon \in \mathcal{E}$ reveals *when* grounding gives way. This joint where/when profile is unavailable to any passive, single-forward-pass detector.

4.4 Risk Score and Adaptive Flagging

The confidence margin $\Delta P = \hat{p}_{(1)} - \hat{p}_{(2)}$ (from the raw softmax, label-free) combines with joint and overall reliability into the hallucination risk $\mathcal{R} = w_J \mathcal{F}_J + w_u(1 - \Delta P) + w_f \mathcal{F}_{\text{overall}}$, $w_J{=}0.45, w_u{=}0.35, w_f{=}0.20$, where $\mathcal{F}_J, \mathcal{F}_{\text{overall}}$ are clipped linear transforms of R_J, R_{overall} (bounds in Appendix F). To avoid alert fatigue under distribution shift, we maintain a sliding window $\mathcal{W} = [\mathcal{R}_{t-k+1}, \ldots, \mathcal{R}_t]$ and stratify $\mathcal{R}_t$ into LOW/MEDIUM/HIGH via percentile thresholds $\tau_{\text{low}}, \tau_{\text{high}}$ recomputed every step ($k{=}50$, $P_{\text{low}}{=}33$, $P_{\text{high}}{=}66$): $L_t{=}$LOW if $\mathcal{R}_t \le \tau_{\text{low}}$, HIGH if $\mathcal{R}_t > \tau_{\text{high}}$, else MEDIUM.

Proposition 4.2 (Threshold Convergence). *Under a stationary distribution, $\tau_{\text{low}}, \tau_{\text{high}}$ converge almost surely to the population percentiles of $\mathcal{R}$ as $k \to \infty$ (Talagrand, 1987).*

5 Correction Methodology: Residual Spectral Grounding (RSG)

Motivation. Correction targets a root cause of fragility: attenuation of high-frequency visual signal during vision-to-language projection. The 1D DFT of H_v along the patch dimension, $\hat{H}_v(f) = \sum_p H_v[p, \cdot] \exp(-2\pi i f p / N_v)$, is empirically dominated by low-frequency modes; the higher-frequency modes encoding fine-grained spatial variation are most vulnerable to attenuation by the projection layer.

Definition 5.1 (Three-Signal Decomposition). *Given $H_v \in \mathbb{R}^{N_v \times d}$: content centroid $\mu_v = \frac{1}{N_v}\sum_p H_v[p,\cdot]$; complexity signal $\sigma_v = \sqrt{\frac{1}{N_v}\sum_p (H_v[p,\cdot]-\mu_v)^2 + \varepsilon_0}$; spectral grounding token $\phi_v = \frac{1}{K}\sum_{k=1}^{K} |\hat{H}_v(f_k^*)|$, the top-K frequency modes by mean channel magnitude.*

μ_v encodes *where* content is centred; σ_v encodes *how spread out* it is; ϕ_v, derived from the Fourier transform, is sensitive to patch *ordering*—the spatial structure most vulnerable to attenuation, making all three geometrically complementary.

Residual injection. RSG injects these signals into the decoder's *own hidden states* at inference time—not the image token stream, and not by replacing information, but by *adding* a corrective term on top of it, in the spirit of skip connections (He et al., 2015). For layers $\ell \geq \ell^* = \lfloor \rho L \rfloor$:

$$\tilde{H}_l^{(\ell)} = H_l^{(\ell)} + \frac{\alpha_\mu}{L-\ell^*}\xi(H_l^{(\ell)})\hat{\mu}_v + \frac{\alpha_\sigma}{L-\ell^*}\xi(H_l^{(\ell)})\hat{\sigma}_v + \frac{\alpha_\phi}{L-\ell^*}\xi(H_l^{(\ell)})\hat{\phi}_v, \tag{2}$$

where $\hat{v} = v/(\|v\|_2 + \varepsilon_0)$ and scale-matching factor $\xi(H) = \frac{1}{N_l}\sum_i \|H[i,\cdot]\|_2$ keeps injections proportional to hidden-state magnitude, preventing collapse (early layers) or explosion (later layers); dividing by $L-\ell^*$ bounds total injected energy per signal regardless of layer count. Dimension mismatches are handled by truncation (preserving highest-variance dimensions) or zero-padding.

We emphasize the form $\tilde{H}_l^{(\ell)} = H_l^{(\ell)} + \Delta$: a small, additive, per-sample term. This deliberately contrasts with concurrent *subtractive* corrections—attention-head suppression (Sarkar et al., 2025), latent filtering (Zhang et al., 2026), frequency ablation (Li et al., 2025)—which risk discarding information the model still needs; RSG never removes a computational path, only supplements a missing one.

Risk-coupled application. RSG applies conditionally: LOW-risk predictions are returned unmodified; MEDIUM/HIGH predictions receive Equation (2) at fixed base strengths $\bar{\alpha}_\mu, \bar{\alpha}_\sigma, \bar{\alpha}_\phi$ (Appendix I). The framework also supports continuous risk-coupling, $\alpha_\star(\mathcal{R}) = \bar{\alpha}_\star(1 + \beta_{\text{risk}}\mathcal{R})$; $\beta_{\text{risk}}=0$ recovers our reported fixed-strength default, with $\beta_{\text{risk}}>0$ left to future work.

Proposition 5.1 (Correction Stability). *For any $\beta_{\text{risk}} \geq 0$, $\left\|\tilde{H}_l^{(\ell)} - H_l^{(\ell)}\right\|_F / \left\|H_l^{(\ell)}\right\|_F \leq (\bar{\alpha}_\mu + \bar{\alpha}_\sigma + \bar{\alpha}_\phi)(1+\beta_{\text{risk}})/(L-\ell^*)$, vanishing as $L-\ell^* \to \infty$. (Proof in Appendix A.)*

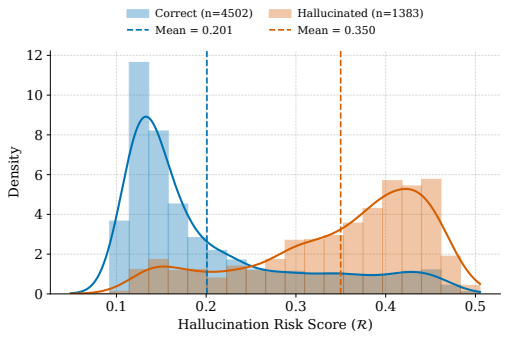
Theorem 5.2 (Spectral Grounding Improves Energy Stability). *Let $H_v' = H_v + \delta\hat{\phi}_v$, $\delta>0$. The Frobenius energy ratio under subspace erasure $\mathcal{P}_{V2}$ satisfies $\gamma_E(H_v') \geq \gamma_E(H_v)$ whenever $\left\langle \hat{\phi}_v, \Pi_k^\perp \hat{\phi}_v \right\rangle > 0$, where $\Pi_k^\perp = \mathbf{I} - U_k U_k^\top$. (Proof in Appendix A.)*

This formalizes our additive intuition: ϕ_v improves stability precisely when it carries information *outside* the model's already-dominant visual subspace—exactly the regime in which the model is hallucinating.

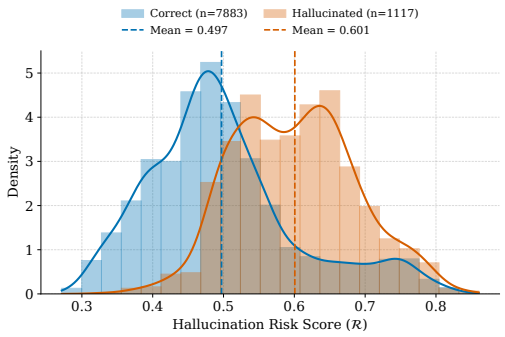
Pipeline and cost. Algorithm 1 integrates detection and correction: standard inference, then 52-attack resilience probing with where/when localization and adaptive flagging, then conditional correction only for flagged samples. Correction costs one additional forward pass with hooks active; the three-signal decomposition costs one FFT and two moment computations, negligible relative to the transformer pass. No gradients or parameter updates are needed, keeping RSG compatible with quantised inference. We avoid fine-tuning because labelled hallucination data is expensive in safety-critical domains, test-time fine-tuning risks catastrophic forgetting, and per-sample correction adapts to each image rather than a fixed training distribution.

6 EMPIRICAL RESULTS

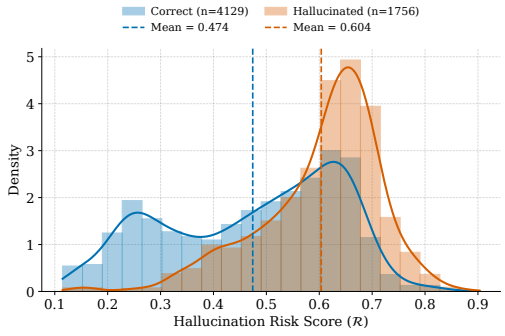
Setup. We evaluate **LLaVA-1.5 (7B)**, **Qwen2.5-VL-3B**, and **Qwen2.5-VL-7B** on **ColorBench** (Liang et al., 2025), **POPE** (object-existence hallucination) and **CableDetection** (Prueba, 2025). All 13 operators run at $\mathcal{E}=\{0.1, 0.3, 0.5, 1.0\}$, SVD rank $k=4$.



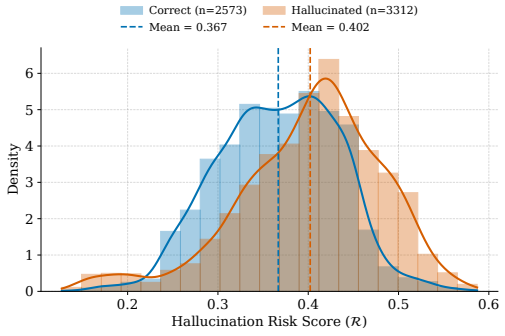
(a) Qwen2.5-VL-7B-Instruct on ColorBench



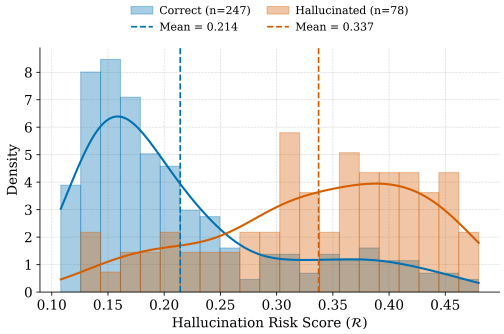
(b) Qwen2.5-VL-3B-Instruct on POPE



(c) Qwen2.5-VL-3B-Instruct on ColorBench



(d) LLaVA-1.5 (7B) on ColorBench



(e) Qwen2.5-VL-7B-Instruct on CableDetection dataset.

Figure 2: Hallucination risk score distributions, correct (blue) vs. hallucinated (orange), across all models and benchmarks.

6.1 Main Results: Where Hallucination Localizes

Figure 2a shows the cleanest case: correct ($\mathcal{R}$=0.201) and hallucinated ($\mathcal{R}$=0.350) predictions occupy almost disjoint regions for Qwen-7B on ColorBench. The full four-model/benchmark comparison shows this separability scales with model capability and is tighter, more unimodal on POPE than on ColorBench or , CableDetection dataset.

Scaling and legibility. Legibility of hallucination grows sharply with capability: on ColorBench, Qwen-7B's correlation (r=0.536) is more than double LLaVA-1.5's (r=0.231), and its risk-score Hedges' g reaches 1.496 vs. 0.478. This strong legibility for Qwen-7B remains consistent even in industrial settings, maintaining r=0.502 and $|g|$=1.353 on the CableDetection dataset. Larger models do not eliminate hallucination, but make its signature far more measurable.

Table 2: Statistical significance of hallucination risk and modality reliability across ColorBench, POPE, and CableDetection.Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

ColorBench	**Qwen-7B**	**Qwen-3B**	**LLaVA-1.5**	**POPE**	**Qwen-3B**	**CableDetection**	**Qwen-7B**
Point-Biserial r (H_{risk})	0.536^{***}	0.363^{***}	0.231^{***}	Point-Biserial r	0.312^{***}	Point-Biserial r	0.502^{***}
Spearman ρ (H_{risk})	0.502^{***}	0.372^{***}	0.259^{***}	Spearman ρ	0.335^{***}	Spearman ρ	0.476^{***}
Hedges $\|g\|$: H_{risk}	1.496^{***}	0.852^{***}	0.478^{***}	Hedges $\|g\|$: H_{risk}	0.995^{***}	Hedges $\|g\|$: H_{risk}	1.353^{***}
Hedges $\|g\|$: R_{Vision}	0.265^{***}	0.393^{***}	0.091^{***}	Hedges $\|g\|$: R_{Vision}	0.070^{*}	Hedges $\|g\|$: R_{Vision}	0.332^{**}
Hedges $\|g\|$: R_{Language}	0.049	0.218^{***}	0.002	Hedges $\|g\|$: R_{Language}	0.043	Hedges $\|g\|$: R_{Language}	0.078
Hedges $\|g\|$: R_{Joint}	0.075^{*}	0.550^{***}	0.099^{***}	Hedges $\|g\|$: R_{Joint}	0.801^{***}	Hedges $\|g\|$: R_{Joint}	0.634^{***}
Hedges $\|g\|$: R_{Overall}	0.002	0.552^{***}	0.058^{*}	Hedges $\|g\|$: R_{Overall}	0.656^{***}	Hedges $\|g\|$: R_{Overall}	0.157

Where hallucination localizes: joint cross-modal dominance. From the Table 2, comparing R_{Vision}, R_{Language}, and R_{Joint} answers *where* collapse occurs. In four of the five settings evaluated, **joint reliability is the strongest predictor**—most dramatically on POPE, where R_{Joint} ($|g|$=0.801) exceeds vision-only ($|g|$=0.070) and language-only ($|g|$=0.043, n.s.) by an order of magnitude. This cross-modal fusion failure extends to industrial tasks like CableDetection, where joint reliability ($|g|$=0.634) again strongly outweighs vision-only ($|g|$=0.332). Qwen-7B on ColorBench is the sole exception, where vision-only ($|g|$=0.265) exceeds joint ($|g|$=0.075)—indicating that at sufficient scale, fine-grained colour-attribute failures may shift toward a purer visual-encoding bottleneck. Language-only reliability is weakest or non-significant everywhere, confirming hallucination across these benchmarks is a visual-grounding failure, not a language-backbone one.

6.2 Fine-Grained Correction Analysis

We compare RSG-corrected Qwen-3B against the base model and subtractive SPIN (Sarkar et al., 2025) on seven ColorBench sub-tasks.

Table 3: Accuracy across ColorBench sub-tasks (*Average* over all seven). Best per column in **bold**.

Model	**Recognition**	**Extraction**	**Obj. Recog.**	**Proportion**	**Comparison**	**Obj. Count**	**Mimicry**	**Average**
Base	0.7105	0.3854	0.7143	0.3875	**0.4653**	0.2816	**0.5000**	0.4921
SPIN	0.3158	0.3646	0.3506	0.1875	0.3762	0.1456	0.2429	.2833
RSG (Ours)	**0.7763**	**0.4583**	**0.7273**	**0.4125**	0.4455	**0.2913**	**0.5000**	**0.5159**

From the Table 3 we observe that RSG raises average accuracy from 49.2% to 51.6% ($+2.4$ points), largest on Color Recognition ($+6.6$) and Extraction ($+7.3$)—exactly the attribute-binding tasks it targets. SPIN, by contrast, collapses accuracy to 28.3%: suppressing attention heads destroys fine-grained colour-object binding information, whereas RSG's additive injection preserves all original information while supplementing what is missing.

We further evaluate on the CableDetection dataset, where we observe that RSG (Ours) recorded **74.46%** , as compared to SPIN: 51.38% and base Qwen (3B) : 72.62%. Clearly, our method boosts accuracy without destroying all original informations.

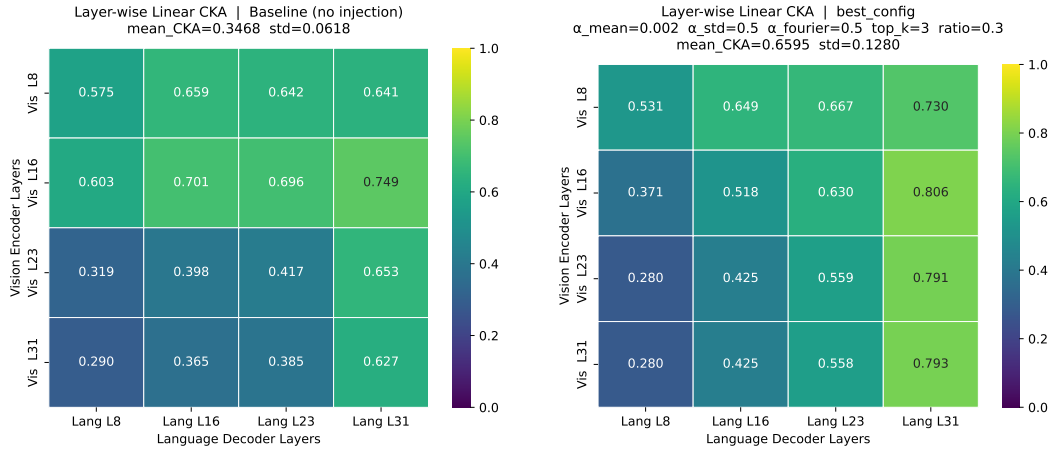
To confirm this reflects genuine grounding rather than a logit artefact, layer-wise Linear CKA between vision-encoder and language-decoder representations rises from 0.3468 (baseline) to 0.6595 (RSG)—a $\sim 90\%$ relative increase, concentrated at the *final* decoder layer (Figure:3), consistent with RSG concentrates corrections toward layers closest to the answer.

All the hyperparameter settings for the mitigation purpose, have been well discussed in Appendix I.

7 Ablation Studies

Window size k. Varying $k \in \{10, 20, 50, 100, 200\}$ (Figure 5), recall and false-positive rate of the HIGH tier *both* improve with k (recall $0.605 \rightarrow 0.665$; FPR $0.355 \rightarrow 0.245$)—unusual, since larger windows reduce percentile-estimation noise rather than trading recall against precision. Gains diminish beyond k=50, our default.

RSG hyperparameters. On Qwen-3B/ColorBench (Figure 6): combining all three signals is never uniquely best and tied-or-worst on six of eight tasks, motivating task-adaptive signal selection; a



(a) Baseline (no injection) (b) With RSG

Figure 3: Layer-wise Linear CKA between vision-encoder and language-decoder representations. RSG increases mean CKA from 0.3468 to 0.6595, concentrated at deep decoder layers.

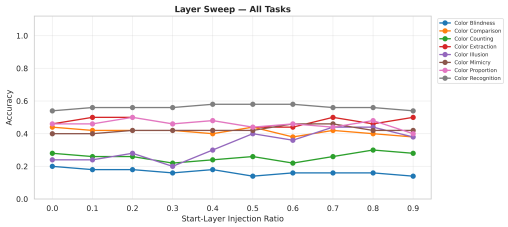


Figure 4: Accuracy as a function of the start-layer injection ratio ρ, per ColorBench task.

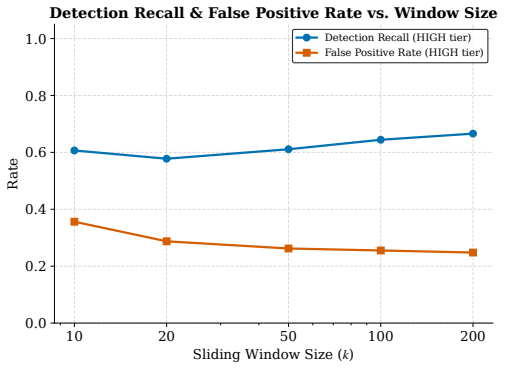


Figure 5: Detection recall/FPR of the HIGH tier vs. window size k.

narrow low-order frequency window ($[0, 7]$) is optimal for half of all tasks, suggesting dominant-energy spectral content concentrates in coarse, large-region structure rather than fine texture; and

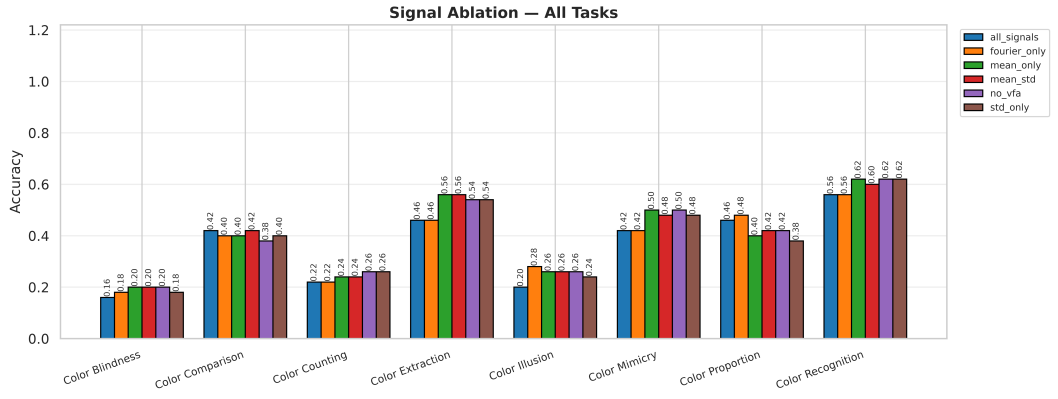


Figure 6: Accuracy under each signal-composition configuration, per ColorBench task.

injection ratio $\rho \in [0.4, 0.6]$ gives the best average trade-off, with Color Illusion specifically requiring $\rho \geq 0.5$ to recover from a sharp low-ρ failure mode.

8 Conclusion

R^2-VLM detects hallucination via *active* representational perturbation—rather than the passive probing used by all prior internal-geometry detectors—and corrects it via **Residual Spectral Grounding**, an additive injection that never suppresses information. Detection localizes *where* hallucination occurs (joint cross-modal reliability dominates in three of four settings) and *when* it occurs (via severity sweeps), backed by Propositions 4.1 and 4.2. Across three VLMs and two benchmarks, our risk score reaches r up to 0.54 and Hedges' $|g|$ up to 1.50 ($p<0.001$); our correction raises average fine-grained accuracy by 2.4 points while nearly doubling cross-modal CKA alignment, where a subtractive baseline degrades the same accuracy by nearly half.

Limitations and future work. Detection overhead scales with the number of operators; efficient approximations (random projections) are a natural next step. Calibration bounds are dataset-specific and validated only on colour- and object-existence-centric benchmarks; the risk-coupled injection strength of Section 5 is a theoretically-supported extension not yet empirically swept ($\beta_{\text{risk}}>0$); and task-adaptive signal selection, suggested directly by our own ablation, is a promising direction. We believe treating hallucination as an actively-measurable, localized geometric fragility opens a productive direction for VLM reliability research.

Ethics Statement

R^2-VLM and RSG are training-free and applied conditionally only to flagged samples, avoiding the compute cost of retraining-based mitigation. We caution that our risk score is a statistically significant but imperfect predictor, and deployments in safety-critical domains should retain human oversight. LLM assistance was used for LaTeX formatting and language refinement during manuscript preparation; all scientific content, code, and results were independently developed and verified by the authors.

Reproducibility Statement

The full formulation of R^2-VLM (Section 4, Appendix B) and RSG (Section 5) is given in the text, with default hyperparameters in Appendix I. All experiments use public benchmarks (ColorBench, POPE) and open-weight models (Qwen2.5-VL-3B/7B, LLaVA-1.5), requiring no proprietary data or access.

Responsible use of LLMs. Claude and Gemini were used to aid in coding for analysis, proof of mathematical claims and in editing text for LaTeX formatting, clarity, and grammatical refinement of the writing.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024.

Souradeep Dutta, Keshav Bulia, and Neena S Nair. Revisiting krisp: A lightweight reproduction and analysis of knowledge-enhanced vision-language models, 2025. URL https://arxiv.org/abs/2511.20795.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024. doi: 10.1109/CVPR52733.2024.01274.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. Halp: Detecting hallucinations in vision-language models without generating a single token, 2026. URL https://arxiv.org/abs/2603.05465.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

Shengzhi Li, Rongyu Lin, and Shichao Pei. Multi-modal preference alignment remedies degradation of visual instruction tuning on language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14188–14200, 2024.

Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Mitigating object hallucinations in mllms via multi-frequency perturbations. *arXiv preprint arXiv:2503.14895*, 2025.

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*, 2023.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. Colorbench: Can vlms see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4803. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d34497330b1fd6530f7afd86d0df9f76-Paper-Datasets_and_Benchmarks_Track.pdf.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*, 2024.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E. Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Prueba. Cablev2 dataset. https://universe.roboflow.com/prueba-uuq4v/cablev2-6kizd, oct 2025. URL https://universe.roboflow.com/prueba-uuq4v/cablev2-6kizd. visited on 2026-09-10.

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*, 2018.

Dvir Samuel, Guy Bar-Shalom, Fabrizio Frasca, Ethan Fetaya, Yftah Ziser, Gal Chechik, and Haggai Maron. Uniprobe: A learnable token-level hallucination detector for large vlms using multi-structural internal representations, 2026. URL https://arxiv.org/abs/2608.10835.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12492–12511, 2025.

Michel Talagrand. The glivenko–cantelli problem. *The Annals of Probability*, pp. 837–870, 1987.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Farooq Ahmad Wani, Alessandro Suglia, Rohit Saxena, Aryo Pradipta Gema, Wai-Chung Kwan, Fazl Barez, Maria Sofia Bucarelli, Fabrizio Silvestri, and Pasquale Minervini. Same answer, different representations: Hidden instability in vlms. *arXiv preprint arXiv:2602.06652*, 2026.

Zhongxing Xu, Chengzhi Liu, Qingyue Wei, Juncheng Wu, James Zou, Xin Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 38:82878–82905, 2026.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Feiran Zhang, Yixin Wu, Zhenghua Wang, Xiaohua Wang, Changze Lv, Xuan-Jing Huang, and Xiaoqing Zheng. Vib-probe: Detecting and mitigating hallucinations in vision-language models via variational information bottleneck. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23509–23521, 2026.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A PROOFS

Proof of Proposition 4.1. $\exp(-D_{\mathrm{KL}}) \leq 1$ since $D_{\mathrm{KL}} \geq 0$; $\max(0, S_{\cos}) \leq 1$ since $S_{\cos} \leq 1$; $\exp(-|\ln \gamma_E|) \leq 1$ since $|\ln \gamma_E| \geq 0$. Each term is non-negative, so $S_i \geq 0$; since weights sum to 1.0, $S_i \leq 1$, giving $S_i \in [0, 1]$. □

Proof of Proposition 5.1. Each signal is ℓ_2-normalised, so $\|\hat{v}\|_2 = 1$. The scale-matching factor satisfies $\xi(H_l^{(\ell)}) \leq \left\|H_l^{(\ell)}\right\|_F / \sqrt{N_l}$. Applying the triangle inequality to Equation (2) and bounding each term by its maximum over $\mathcal{R} \in [0, 1]$ yields the stated bound. □

Proof of Theorem 5.2. The subspace erasure operator gives $\mathcal{P}_{V2}(H_v; k, 1) = \Pi_k^{\perp} H_v$, so after injection $\mathcal{P}_{V2}(H'_v; k, 1) = \Pi_k^{\perp} H_v + \delta \Pi_k^{\perp} \hat{\phi}_v$ and $E(\mathcal{P}_{V2}(H'_v)) = \left\|\Pi_k^{\perp} H_v\right\|_F^2 + 2\delta \left\langle \Pi_k^{\perp} H_v,\, \Pi_k^{\perp} \hat{\phi}_v \right\rangle_F + \delta^2 \left\|\Pi_k^{\perp} \hat{\phi}_v\right\|_F^2$. Since $\left\|\Pi_k^{\perp} \hat{\phi}_v\right\|_F^2 = \left\langle \hat{\phi}_v,\, \Pi_k^{\perp} \hat{\phi}_v \right\rangle > 0$ by assumption, energy strictly increases for any $\delta > 0$. □

B Operator Implementation Details

Vision: (V1) $\mathcal{P}_{V1}(H_v; \varepsilon) = H_v + \varepsilon \mathbf{Z}$, $\mathbf{Z} \sim \mathcal{N}(0, I)$. (V2) subspace erasure, Section 4.2. (V3) anisotropic SVD-spectrum rescaling (amplify modes $i \leq k$ by ε, suppress $i > k$ by $1/(\varepsilon + \varepsilon_0)$). (V4) random permutation of a fraction ε of visual token rows. (V5) $\mathcal{P}_{V5}(H_v; \varepsilon) = H_v - \varepsilon \hat{d}^{\top}$, $\hat{d}$ the unit last-token hidden state. (V6) $\mathcal{P}_{V6}(H_v; \varepsilon) = (1 - \varepsilon) H_v + \varepsilon \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\top}$.

Language: (L1) semantic drift along top-k right-singular subspace of H_l. (L2) $(1 - \varepsilon) H_l + \varepsilon \mathbf{1} \bar{\mathbf{h}}_l^{\top}$. (L3) Bernoulli($\varepsilon$) token-row ablation. (L4) $H_l + \varepsilon \mathbf{f} \odot \mathbf{Z}$, $f_i = \sin(\pi i / N_l)$.

Joint: (J1) cross-modal subspace injection between H_v, H_l. (J2) energy equalisation toward equal Frobenius norm. (J3) joint manifold projection of $[H_v; H_l]$ away from its top-k joint singular subspace.

Table 4: Complete perturbation operator inventory.

ID	Modality	Operator Form	Property Tested
V1	Vision	$H_v + \varepsilon \mathbf{Z}$, $\mathbf{Z} \sim \mathcal{N}(0, I)$	Isotropic noise robustness
V2	Vision	$H_v - \varepsilon U_k U_k^{\top} H_v$	Dominant subspace dependence
V3	Vision	$U(\Sigma \odot \Lambda) V^{\top}$	Spectral distribution sensitivity
V4	Vision	Token permutation on $\lfloor \varepsilon N_v \rfloor$ tokens	Positional sensitivity
V5	Vision	$H_v - \varepsilon \hat{d}^{\top}$	Vision-to-answer pathway
V6	Vision	$(1 - \varepsilon) H_v + \varepsilon \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\top}$	Rank diversity requirement
L1	Language	$H_l + \varepsilon \mathbf{1} (V_k^{(l)} \mathbf{c})^{\top}$	Semantic drift robustness
L2	Language	$(1 - \varepsilon) H_l + \varepsilon \mathbf{1} \bar{\mathbf{h}}_l^{\top}$	Token diversity requirement
L3	Language	$\mathbf{m} \odot H_l$, $m_i \sim \mathrm{Bern}(1 - \varepsilon)$	Token ablation robustness
L4	Language	$H_l + \varepsilon \mathbf{f} \odot \mathbf{Z}$	Positional frequency sensitivity
J1	Joint	Cross-subspace injection	Cross-modal contamination
J2	Joint	Energy equalisation	Modality energy balance
J3	Joint	Joint manifold projection	Shared embedding geometry

C Latency

The detection module requires 52 SVD decompositions and 52 LM-head projections per sample, resulting in a modest increase in inference latency. On the ColorBench dataset, the resulting per-sample latency is approximately 2148.6 ms for Qwen-3B, 3285.7 ms for Qwen-7B, and 7521.8 ms for LLaVA-1.5 (7B). In contrast, hallucination mitigation introduces negligible additional latency for flagged samples, as it only involves a weighted residual connection.

D End to End R^2-VLM algorithm

The detailed end to end R^2-VLM detect and correct framework algorithm has been presented in Algorithm 1

E Dataset Curation and VQA Formulation

To systematically evaluate the model's visual reasoning capabilities on industrial wiring data, we constructed a Visual Question Answering (VQA) dataset. Leveraging the structured annotations

Algorithm 1 R^2-VLM: Full Detect-and-Correct Pipeline

Require: Input (I, T); VLM $\mathcal{M}$; operators $\{\mathcal{P}_m\}_{m=1}^{13}$; magnitudes $\mathcal{E}$; base strengths $\bar{\alpha}_\mu, \bar{\alpha}_\sigma, \bar{\alpha}_\phi$; window $\mathcal{W}$

Run $\mathcal{M}(I, T) \to H_{\text{full}}$, logits z, answer $\hat{y}$, margin ΔP; extract H_v, H_l
for each $\mathcal{P}_m, \varepsilon \in \mathcal{E}$ **do**
    Apply $\mathcal{P}_m(H_{\text{full}}; \varepsilon) \to H'$; compute $\gamma_E, S_{\cos}, D_{\text{KL}}, S_i$ (Eq. (1))
end for
Compute $R_v, R_l, R_J, R_{\text{overall}}$, then $\mathcal{R}$; percentile-flag L_t using $\mathcal{W}$
if $L_t =$ LOW **then**
    return $\hat{y}$
else
    Compute μ_v, σ_v, ϕ_v; re-run $\mathcal{M}(I, T)$ with RSG hooks (Eq. (2)); **return** $\tilde{y}$
end if

from our ground-truth label files, we programmatically generated a diverse set of question-answer pairs for each image from the CableDetection dataset (Prueba, 2025). The question templates were designed to target four distinct semantic dimensions: counting, existence, structural/spatial attributes, and multi-label identification. The exact question templates utilized during the curation process are detailed below:

1. **Object Counting Tasks:** Designed to evaluate the model's granular quantification capabilities.
    - "How many wires in total are visible in this image?"
    - "How many ⟨*color*⟩ wires are in this image?"
2. **Existence and Verification Checks:** Formulated as binary classification tasks to test both affirmative and negative existential reasoning.
    - "Is there a ⟨*color*⟩ wire in this image?"
    - "Is there no ⟨*color*⟩ wire in this image?"
3. **Structural and Spatial Attributes:** Aimed at evaluating high-level topological features and spatial orientations within the scene.
    - "Does this image contain a wire connection or junction?"
    - "Is the cable in this image positioned at an angle?"
4. **Color Identification (Multi-label):** Evaluates the model's ability to perform open-ended or multi-class recognition of all active attributes.
    - "Which wire colors are present in this image?"
    - "Which wire color(s) appear in this image?"

The placeholder ⟨*color*⟩ was dynamically mapped to the English color terms (e.g., *red*, *blue*, *green*) present in the corresponding image metadata. Ground-truth answers were automatically derived from the label files to ensure absolute label fidelity.

F LINEAR TRANSFORMATION

$$\mathcal{F}_J = \text{clip}\left(\frac{\tau_J^{\max} - R_J}{\tau_J^{\max} - \tau_J^{\min} + \varepsilon_0}, 0, 1\right), \tag{3}$$

$$\mathcal{F}_{\text{overall}} = \text{clip}\left(\frac{\tau_O^{\max} - R_{\text{overall}}}{\tau_O^{\max} - \tau_O^{\min} + \varepsilon_0}, 0, 1\right), \tag{4}$$

and the weights are $w_J = 0.45$, $w_u = 0.35$, $w_f = 0.20$ (summing to 1.0). The upper and lower bounds for the joint thresholds ($\tau_J^{\max} = 0.85$, $\tau_J^{\min} = 0.60$) and overall thresholds ($\tau_O^{\max} = 0.88$, $\tau_O^{\min} = 0.75$), alongside the component weights, are empirically calibrated min-max bounds based on the reliability distributions observed on the ColorBench validation set across the tested models. Furthermore, this calibration strategy exhibits strong generalization, maintaining highly effective performance, also when evaluated on the POPE dataset.

G Dataset Statistics

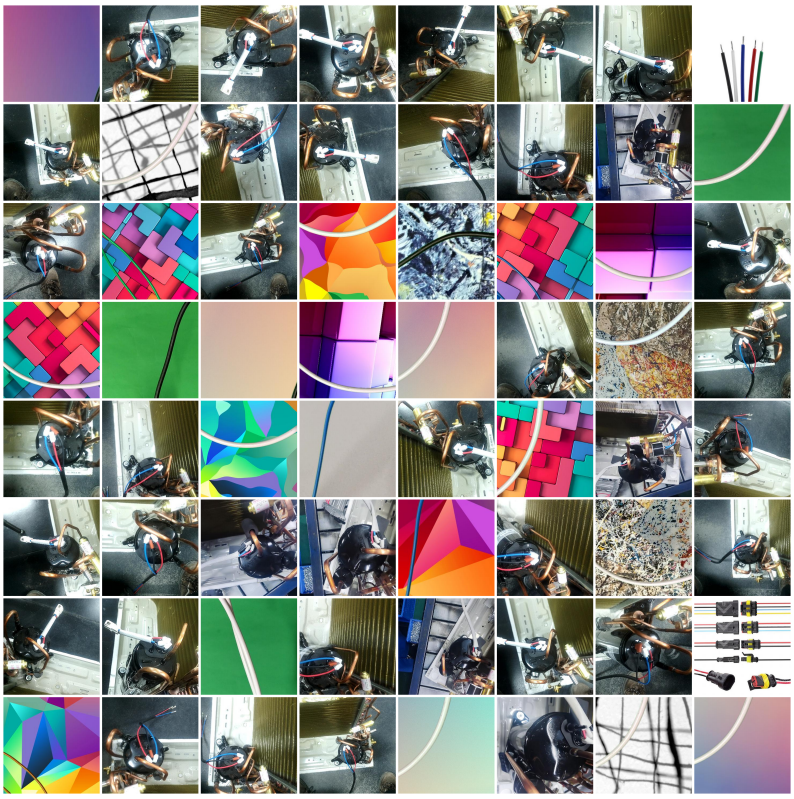


Figure 7: A glimpse of the CableDetection Dataset.

Table 5 summarises the three visual question-answering (VQA) resources referenced in this work: the color-perception benchmark **ColorBench** Liang et al. (2025), the hallucination-probing benchmark **POPE** Li et al. (2023), and our own **curated Industrial Wire-QA** dataset, constructed to evaluate fine-grained color and structural reasoning in industrial wiring imagery.

Table 5: Comparative dataset statistics. Values for ColorBench and POPE are drawn from their respective publications; values for the curated Wire-QA dataset are computed directly from the released annotation file.

Property	ColorBench	POPE	Wire-QA (Ours)
Domain	Natural images (color reasoning)	Natural images (COCO / A-OKVQA / GQA)	Industrial wiring / cable imagery
Source images	1,448 instances	500 images per source dataset (1,500 total)	14 unique images
Total QA pairs	5,814	27,000 (3 datasets $\times$ 3 splits $\times$ 3,000)	325
Avg. QA / image	≈ 4.0	6 per image per split ($\times$ 3 splits = 18)	26
# Task / question types	11 (3 dimensions: Perception, Reasoning, Robustness)	1 (binary object existence)	8 (counting, presence, absence, structural, listing)
Answer format	Multiple-choice (3–6 options)	Binary (Yes/No)	Mixed: numeric (0–3), Yes/No, True/False, color name(s)
Sampling strategy	Curated per task template	Random / Popular / Adversarial negative sampling	Exhaustive per-color template over fixed label set
Target model behavior probed	Color perception & reasoning robustness	Object-existence hallucination	Color attribute & structural (junction/angle) hallucination

Table 6 decomposes our curated dataset by question type, confirming internal consistency between the number of source images, the fixed color vocabulary, and the total question count.

H Failure Analysis: Detection Legibility at Lower Model Scales

While R^2-VLM demonstrates strong separability between grounded and hallucinated predictions for larger models, empirical analysis reveals a significant boundary condition at lower parameter scales when processing complex, domain-specific imagery.

Figure 8 illustrates the hallucination risk score ($\mathcal{R}$) distribution for Qwen2.5-VL-3B on the industrial Wire-QA dataset. Unlike the clean margin observed with Qwen2.5-VL-7B on the same dataset

Table 6: Breakdown of the curated Wire-QA dataset (364 total QA pairs, 14 images, 7 wire colors: Black, Blue, Green, Orange, Red, White, Yellow).

Question Type	Template	Count	Answer Space
`count_color`	"How many {color} wires...?"	$7 \times 14 = 98$	$\{0, 1, 2, 3\}$
`presence_check`	"Is there a {color} wire...?"	$7 \times 14 = 98$	Yes / No
`absence_check`	"Is there no {color} wire...?"	$7 \times 14 = 98$	True / False
`total_wires`	"How many wires in total...?"	$1 \times 14 = 14$	$\{0, 1, 2, 3\}$
`connection_check`	"...wire connection or junction?"	$1 \times 14 = 14$	Yes / No
`angle_check`	"...cable positioned at an angle?"	$1 \times 14 = 14$	Yes / No
`list_colors`	"Which wire colors are present...?"	$1 \times 14 = 14$	Color name(s) / None
`color_identification`	"Which wire color(s) appear...?"	$1 \times 14 = 14$	Color name(s) / None
Total		**364**(325 excluding corrupted files)	—

(Section 6.1), the 3B model exhibits severe distributional overlap between correct (mean $\mathcal{R} = 0.436$) and hallucinated (mean $\mathcal{R} = 0.492$) predictions. The density curves are highly entangled across the entire scoring range, lacking a decisive threshold for adaptive flagging and rendering the risk score largely uninformative.

This corroborates the capability-scaling trend established earlier: hallucination legibility is intrinsically tied to model capacity. We hypothesize that lower-parameter models lack the robust, highly structured internal geometry necessary to produce a clean differential signal under perturbation. When a 3B model encounters challenging structural reasoning tasks—such as fine-grained cable counting or junction detection—its visual and cross-modal representations are inherently noisier and more fragile, even when answering correctly. Consequently, our structural stress tests record a high baseline fragility across all predictions, blurring the diagnostic boundary between genuine visual grounding and spurious correlation. This suggests that training-free, geometry-based detection methods require a minimum representational capacity to be viable, establishing a practical lower bound for the application of R^2-VLM.

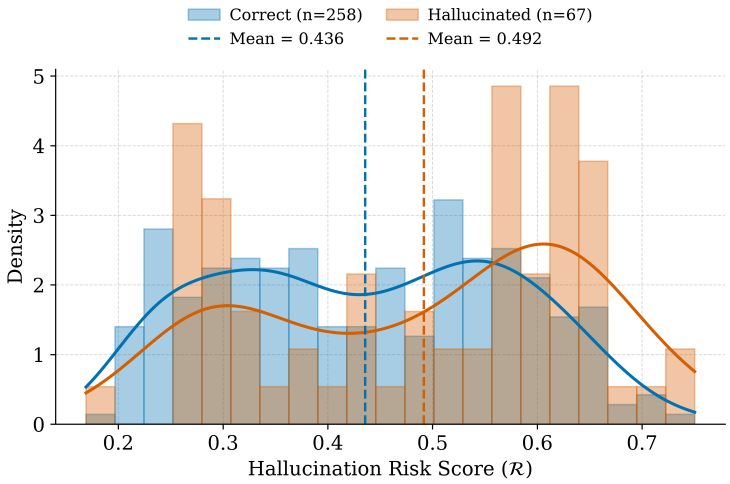


Figure 8: Failure Case: Hallucination risk distribution for Qwen2.5-VL-3B on the industrial Wire-QA dataset. The severe overlap between correct and hallucinated densities demonstrates degraded detection legibility at lower parameter scales.

I Hyperparameter Settings

Table 7 tabulates all the hyper parameters used for hallucination detection. Table 8 and Table 9 tabulates all the hyper parameters used in RSG module for ColorBench and CableDetection dataset respectively. Max_token is kept 1 across all baselines.

Table 7: Default detection hyperparameters.

Parameter	Value	Description
$\mathcal{E}$	$\{0.1, 0.3, 0.5, 1.0\}$	Perturbation magnitudes
k (SVD rank)	4	Truncation rank for all SVD operators
w_J, w_u, w_f	$0.45, 0.35, 0.20$	Risk score weights
α, β, γ	$0.40, 0.30, 0.30$	Decision/direction/energy weights in S_i
k_{window}	50	Sliding window size
$P_{\text{low}}, P_{\text{high}}$	$33, 66.6$	Flagging percentile thresholds
ε_0	10^{-8}	Numerical stabiliser
τ (softmax temp.)	1.0	Temperature for option probabilities

Table 8: Default RSG correction hyperparameters for ColorBench dataset.

Parameter	Symbol	Value	Description
`alpha_mean`	$\bar{\alpha}_\mu$	0.5	Base strength for content centroid μ_v
`alpha_std`	$\bar{\alpha}_\sigma$	0.02	Base strength for complexity signal σ_v
`alpha_fourier`	$\bar{\alpha}_\phi$	0.02	Base strength for spectral grounding token ϕ_v
`top_k_freq`	K	3	Dominant frequency modes retained for ϕ_v
`inj_layer_ratio`	ρ	0.3	Total hooked layers $\ell^* = \max\{1, L - \lfloor \rho L \rfloor\}$

Table 9: Default RSG correction hyperparameters for CableDetection dataset.

Parameter	Symbol	Value	Description
`alpha_mean`	$\bar{\alpha}_\mu$	0.05	Base strength for content centroid μ_v
`alpha_std`	$\bar{\alpha}_\sigma$	0.06	Base strength for complexity signal σ_v
`alpha_fourier`	$\bar{\alpha}_\phi$	0.05	Base strength for spectral grounding token ϕ_v
`top_k_freq`	K	3	Dominant frequency modes retained for ϕ_v
`inj_layer_ratio`	ρ	1	Total hooked layers $\ell^* = \max\{1, L - \lfloor \rho L \rfloor\}$

J Device Configuration

All experiments have been performed on a Quadro RTX 6000 (24GB VRAM) GPU , Intel(R) Xeon(R) Gold 6230 CPU with 16GB RAM in Ubuntu environment.