Same Answer, Different Representations: Hidden instability in VLMs

Anonymous ACL submission

Abstract

The robustness of Vision Language Models (VLMs) is commonly assessed through output-level invariance, implicitly assuming that stable predictions reflect stable multimodal processing. In this work, we argue that this assumption is insufficient. We introduce a representation-aware and frequency-aware evaluation framework that measures internal embedding drift, spectral sensitivity, and structural smoothness (spatial consistency of vision tokens), alongside standard label-based metrics. Applying this framework to modern VLMs across the SEEDBench, MMMU, and POPE datasets reveals three distinct failure modes. First, models frequently preserve predicted answers while undergoing substantial internal representation drift; for perturbations such as text overlays, this drift approaches the magnitude of inter-image variability, indicating that representations move to regions typically occupied by unrelated inputs despite unchanged outputs. Second, robustness does not improve with scale; larger models achieve higher accuracy but exhibit equal or greater sensitivity, consistent with sharper yet more fragile decision boundaries. Third, we find that perturbations affect tasks differently: they harm reasoning when they disrupt how models combine coarse and fine visual cues, but on the hallucination benchmarks, they can reduce false positives by making models generate more conservative answers.

1 Introduction

Vision Language Models (VLMs) are increasingly deployed in real-world applications that require multimodal perception and reasoning, enabling systems to answer questions, follow instructions, and understand diverse visual inputs (Bai et al., 2024; Deitke et al., 2024). As these systems are used more widely, it becomes critical to understand whether they can reliably maintain their predictions when images undergo minor transformations that preserve semantic interpretation.

Most existing robustness evaluations for VLMs implicitly assume that output-level robustness (i.e., whether a model gives the same answer to slightly perturbed versions of the same image) indicates robust processing (Fang et al., 2022). However, recent studies suggest that output-level stability masks hidden instability, particularly in overparameterised models where correct predictions can persist despite significant volatility in internal representations (Chandhok et al., 2025; Liu et al., 2025). More broadly, research on the latent dynamics of large language models (LLMs) shows that hidden representations can drift substantially even when output tokens remain correct (Wang et al., 2025; Khanmohammadi et al., 2025; Dies et al., 2025). This *representation drift* is concerning because it may mask decision boundary instability that could lead to failures in downstream applications or brittle behaviour under additional perturbations. In this work, we aim to answer: *Does output-level robustness in VLMs similarly mask representation drift?* Specifically, do VLMs maintain predictions under image perturbations while exhibiting representation drift?

VLMs present a unique challenge compared to LLMs: predictions emerge from the interaction of multiple components (i.e., vision encoders, connector layers, and language model backbones), making internal instability more complex and harder to diagnose (Karamcheti et al., 2024). To study this systematically, we examine two different failure modes: *reasoning stability* (whether models maintain correct logical deductions under perturbations) and *hallucination dynamics* in hallucination tasks (whether models correctly verify object existence when images are altered). We evaluate reasoning stability using SEED-Bench (Li et al., 2024) and MMMU (Yue et al., 2024), and hallucination dynamics using POPE (Li et al., 2023b). We evaluate multiple VLMs of different scales within the Qwen3-VL and LLaVA families on these bench-

marks under multiple perturbation types: geometric transformations (i.e., scaling, rotation, translation), spatial modifications (i.e., cropping, padding), and text overlays. We develop an evaluation framework that jointly measures label stability, embedding consistency across multiple representation levels, and the structural properties of internal activations.

We identify several phenomena: i) A substantial disconnect between output and representation stability; 37.6% of images are affected by at least one perturbation type, and in some cases, the model preserves its predicted answer under perturbation despite a substantial representation drift. ii) Text overlays are particularly disruptive (19.2% answer flip rate), while geometric perturbations cause smaller but measurable changes (6–8%). iii) Most perturbations produce representation drifts comparable in magnitude to inter-image variability, suggesting that such perturbations move internal representations to regions that are typically occupied by different inputs despite unchanged outputs. iv) Model scale does not improve robustness; larger models achieve higher accuracy but exhibit comparable or greater representation drift under perturbations. v) On reasoning tasks, perturbations lead to more random errors. In contrast, on hallucination tasks, perturbations cause the model to make more conservative predictions, reducing false positives.

These observations suggest that assessing VLM robustness requires more than checking output invariance. Our evaluation framework provides a more complete diagnosis of where and why VLMs fail under perturbations, measuring their robustness beyond output consistency.

2 Related Work

Robustness and Internal Consistency in VLMs. Robustness in vision-language models has been studied primarily through the lens of task accuracy and output consistency. Extensive work has characterised VLM failure modes under adversarial attacks (Zhao et al., 2023), geometric transformations (Ishmam et al., 2025), and hallucination triggers (Guan et al., 2024; Li et al., 2023b). However, these evaluations typically treat the model as a black box, assuming that output stability implies robust processing. In parallel, research on Large Language Models (LLMs) has challenged this assumption, demonstrating that models can preserve identical outputs while undergoing substantial drift in hidden states and confidence margins (Turpin et al., 2023; Wang et al., 2025; Nishida et al., 2025). Such *latent instability* has been linked to calibration failures and instability in chain-of-thought reasoning. We bridge these two lines of inquiry. Unlike prior VLM benchmarks that focus exclusively on output correctness (Yue et al., 2024; Duan et al., 2024), we analyze how internal representations evolve under perturbations. We explicitly measure representation drift across multiple context and answer embedding regimes, showing that output-level invariance in VLMs often masks significant multimodal instability.

Spectral Analysis and Structural Drift. A large body of work defines robustness through the frequency domain. Early analyses demonstrated that CNNs and Vision Transformers (ViTs) exhibit a strong reliance on low-frequency features, often failing when high-frequency textures are perturbed (Yin et al., 2019; Shao et al., 2021). Furthermore, discretization steps in ViTs, such as patchification, were shown to introduce aliasing artifacts that exacerbate spectral sensitivity (Abraham et al., 2025; Qian et al., 2021). Our work formalises the intuition that "benign" natural perturbations (e.g., rotation, scaling) are actually sources of induced spectral drift. We extend frequency analysis from pure vision backbones to the full VLM pipeline, demonstrating that robustness failures are better explained by *cross-frequency misalignment* rather than simple low-frequency reliance. To quantify the structural impact of these shifts, we adopt *Dirichlet energy*, traditionally used to measure spatial smoothness in graph neural networks (Zhou et al., 2021), and apply it here to track the structural reorganization of vision tokens under perturbation.

Evaluation Protocols: Generation vs. Scoring. VLM evaluation protocols often rely on free-form text generation, which complicates robustness analysis due to decoding stochasticity and formatting ambiguity. To isolate decision stability from generation noise, we adopt a multiple-choice scoring protocol based on the log-likelihood of answer options, which is popular in VLM (Li et al., 2024) and LLM (Gao et al., 2024) evaluation benchmarks. This approach allows us to track *margin dynamics*, the confidence gap between the correct answer and the top distractor. By combining decision-level margins with representation-level metrics, our framework provides a diagnostic view of robustness, distinguishing between cases where the model is robust (*stable internal state*) versus cases

where it is merely fortuitous (*drifted state but preserved prediction*).

3 Methodology and Experiments

Our evaluation framework is designed to decouple *decision instability* (output changes) from *representation instability* (representation drift). We conduct experiments on SEEDBench (Li et al., 2024), MMMU (Yue et al., 2024), and POPE (Li et al., 2023b), restricting our scope to multiple-choice and binary decision subsets. This setting is particularly well-suited for robustness analysis because i) the semantic task remains fixed across perturbations, ii) decisions are discrete and directly comparable, and iii) confidence margins can be measured reliably. A detailed description of the datasets, models, perturbation examples, and prompt templates is provided in Section A.

3.1 Scoring Protocol and Margin Dynamics

We use a log-likelihood scoring-based evaluation protocol. More formally, let x denote the visual input. For SEEDBench and POPE, x consists of a single image, while for MMMU, x represents a sequence of images, denoted as $x = \{I_1, I_2, \dots\}$. Given a question q and a set of answer options $\mathcal{A}$, we construct a prompt $p(q, \mathcal{A})$ and compute the conditional log-likelihood for each option $o \in \mathcal{A}$:

$$s(o \mid x) = \log P_\theta(o \mid x, p(q, \mathcal{A})), \qquad (1)$$

where θ represents the frozen VLM parameters. The predicted label is $\hat{y}(x) = \arg\max_{o \in \mathcal{A}} s(o \mid x)$. Unlike free-form generation, this allows us to track *margin dynamics*. We define the margin as the log-probability gap between the predicted option o^* and its strongest competitor:

$$\text{margin}(x) = s(o^* \mid x) - \max_{o \neq o^*} s(o \mid x). \qquad (2)$$

Intuitively, a positive but shrinking margin indicates latent instability even when the output label remains unchanged.

3.2 Robustness Metrics

We evaluate robustness under six families of *natural* perturbations v'. The perturbation families include translation (cyclic shifts), padding/cropping, scaling, rotation, and text overlays. For text overlays, we evaluate three variants with identical geometry but different content: semantic directives (`TextOverlay`), random strings (`RandomText`), and empty boxes (`BoxOverlay`). Parameter ranges and specific values are detailed in Section A.6. We quantify stability using two metrics: *Instance Flip Rate* (AVg) and *Image Vulnerability* (Ve).

Instance Flip Rate. AVg measures the failure frequency across all perturbation trials. Let $\mathcal{I}_{v'}$ be the set of all perturbed pairs (x, x') for perturbation type v':

$$\text{AVg}_{v'} = \frac{1}{|\mathcal{I}_{v'}|} \sum_{(x,x') \in \mathcal{I}_{v'}} \mathbb{1}\{\hat{y}(x') \neq \hat{y}(x)\}. \qquad (3)$$

Image Vulnerability. Measures the fraction of unique images that are vulnerable to *at least one* instance of perturbation v'. Let $\mathcal{X}$ be the set of unique base samples:

$$\text{Ve}_{v'} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{1}\{\exists x' \in v'(x) : \hat{y}(x') \neq \hat{y}(x)\}. \qquad (4)$$

3.3 Representation-Aware Analysis

We employ four diagnostic tools to measure internal drift, namely *Embedding Stability*, *Structural Smoothness* (also referred to as *Dirichlet Energy*), *Perturbation and Control Drift*, and *Drift-to-Prior*.

Embedding Stability. We measure the cosine distance and ℓ_2 norm between the base embedding $e(x)$ and perturbed embedding $e(x')$ across five distinct regimes—isolating where in the pipeline the visual grounding erodes. We extract hidden states from the final layer, before logits, at five hook points: `ctx_open` and `ctx_mcq` (last context token), `ans_open` and `ans_mcq` (mean-pooled answer tokens), and `ans_mcq_free` (constrained free generation). See Section A.7 for more details.

Structural Smoothness (Dirichlet Energy). To quantify spatial reorganization, we treat vision tokens of image x as nodes in a grid graph Z. Let $\mathbf{z}_{i,j} \in \mathbb{R}^d$ be the embedding of the patch at coordinates (i, j), and let $Z = \{\mathbf{z}_{i,j}\}$ denote the full grid of patch embeddings. The Dirichlet energy $E(Z)$ of Z measures the total variation between adjacent patches:

$$E(Z) = \sum_{(i,j) \sim (i',j')} ||\mathbf{z}_{i,j} - \mathbf{z}_{i',j'}||_2^2 \qquad (5)$$

We report the energy gap $\Delta E = E(Z_{\text{pert}}) - E(Z_{\text{base}})$. A positive ΔE implies the perturbation has introduced high-frequency structural noise, while a negative ΔE implies over-smoothing.

Perturbation Drift vs. Control Drift. To contextualise perturbation-induced representation drift, we compare it against a control baseline. For each base image, we compute distances between its embedding across regimes and the embeddings of randomly sampled other images, yielding a control-drift distribution. We quantify separation using Cohen's d (Cohen, 1988). Large negative values indicate that perturbation-induced drift is substantially smaller than inter-image variability, while values closer to zero indicate non-local displacement in representation space (detailed in Section A.8).

Drift-to-Prior (POPE). To distinguish between visual hallucinations and language bias, we define a *Blank Image Baseline*. We calculate the prior score $S_{\text{prior}} = P(\text{“Yes”} \mid q, \text{blank})$, and measure if perturbations shift the margin closer to S_{prior}, effectively forcing the model to abandon visual evidence in favour of its base language prior (Section 7).

4 Natural Perturbations: Output Instability and Latent Representational Drift

We now analyse the effect of natural visual perturbations on both model outputs and internal representations. Our goal is to answer two questions: i) how often do predictions change under meaning-preserving transformations, and ii) what happens internally when predictions do *not* change.

Unless stated otherwise, results in this section are reported for a single representative model, `Qwen3-VL-2B Instruct` and SEEDBench dataset, while the same evaluation protocol is applied consistently across datasets (SEEDBench, MMMU, and POPE) and architectural families; additional results can be found in Section C and Section D. Cross-dataset and cross-model consistency is analysed explicitly in Section 5.

4.1 Label Instability Under Natural Perturbations

We first report results on SEEDBench, which we use as a primary illustrative benchmark due to its diverse visual reasoning tasks and multiple-choice structure. Each experiment is conducted over *four independent runs*, each using a disjoint random subset of 3,500 images. In total, this corresponds to approximately 14,000 evaluated samples, and all reported statistics are averaged across runs.

The average base accuracy on unperturbed images is 61.7%. Table 1 reports the instance flip rate (AVg) and image-level flip probability (Ve) for each perturbation type.

Even simple geometric perturbations induce non-trivial instability. For translation and pad/crop, approximately 16–17$\%$ of images exhibit at least one perturbation instance that flips the predicted answer. Rotation further increases instability, while text overlays are the most disruptive: they exhibit the highest instance flip rate (AVg $\approx 19.2\%$) and image-level flip probability (Ve $\approx 23.9\%$).

When considering the union of all perturbations, over one third of images (Ve $\approx 37.6\%$) experience at least one decision flip. This suggests that robustness failures are not isolated corner cases but arise across a broad range of natural transformations.

These perturbations preserve high-level semantic content and visual plausibility. Thus, the observed flips reflect sensitivity to representational changes rather than semantic corruption. Notably, the union AVg is lower than the maximum single-perturbation AVg because it aggregates across heterogeneous perturbation families, while Ve captures worst-case image-level vulnerability.

Revisiting Overlays: Semantics vs. Occlusions. Decomposing overlays reveals a clear ordering. `TextOverlay` (semantic, e.g., “Answer is X”) is consistently the most disruptive variant, followed by `RandomText`, while `BoxOverlay` is comparatively benign. This indicates that the majority of overlay-induced failures cannot be attributed to occlusion alone: *semantic or instruction-like visual content* substantially amplifies decision instability beyond what is explained by masking(BoxOverlay) or edge injection(RandomText).

4.2 Correctness Transitions: Perturbations Can Fix and Break Predictions

To understand whether perturbations systematically degrade performance or move predictions bidirectionally, we distinguish four transition types: R→W (correct → wrong), W→R (wrong → correct), R→R (stays correct), and W→W (stays wrong). Table 2 reports transition counts for each perturbation family.

All perturbations exhibit both R→W and W→R transitions, showing they induce movement across decision boundaries in both directions rather than acting as uniform noise. For TextOverlay, W→R transitions (881) are comparable to R→W (685), indicating overlays can correct wrong predictions despite increasing instability. This bidirectionality

suggests non-trivial representational shifts rather than systematic degradation.

4.3 Embedding Invariance: The Answer Can Stay While the Model Moves

Decision stability alone can give a misleading impression of robustness. To probe internal behaviour, we analyse representation drift across multiple context- and answer-conditioned representation regimes (Section A.7). Table 3 reports the mean cosine similarity and L2 distance between base and perturbed embeddings, averaged across perturbations and runs.

Pattern 1: TextOverlay induces the largest drift. `TextOverlay` produces substantially larger drift than geometric perturbations across all regimes. In `ctx_open`, cosine similarity drops to 0.866 with L2 distance exceeding 1000, compared to ~ 0.99 for translation and pad/crop.

Pattern 2: MCQ conditioning stabilises context representations. Context embeddings under MCQ prompts (`ctx_mcq`) are more stable than under open prompts (`ctx_open`), indicating explicit option conditioning constrains the representation space. However, this does not eliminate answer-level instability.

Predictions may remain unchanged while representations undergo substantial movement, highlighting a gap between output-level robustness and internal stability that motivates the control-drift analysis next.

4.4 Drift Versus Control Drift

To assess whether perturbation-induced drift is merely local or comparable to inter-image variation, we compare it against a control baseline. For each image, we compute distances to randomly sampled other base images and quantify separation using Cohen's d.

Tables 4 (and 6 in the Appendix) summarise results for the `ans_mcq_free` regime, which exhibits the strongest and most interpretable effects.

Translation, pad/crop, and scale produce drift well separated from control drift (large negative Cohen's d), indicating local deformations, while TextOverlay produces drift substantially overlapping with control drift (Figure 1, Figure 4(Appendix) and Table 4). Mean perturbation drift (≈ 0.177) approaches control mean (≈ 0.228), yielding Cohen's $d \approx -0.34$,

Perturbation	AVg (flip-rate)	Ve (vulnerability)
Translation	0.062	0.168
Pad/Crop	0.065	0.169
Scale	0.079	0.079
Scale+Pad	0.080	0.100
Rotation	0.122	0.166
TextOverlay(semantic)	0.192	0.239
TextOverlay(random)	0.064	0.086
TextOverlay(empty)	0.043	0.044
Any (union)	0.079	0.376

Table 1: Label invariance statistics on SEEDBench samples using log-likelihood MCQ scoring.

Perturbation	R→W	W→R	R→R	W→W	GT inst.
Translation	694	618	16538	10094	27944
Pad/Crop	725	642	16507	10070	27944
Scale	95	106	2059	1233	3493
Scale+Pad	232	181	4076	2497	6986
Rotation	390	262	3918	2416	6986
TextOverlay	685	881	5738	3126	10430
Any	2821	2690	48836	29436	83783

Table 2: Prediction correctness transitions under natural perturbations. R→W: base-correct predictions that become wrong under perturbation; W→R: base-wrong predictions that become correct. Counts are reported at the perturbation-instance level for 3500 base samples.

implying text overlays move representations significantly toward regions of unrelated images.

Even with unchanged predictions, representations may no longer reside in neighborhoods of the original image. This non-local movement reduces stability and explains why subsequent perturbations easily induce failure.

These results expose a gap between output robustness and internal stability: geometric transformations trigger flips despite small changes, while text overlays induce inter-image-level drift, motivating the spectral analysis next.

These findings motivate a deeper analysis of hidden representations and their spectral structure, which we explore next.

4.5 Vision-Token Smoothness and Dirichlet Energy

To complement embedding drift, we analyze *Dirichlet energy* to quantify spatial reorganization. We extract vision tokens from the vision encoder's output, then reshape the flattened sequence back into a 2D grid. We then compute $\Delta E_{\text{dir}} = E_{\text{dir}}(x') - E_{\text{dir}}(x)$ for each perturbation, measuring the deviation in local structural smoothness. Different perturbations induce distinct patterns: Translation shows small positive shifts

Type	ctx_open (cos / L2)	ctx_mcq (cos / L2)	ans_open (cos / L2)	ans_mcq (cos / L2)	ans_mcq_free (cos / L2)
Translation	0.989 / 226	0.995 / 104	0.980 / 141	0.994 / 82	0.991 / 65
Pad/Crop	0.988 / 245	0.995 / 113	0.979 / 153	0.993 / 88	0.991 / 70
Scale	0.984 / 283	0.993 / 132	0.977 / 171	0.991 / 103	0.990 / 80
Scale+Pad	0.981 / 299	0.992 / 145	0.974 / 184	0.990 / 113	0.988 / 89
Rotation	0.963 / 458	0.976 / 262	0.959 / 276	0.983 / 182	0.980 / 142
TextOverlay	0.866 / 1027	0.918 / 551	0.922 / 465	0.963 / 310	0.820 / 484
Any	0.970 / 360	0.984 / 181	0.970 / 201	0.988 / 124	0.968 / 128

Table 3: Embedding invariance summary (mean similarity to base) across five representation regimes.

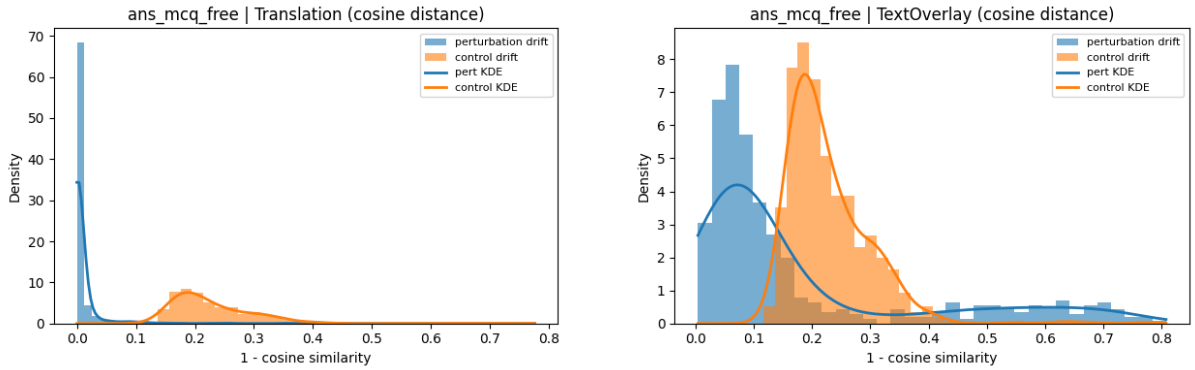

Figure 1: Cosine distance ($1 - \cos$), Drift versus control drift for the `ans_mcq_free` embedding under Translation and Textoverlay perturbation. Blue shows perturbation-induced drift relative to the base image; orange shows control drift (base image versus randomly sampled other images). Left: Translation. Right: Textoverlay. Unlike geometric perturbations, the Textoverlay perturbation-induced distribution does not remain well separated from control drift, indicating that the representation no longer stays local in embedding space.

Perturbation	μ_{pert}	σ_{pert}	μ_{ctrl}	Cohen's d
Translation	0.0093	0.0338	0.2283	-3.8838
Pad/Crop	0.0104	0.0381	0.2283	-3.7735
Scale	0.0109	0.0499	0.2283	-3.5028
Scale+Pad	0.0129	0.0427	0.2283	-3.6292
Rotation	0.0200	0.0422	0.2283	-3.5210
TextOverlay	0.1773	0.2022	0.2283	-0.3355

Table 4: Drift versus control drift for the `ans_mcq_free` embedding measured using cosine distance ($1 - \cos$). μ_{pert} and σ_{pert} denote mean and standard deviation of perturbation-induced drift (base vs. perturbed), while μ_{ctrl} denotes mean control drift (base vs. randomly sampled images). Cohen's d quantifies standardised separation between distributions, with values near 0 indicating stronger overlap.

(10.34 ± 67.49), TextOverlay produces negative shifts (-33.87 ± 60.14), and Rotation exhibits the strongest effect (-72.73 ± 99.95). Flip-inducing instances consistently show larger absolute deviations, indicating that spatial reorganization correlates with decision failures. Dirichlet energy and embedding drift capture complementary aspects: drift measures *where* representations move, while Dirichlet energy measures *how* spatial structure reorganises. Together, these show that robustness failures arise from structural and spectral drift that misaligns visual representations with decision boundaries. Complete analyses are in Appendix E.

5 Scaling Behaviour and Cross-Dataset Robustness

This section evaluates whether the robustness phenomena identified in the previous section persist across model scale, datasets, and architectural families. All experiments use the same perturbation suite, scoring protocol, and metrics from Section 3.

5.1 Model Scaling on SEEDBench

We extend the natural perturbation analysis to multiple model scales within the Qwen3-VL (Instruct) family, evaluating `Qwen3-VL-2B`, `4B`, `8B`, and `32B` on the same SEEDBench subset. This allows us to examine whether robustness improves monotonically with model capacity.

Base accuracy improves from 2B to 8B, but slightly degrades at 32B due to the restriction of answers to fixed multiple-choice options under log-likelihood scoring. More importantly, relative stability does not scale linearly with accuracy; larger models often exhibit comparable or higher flip rates

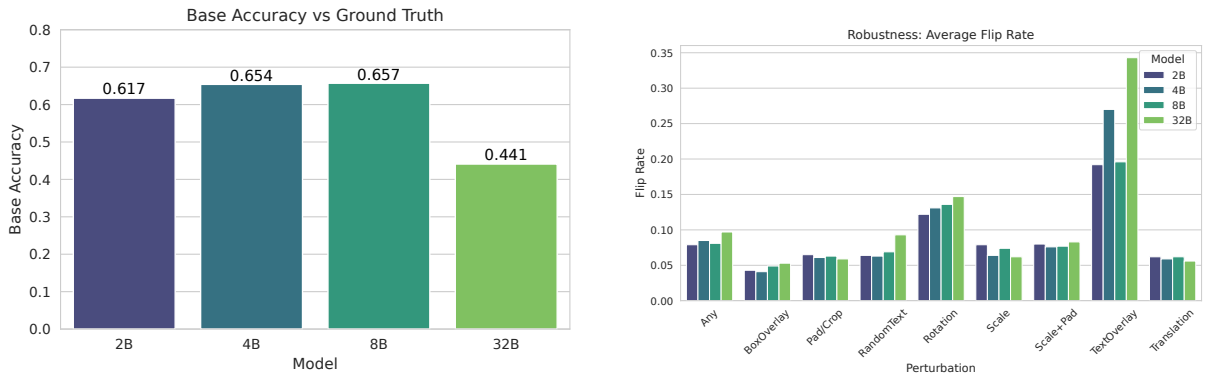

Figure 2: **Qwen3-VL (Instruct) scaling on SEEDBench.** Left: base accuracy versus ground truth. Right: average flip rate under natural perturbations (lower is better).

under natural perturbations (Figure 2). This decoupling is most pronounced for semantic text overlays, where the largest models exhibit the highest instability.

Figure 5 in Appendix C.1 decomposes flip behaviour into correctness transitions. As model size increases, generally, both error injection (R→W) and correction (W→R) rates rise under perturbations. This indicates that larger models develop sharper but more fragile decision boundaries, rather than uniformly improved robustness.

Cross-Dataset Validation on MMMU To assess whether scaling-related robustness trends generalise beyond SEEDBench, we repeat the analysis on the MMMU benchmark, which features multi-disciplinary reasoning and more complex visual–textual dependencies. Despite substantial differences in task structure and difficulty between SEEDBench and MMMU, the qualitative robustness trends remain consistent. All the results are reported in Appendix C.2, in particular see Figure 6 and 7. Accuracy improves with scale, while robustness under natural perturbations fails to improve monotonically. Semantic text overlays again induce the largest instability, and correctness transitions exhibit the same bidirectional behaviour observed on SEEDBench. This consistency suggests that the observed robustness failures are not dataset-specific artifacts.

Cross-Architecture Robustness: LLaVA-OneVision Models To assess whether the observed robustness phenomena are specific to the Qwen3-VL family, we extend our analysis to the LLaVA-OneVision architecture. We evaluate `llava-onevision-qwen2-0.5b` and `llava-onevision-qwen2-7b` using the same perturbation suite and evaluation protocol.

Although LLaVA-OneVision employs a different vision backbone and multimodal fusion strategy, its robustness behaviour closely mirrors that of Qwen3-VL. The results can be found in Appendix C.3; these results demonstrate that robustness failures persist across architectural families.

6 Frequency Dynamics and Spectral Drift

To better understand the mechanism behind the observed robustness failures, we probe the role of frequency content in VLM decision-making. Specifically, we test two competing hypotheses: (H1) *Low-frequency dominance*, suggesting VLMs rely primarily on global shapes and should be robust to high-frequency noise (Yin et al., 2019; Paul and Chen, 2021; Naseer et al., 2021; Gavrikov et al., 2025); (H2) *Cross-frequency sensitivity*, suggesting that decision boundaries depend on the alignment of both low- and high-frequency components. We conduct three controlled experiments on SEEDBench using a frequency-aware evaluation suite: (i) random band-limited noise injection, (ii) progressive frequency ablation, and (iii) frequency-constrained adversarial attacks (PGD) (Madry et al., 2018). Detailed formulations and per-sample results are provided in Appendix F.

Our analysis strongly supports H2 (Cross-frequency sensitivity). First, random noise injection reveals that high-frequency perturbations are just as effective at inducing flips as low-frequency ones (Figure 17 in Appendix), contradicting the low-frequency dominance hypothesis. Second, frequency ablations show that margins degrade smoothly rather than abruptly, indicating that VLMs do not rely on a single "truth" band but require spectral coherence across components (Fig-

ure 19). Third, frequency-constrained PGD attacks achieve high success rates ($> 79\%$) in both low- and high-frequency regimes (Table 8 in Appendix), confirming that adversarial vulnerability is broadband. These findings suggest that "benign" natural perturbations (like rotation or text overlay) induce failures not by destroying semantic content, but by causing *spectral and phase drift* that decouples the visual encoding from the model's reasoning priors.

7 Does Drift Cause Hallucination? (POPE Analysis)

We analyse robustness effects on the POPE (Adversarial) dataset to examine the relationship between visual instability and hallucination. This analysis complements the reasoning benchmarks by focusing on object existence verification.

To obtain the "skeptic prior", we used a blank-image baseline; we found that the model predicts "No" for nearly all samples in the absence of visual input, revealing a strong skeptical prior against object existence.

We found for the samples where the clean model hallucinates object presence (Figure 12 in Appendix), perturbations consistently shift prediction margins toward the negative spectrum (average flip margin ≈ -0.96). This *drift-to-prior* behaviour indicates that hallucinations in this setting arise from fragile visual features rather than from language bias. By degrading these spurious features, perturbations force the model to revert to its skeptical prior, reducing false positives at the cost of recall. Detailed discussion is in Appendix D

In particular, we found that across datasets, model scales, and architectural families, robustness under natural visual perturbations fails to scale monotonically with capacity. Despite substantial gains in clean accuracy, larger models consistently exhibit increased sensitivity to perturbations and higher rates of decision boundary crossings. The persistence of these effects suggests that robustness failures arise from shared design principles in modern VLMs, such as patch-based vision encoders and tightly coupled vision–language reasoning, rather than from idiosyncrasies of individual benchmarks or models.

8 Discussion and Conclusion

We presented a systematic analysis of robustness in vision-language models, examining how natural, spectral, and adversarial perturbations affect both outputs and internal representations across SEED-Bench, MMMU, and POPE. Our findings challenge the assumption that high base accuracy or increased model scale reliably confer robustness.

Across all experiments, a consistent pattern emerges: perturbations induce substantial internal drift—captured by embedding displacement and Dirichlet energy changes—well before any observable prediction flip. This latent instability explains why models may appear output-invariant while becoming increasingly fragile to subsequent perturbations. Importantly, this instability is not task-specific but reflects a shared vulnerability in multimodal alignment, manifesting differently depending on task geometry. In reasoning tasks (SEEDBench, MMMU), high-frequency corruption degrades fine-grained discrimination. In hallucination detection (POPE), the same perturbations can suppress spurious features, reducing false positives while eroding evidence for subtle true objects.

Contrary to expectations, larger models do not exhibit lower flip rates under semantic or phase-altering transformations such as text overlays and rotations. This scale-robustness decoupling suggests that increased capacity sharpens decision boundaries without stabilizing the underlying representations. Text overlays prove particularly disruptive, creating explicit alignment conflicts between visual evidence and linguistic cues that disproportionately affect larger models—a deficiency already highlighted by Li et al. (2023a).

Our frequency-aware framework unifies natural, semantic, and adversarial robustness failures under a single trajectory-level perspective. By interpreting perturbations as inducing spectral and representational drift rather than mere pixel corruption, we reveal that robustness evaluation must extend beyond output accuracy to include representation stability, structural smoothness, and asymmetric error dynamics. We hope these findings motivate robustness evaluations and architectural designs that explicitly account for multimodal alignment stability in vision-language foundation models.

Limitations

Our analysis focuses on a specific set of vision-language models and benchmarks, which may not fully capture the diversity of architectures and task distributions in deployment. While we examine multiple perturbation families, the parameter

ranges and specific transformations were chosen to reflect natural variations rather than exhaustive adversarial exploration—more extreme perturbations or targeted attacks may reveal different failure modes—as originally showcased in Goodfellow et al. (2014).

The frequency-domain analysis provides interpretable insights but relies on specific metrics (Dirichlet energy, spectral norms) that may not capture all aspects of representational drift. Alternative measures of trajectory stability or decision margin erosion could complement our findings. Additionally, our hook points target the final LLM layer, which may not fully characterise the early fusion dynamics within vision encoders or cross-attention mechanisms.

Our task-dependent observations (e.g., high-frequency noise reducing POPE hallucinations) suggest that robustness cannot be evaluated uniformly across all applications. A perturbation that improves calibration in one context may degrade performance in another, complicating the development of universal robustness interventions. Finally, while we document the scale-robustness decoupling, we do not propose architectural modifications or training objectives to address it—identifying effective mitigation strategies remains an open challenge.

References

Sophia J. Abraham, Jonathan D. Hauenstein, and Walter J. Scheirer. 2025. Wavelet-Based Mechanistic Interpretability of Vision Transformers via Frequency-Aware Ablations . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4830–4834, Los Alamitos, CA, USA. IEEE Computer Society.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2024. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Shivam Chandhok, Wan-Cyuan Fan, Vered Shwartz, Vineeth N. Balasubramanian, and Leonid Sigal. 2025. Response wide shut? surprising observations in basic vision language model capabilities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25530–25545, Vienna, Austria. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Routledge.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409.

Samantha Dies, Courtney Maynard, Germans Savcisens, and Tina Eliassi-Rad. 2025. Representational stability of truth in large language models.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 11198–11201, New York, NY, USA. Association for Computing Machinery.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6216–6234. PMLR.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper, and Janis Keuper. 2025. Can we talk models into seeing the world differently? In *The Thirteenth International Conference on Learning Representations*.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. 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)*, pages 14375–14385.

Md Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, Abu Raihan

Mostofa Kamal, and Md. Azam Hossain. 2025. Visual robustness benchmark for visual question answering (vqa). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6623–6633.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: investigating the design space of visually-conditioned language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Reza Khanmohammadi, Erfan Miahi, Mehrsa Mardikoraem, Simerjot Kaur, Ivan Brugere, Charese Smiley, Kundan S Thind, and Mohammad M. Ghassemi. 2025. Calibrating LLM confidence by probing perturbed representation stability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10459–10525, Suzhou, China. Association for Computational Linguistics.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Xiujun Li, Yujie Lu, Zhe Gan, Jianfeng Gao, William Yang Wang, and Yejin Choi. 2023a. Text as images: Can multimodal large language models follow printed instructions in pixels? *arXiv preprint arXiv:2311.17647*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yunqi Liu, Xue Ouyang, and Xiaohui Cui. 2025. Gleam: Enhanced transferable adversarial attacks for vision-language pre-training models via global-local transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1665–1674.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 23296–23308. Curran Associates, Inc.

Yuto Nishida, Masaru Isonuma, and Yusuke Oda. 2025. Instability in downstream task performance during LLM pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22883–22895, Suzhou, China. Association for Computational Linguistics.

Sayak Paul and Pin-Yu Chen. 2021. Vision transformers are robust learners. *AAAI*.

Shengju Qian, Hao Shao, Yi Zhu, Mu Li, and Jiaya Jia. 2021. Blending anti-aliasing into vision transformer. In *Advances in Neural Information Processing Systems*, volume 34, pages 5416–5429. Curran Associates, Inc.

Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. 2021. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. 2025. Latent space chain-of-embedding enables output-free LLM self-evaluation. In *The Thirteenth International Conference on Learning Representations*.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. 2019. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. 2021. Dirichlet energy constrained learning for deep graph neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 21834–21846. Curran Associates, Inc.

A Detailed Experimental Setup

System configuration and reproducibility. All experiments are conducted with a fixed random seed (seed = 0). Unless otherwise specified, evaluations are performed on a single NVIDIA A100 GPU (80GB memory) using FP16 precision. For Qwen3-VL-32B-Instruct, evaluation is distributed across four NVIDIA A100 GPUs (80GB each). Log-likelihood scoring is performed with a batch size of 4.

A.1 Models

We evaluate the following vision language models in a zero-shot setting:

- Qwen3-VL (Instruct): 2B, 4B, 8B, 32B
- LLaVA-OneVision: 0.5B, 7B

All models are evaluated without fine-tuning.

A.2 Datasets and Sample Counts

Dataset	Split	Base Samples	Task
SEEDBench	Full	14000	MCQ
MMMU	Validation	847	Multi-image MCQ
MMMU	Validation	3000	Multi-image MCQ
POPE	Adversarial	3000	Yes/No

Table 5: Datasets and evaluation splits used in the paper.

Each base sample is evaluated under multiple perturbation instances A.6.

A.3 Prompt Templates (Prompt Regimes)

We use chat-style multimodal prompts where images are embedded in the user message content. The textual instruction depends on the representation probes regime A.7. In all regimes, the image(s) are provided as part of the user content, followed by the text prompt.

`ans_mcq` regime This regime explicitly lists the options and instructs the model to select from them:

```
User:
  [IMAGE]
  Question: <question>
  Options:
  A. <option A>
  B. <option B>
  C. <option C>
  D. <option D>
  Please select the correct answer from
  the options above.
Assistant:
  <option text>
```

`ans_open` regime This regime does not surface the options and requests a concise answer:

```
User:
  [IMAGE]
  Question: <question>
  Please answer the question concisely.
```

`ans_mcq_free` regime This regime provides the options for context but requests a free-form answer:

```
User:
  [IMAGE]
  Question: <question>
  Options (for your reference; answer freely):
  A. <option A>
  B. <option B>
  C. <option C>
  D. <option D>
  Provide the best answer in your own words.
```

Multi-image MMMU formatting. For MMMU, questions may include markers such as `<image 1>`, `<image 2>`, etc. We interleave images into the user message at the referenced positions; any remaining images are appended afterward. The instruction line is still determined by the prompt regime above.

POPE (Yes/No). For POPE, the user message contains the image and a binary question. The assistant response is restricted to a single token (`Yes` or `No`).

```
User:
  [IMAGE]
  Question: <question>
  Answer with exactly one word: Yes or No.

Assistant:
  Yes / No
```

A.4 Visual Perturbation Examples

We ensure that all applied perturbations remain semantically benign, meaning a human observer would still easily recognise the original content. As shown in Figure 3, while operations like rotation ($-30°$) or padding significantly shift pixel distributions, the core visual evidence required for reasoning remains intact.

A.5 Data Pre-processing

We utilise the official evaluation splits for SEEDBench (Image), MMMU (Validation), and POPE (Adversarial). For MMMU, we process multi-image examples by applying the perturbation v' to all image components $\{I_k\}$ within the sequence x simultaneously.

A.6 Perturbation Hyperparameters

We define the perturbation families with the following parameter sweeps:

Figure 3: Visual perturbation examples applied to a SEEDBench sample used in our robustness evaluation. The figure displays the **original base image (Top-Left)** alongside three geometric transformations: **translation (Top-Right)**, **padding and cropping (Bottom-Left)**, and $-30°$ **rotation (Bottom-Right)**. These perturbations serve to test the model's structural consistency under spatial variation without altering semantic content.

- **Translation (cyclic):** horizontal wrap-around (cyclic) shifts by $\Delta x \in \{-16, -12, \ldots, 16\} \setminus \{0\}$ pixels.
- **Pad/Crop:** symmetric padding or cropping by $n \in \{-16, -12, \ldots, 16\} \setminus \{0\}$ pixels.
- **Scale:** rescaling by a factor α (default $\alpha = 0.9$) followed by resizing back to the original resolution.
- **Scale+Pad:** rescaling followed by padding with a uniform background.
- **Rotation:** in-plane rotation by $\pm 30°$ with interpolation.
- **Text Overlays (three variants).** To disentangle occlusion and edge injection from *semantic steering*, we evaluate three overlay families with identical geometry but different content:
  - **Semantic overlay (`TextOverlay`)**: short directive phrases such as "Answer is A/B/C/D" rendered on the image.
  - **Random-text overlay (`RandomText`)**: same-sized text region filled with random strings of comparable length and ink density.
  - **Empty-box overlay (`BoxOverlay`)**: the same box region rendered without text, controlling for occlusion and shape.

Each perturbation type is applied multiple times per image, producing a set of perturbed variants $x' \in v'(x)$ for each base image x, here v' is perturbation type .

Unless explicitly stated otherwise, `TextOverlay` refers to the semantic overlay variant. We report `RandomText` and `BoxOverlay` separately when analyzing overlay-specific effects.

A.7 Representation Probes

We extract hidden states $h \in \mathbb{R}^d$ from the final layer of the LLM backbone at five specific hook points:

- **`ctx_open`:** The last token of the visual context under an open-ended prompt.
- **`ctx_mcq`:** The last token of the visual context under the specific MCQ prompt template.
- **`ans_open`:** The mean-pooled embedding of the generated answer tokens (greedy decoding).
- **`ans_mcq`:** The mean-pooled embedding of the generated answer conditioned by MCQ.
- **`ans_mcq_free`:** The mean-pooled embedding of the generated answer when the model is forced to answer freely, constrained to the option set.

A.8 Control Drift Baseline

To normalise embedding distances, we compute Cohen's d between the perturbation drift distribution $D_{pert} = \{\|e(x) - e(x')\|\}$ and a control distribution $D_{ctrl} = \{\|e(x_i) - e(x_j)\|\}$ derived from 1,000 random pairs of unrelated images Table 6.

B Drift versus Control Drift, L2 results

C Model Scaling

C.1 SeedBench Dataset

Here we report some additional results on model scaling for SeedBench dataset. Figure 5 shows that as model size increases, both error injection

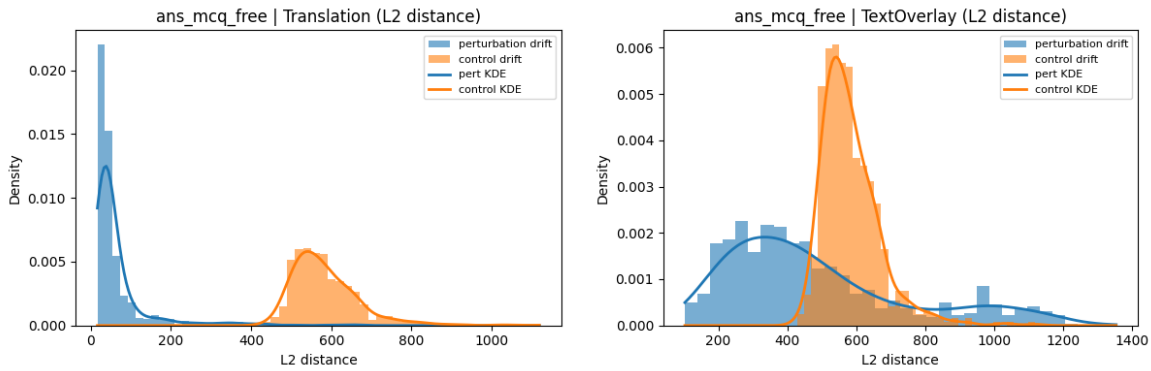

Figure 4: L2 distance, Drift versus control drift for the `ans_mcq_free` embedding under Translation and text overlay perturbations. Blue shows perturbation-induced drift relative to the base image; orange shows control drift (base image versus randomly sampled other images). Left: Translation. Right: Textoverlay. Unlike geometric perturbations, the perturbation-induced distribution does not remain well separated from control drift, indicating that the representation no longer stays local in embedding space.

(R→W) and correction (W→R) rates rise, indicating larger models develop sharper but more fragile decision boundaries rather than uniformly improved robustness.

C.2 Cross-Dataset Validation on MMMU

To assess whether scaling-related robustness trends generalise beyond SEEDBench, we repeat the analysis on the MMMU benchmark, which features multi-discipline reasoning and more complex visual–textual dependencies. Despite substantial differences in task structure and difficulty between SEEDBench and MMMU, the qualitative robustness trends remain consistent (Figure 6 and 7).

Perturbation	μ_{pert}	σ_{pert}	μ_{ctrl}	Cohen's d
Translation	64.42	78.78	584.19	−6.3522
Pad/Crop	70.98	84.18	584.19	−6.0757
Scale	77.47	104.29	584.19	−5.3325
Scale+Pad	91.12	103.06	584.19	−5.2256
Rotation	140.57	98.37	584.19	−4.8316
TextOverlay	483.77	266.40	584.19	−0.5080

Table 6: Drift versus control drift for the `ans_mcq_free` embedding measured using L2 distance. TextOverlay exhibits substantially reduced separation from control drift, consistent with non-local displacement in representation space.

C.3 Cross-Architecture Robustness: LLaVA-OneVision Models

To assess whether the observed robustness phenomena are specific to the Qwen3-VL family, we extend our analysis to the LLaVA-OneVision architecture. We evaluate `llava-onevision-qwen2-0.5b` and `llava-onevision-qwen2-7b` using the same perturbation suite and evaluation protocol.

Although LLaVA-OneVision employs a different vision backbone and multimodal fusion strategy, its robustness behaviour closely mirrors that of Qwen3-VL (Figure 8 and 9). Accuracy improves with scale, while robustness remains comparable or degrades, particularly under semantic text overlays. Correctness transitions again show increased error injection and correction at larger scales, indicating sharper yet more fragile decision boundaries. These results demonstrate that robustness failures persist across architectural families.

D Semantic Robustness and Error Asymmetry on POPE

We analyze robustness on POPE (Adversarial) to examine the relationship between visual instability and hallucination. Unlike multiple-choice reasoning tasks, POPE poses a binary decision problem with a highly asymmetric label space. In such settings, representation drift does not induce random label switching. Instead, misaligned or unfamiliar visual representations tend to collapse predictions toward the model's language prior, which in POPE is strongly biased toward negative ("No") responses. This allows us to study how the same underlying representational instability manifests differently depending on task geometry.

Having established that representation drift can both degrade reasoning performance and, in some cases, suppress hallucinations, we now provide a deeper analysis of *semantic robustness* using the POPE benchmark. Unlike SEEDBench and

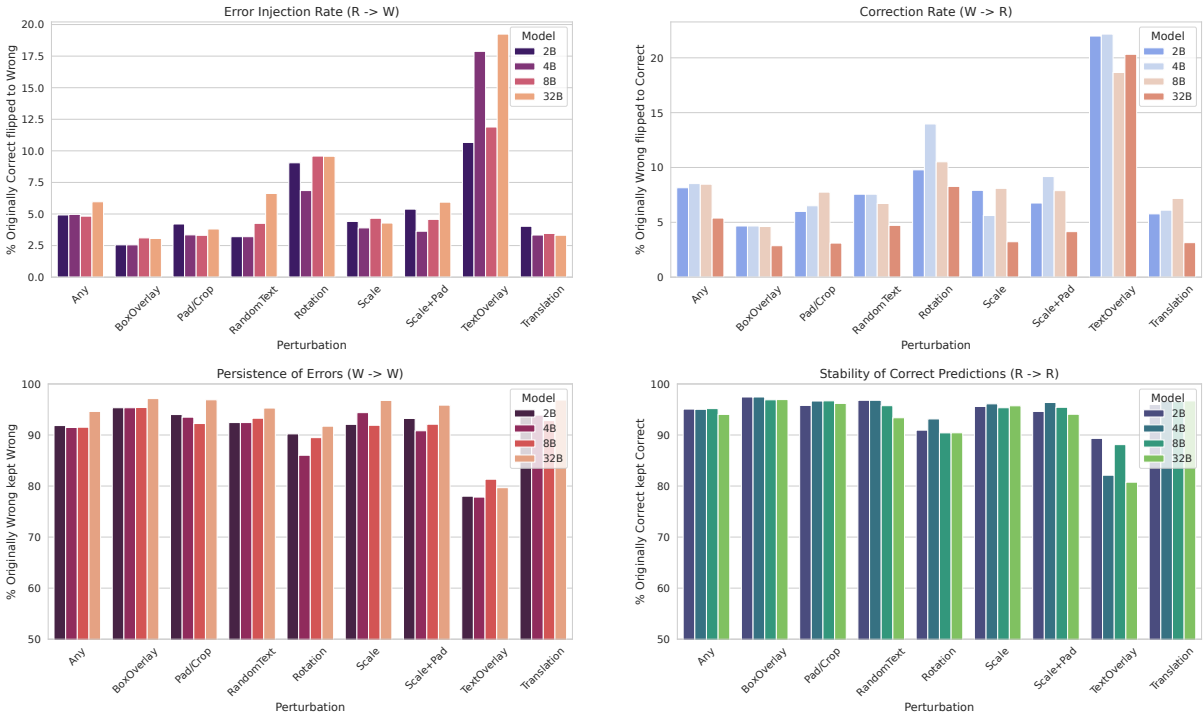

Figure 5: Correctness transition statistics under natural perturbations for different Qwen3-VL model scales on SEEDBench.

MMMU, POPE focuses on object existence verification and enables fine-grained analysis of asymmetric error dynamics under perturbations.

All experiments in this section are conducted on the adversarial split of POPE using `Qwen3-VL-2B` and `Qwen3-VL-8B` Instruct models. Perturbations are applied exclusively to the image, while the semantic content of the query remains unchanged.

D.1 Overall Semantic Stability Under Perturbations

Figure 10 reports base accuracy and average flip rates across perturbation types. Although both models achieve high base accuracy on clean images, natural perturbations induce substantial semantic instability. Rotation consistently emerges as the most disruptive transformation, followed by text overlays and scale-related perturbations.

Notably, increased model capacity does not guarantee improved semantic robustness: the 8B model exhibits comparable or higher flip rates than the 2B model for several perturbations, reinforcing the accuracy–robustness decoupling observed earlier.

D.2 Asymmetric Error Dynamics

A key strength of POPE is its ability to distinguish asymmetric semantic errors. We analyse four complementary transition types: (i) true positives flipping to false negatives (TP $\rightarrow$ FN), (ii) true negatives flipping to false positives (TN $\rightarrow$ FP), (iii) correction of hallucinations (FP $\rightarrow$ TN), and (iv) recovery of missed detections (FN $\rightarrow$ TP).

Figure 12 reveals pronounced asymmetry. Perturbations such as rotation and scale predominantly increase TP $\rightarrow$ FN errors, indicating that correct affirmative detections are fragile under visual transformations. In contrast, text overlays and crop-based perturbations more frequently induce TN $\rightarrow$ FP transitions, increasing hallucination rates.

Importantly, some perturbations also *reduce* hallucinations: non-zero FP $\rightarrow$ TN correction rates indicate that perturbations can accidentally suppress false positives, consistent with the drift-to-prior behaviour identified earlier.

D.3 Representation Drift and Smoothness Under Semantic Perturbations

To connect semantic instability with internal behaviour, we measure embedding drift and changes in Dirichlet energy under perturbations. Figure 11 shows that perturbations inducing higher semantic flip rates also exhibit larger embedding displacement and greater changes in the token smoothness.

Rotation induces the largest embedding drift and the strongest decrease in Dirichlet energy, aligning with its dominant impact on TP $\rightarrow$ FN errors.

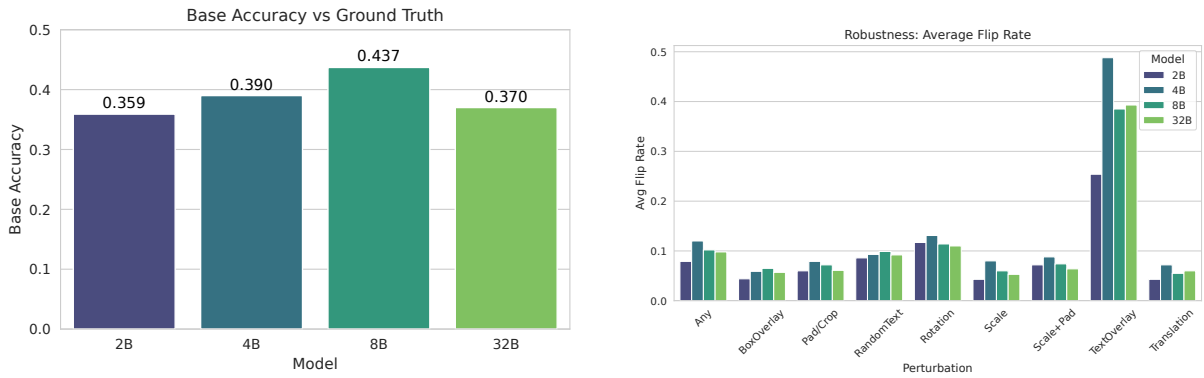

Figure 6: **Qwen3-VL (Instruct) scaling on MMMU.** Left: base accuracy versus ground truth. Right: average flip rate under natural perturbations.

Text overlays, while visually localised, produce consistent representation shifts, explaining their disproportionate effect on hallucination dynamics.

These results extend our earlier analysis by showing that semantic robustness failures are tightly coupled to both global representation drift and local structural reorganization, even when high-level image semantics appear preserved.

E Vision-Token Smoothness and Dirichlet Energy

Embedding drift captures *global* movement in representation space, but it does not directly characterise how visual information is organised *locally* within the vision encoder. To study structural stability at the token level, we analyse the *Dirichlet energy* of vision tokens arranged on their spatial grid.

This provides a complementary diagnostic of robustness: embedding drift measures *where* representations move in latent space, whereas Dirichlet energy measures *how* visual features are spatially structured and smoothed across neighbouring tokens. Together, these views allow us to distinguish global representation shifts from local structural reorganization.

E.1 Dirichlet Energy on Vision Tokens

Following the general definition introduced in Section 3, we instantiate Dirichlet energy on the spatial grid of vision tokens to quantify local structural organization within the vision encoder. Dirichlet energy quantifies spatial smoothness, low values correspond to locally consistent token representations, while high values indicate sharp spatial variation or token misalignment. For each perturbation instance, we compute the change in Dirichlet energy

Perturbation	ΔE_{dir} (all)	ΔE_{dir} (flips)
Translation	10.34 ± 67.49	15.26 ± 68.56
Pad/Crop	8.65 ± 70.13	11.76 ± 71.31
Scale	20.70 ± 82.47	21.02 ± 88.59
Scale+Pad	9.86 ± 82.29	14.38 ± 80.81
TextOverlay	-33.87 ± 60.14	-34.27 ± 61.45
Rotation	-72.73 ± 99.95	-66.76 ± 101.21

Table 7: Dirichlet energy change ΔE_{dir} (mean $\pm$ std).

$$\Delta E_{\text{dir}} = E_{\text{dir}}(x') - E_{\text{dir}}(x), \tag{6}$$

where x is the base image and x' its perturbed version.

E.2 Dirichlet Energy Distributions

Figures 13 and 14 show the distribution of ΔE_{dir} for Translation and TextOverlay perturbations, respectively. For each perturbation, we report distributions over all instances and over the subset of instances that induce label flips. (Combined visualizations across perturbation types are provided in the Appendix.)

Translation exhibits a small positive mean shift in ΔE_{dir} (10.34 ± 67.49), accompanied by substantial variance. This indicates that translations do not uniformly smooth or disrupt spatial structure, but instead introduce heterogeneous local misalignment. Flip-inducing translation instances show slightly larger positive shifts, consistent with phase-driven token misalignment rather than large-magnitude distortions. In contrast, TextOverlay produces a qualitatively different pattern. The mean ΔE_{dir} is negative (-33.87 ± 60.14), indicating a systematic reorganization of local token neighbourhoods. This shift occurs for both flip

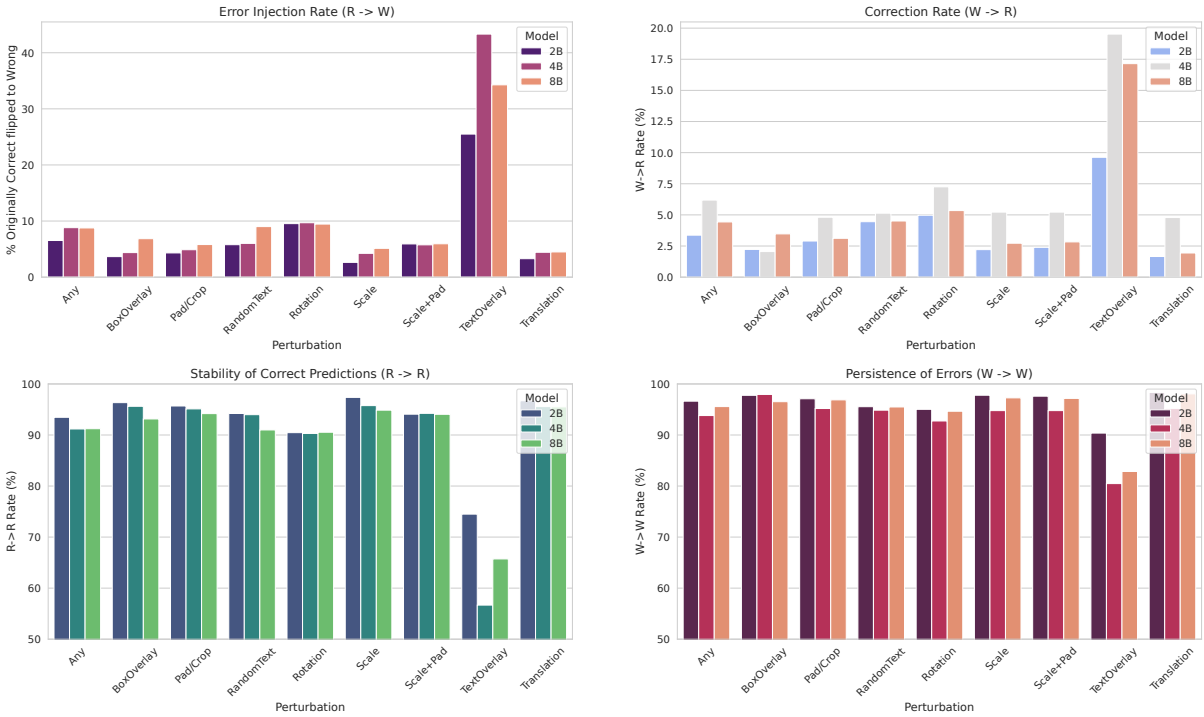

Figure 7: Correctness transition statistics under natural perturbations for different Qwen3-VL model scales on MMMU.

and non-flip instances, suggesting that overlays alter spatial structure even when the final prediction remains unchanged.

Unlike geometric perturbations, which primarily redistribute spatial alignment, text overlays introduce sharp edges and high-contrast strokes that interact directly with the patch grid. The resulting effect is not simple displacement but a restructuring of local token relationships.

E.3 Dirichlet Energy and Decision Instability

Table 7 summarises mean ΔE_{dir} across all perturbation types, reported separately for all instances and for flip-inducing instances. Across perturbation families, flip-inducing instances are associated with larger absolute Dirichlet deviations, indicating more pronounced spatial reorganization of vision tokens, rather than providing a standalone predictor of failure.

Dirichlet deviations than non-flip instances. This indicates that perturbations that substantially reorganise local token structure, either sharpening or smoothing relative to the base, are more likely to trigger decision boundary crossings.

Specifically, Dirichlet energy does not serve as a universal failure threshold; rather, it functions as a proxy for structural reorganization. Label flips are associated with significant changes in the spatial relationships between vision tokens, even when global embedding drift remains moderate.

Many natural perturbations are not defined in the frequency domain, yet they systematically alter frequency content or phase relationships after interpolation, resampling, and discretization. Thus we also examined correlations between Dirichlet energy change, embedding drift, and frequency-band energy shifts. Across perturbation types, correlations between ΔE_{dir} and embedding drift are modest but consistently non-zero, and are stronger for flip-inducing instances. For example for flip inducing instances:

- Translation: $\text{corr}(\Delta E_{\text{dir}}, \text{drift}) = 0.057$
- Scale: $\text{corr}(\Delta E_{\text{dir}}, \text{drift}) = 0.149$
- Rotation: $\text{corr}(\Delta E_{\text{dir}}, \text{drift}) = 0.065$

Importantly, Dirichlet energy is not intended as a superior predictor of failures compared to embedding drift. Rather, it provides complementary structural evidence that global drift is accompanied by localised reorganization of visual tokens, which is invisible to pooled embedding metrics. While these correlations do not imply direct causality, they indicate that structural token reorganization and global representation drift are related but distinct phenomena. Perturbation-specific frequency

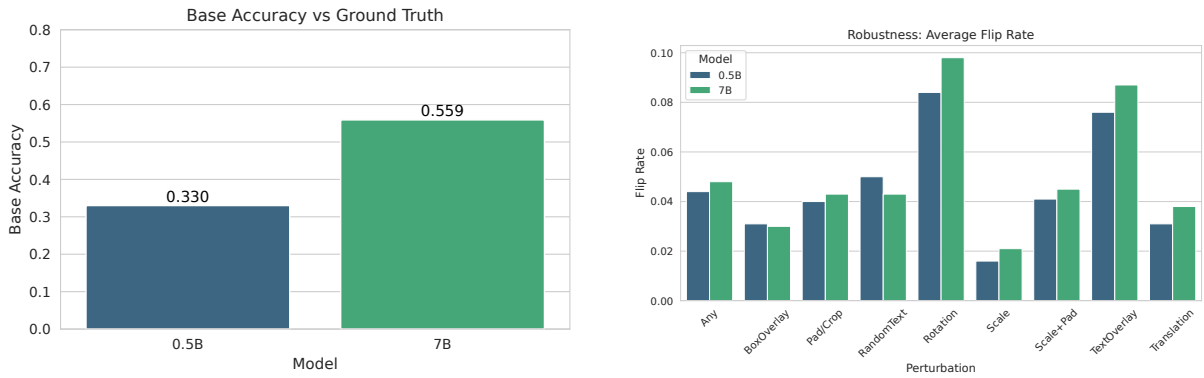

Figure 8: **LLaVA-OneVision scaling on SEEDBench.** left: base accuracy versus ground truth. Right: average flip rate under natural perturbations.

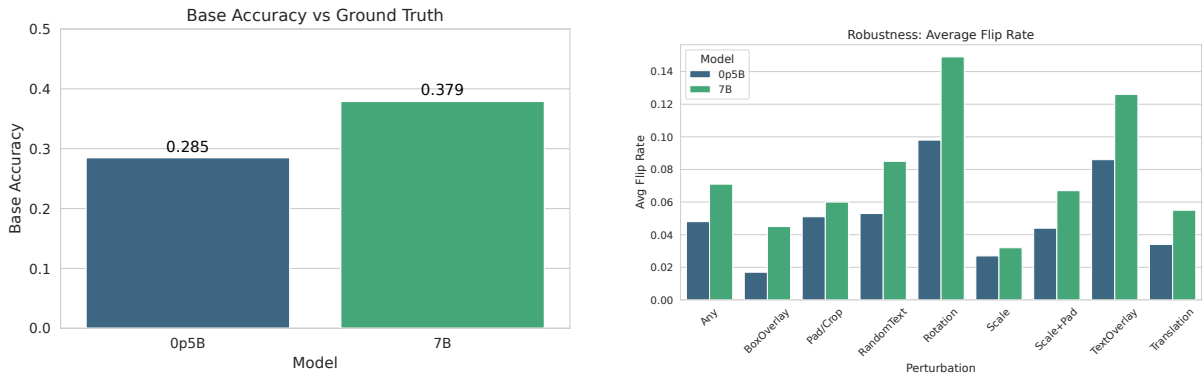

Figure 9: **LLaVA-OneVision scaling on MMMU.** Left: base accuracy versus ground truth. Right: average flip rate under natural perturbations.

(Figure 15) effects also align with Dirichlet trends, interpolation-based transformations amplify low-frequency components, while text overlays inject broadband energy, consistent with both observed Dirichlet shifts and embedding drift.

E.4 Structural Drift Complements Representation Drift

Dirichlet energy provides a complementary diagnostic to embedding-based analyses:

- Embedding drift captures *where* representations move in latent space.
- Dirichlet energy captures *how* spatial structure within vision tokens is reorganised.

Together, these results reinforce our central thesis: robustness failures in VLMs arise from *structural and spectral drift of vision tokens* that misalign visual representations with language-conditioned decision boundaries. This drift can accumulate even when output predictions remain unchanged, exposing a hidden reduction in stability and decision margin.

F Detailed Frequency-Aware Robustness Suite

This appendix details the experimental setup and full results for the frequency analysis summarised in Section 6. This section is not intended as an independent robustness benchmark. Rather, it serves as a controlled hypothesis test of the spectral-drift interpretation motivated by the natural perturbation and Dirichlet analyses (Sections 4). We move beyond geometric transformations and explicitly probe the role of *frequency content* in VLM robustness.

Our objective is to disentangle two competing hypotheses:

- **H1 (Low-frequency dominance).** VLM decisions rely primarily on low-frequency visual content; high-frequency perturbations should therefore have limited impact (Paul and Chen, 2021; Naseer et al., 2021; Gavrikov et al., 2025).
- **H2 (Cross-frequency sensitivity and interaction).** Both low- and high-frequency

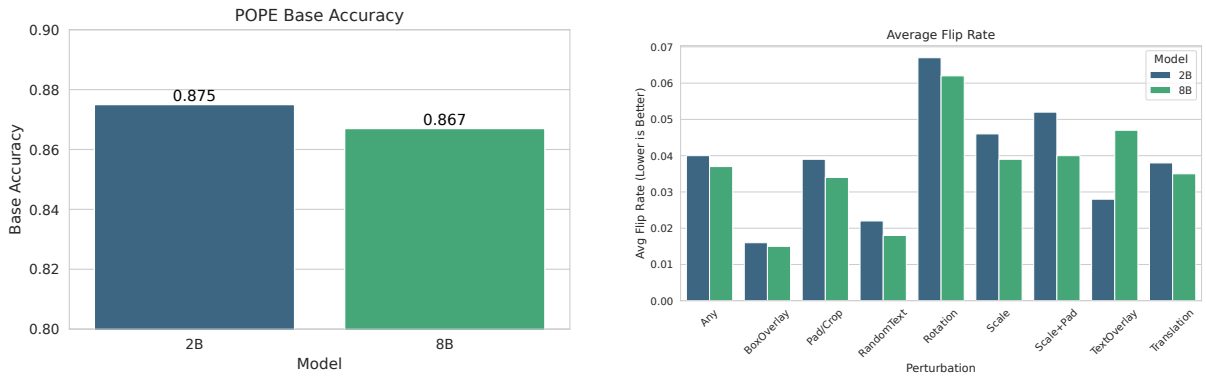

Figure 10: POPE adversarial split. **Left**: base accuracy (higher is better). **Right**: average flip rate under perturbations (lower is better). Semantic robustness does not improve monotonically with model scale.

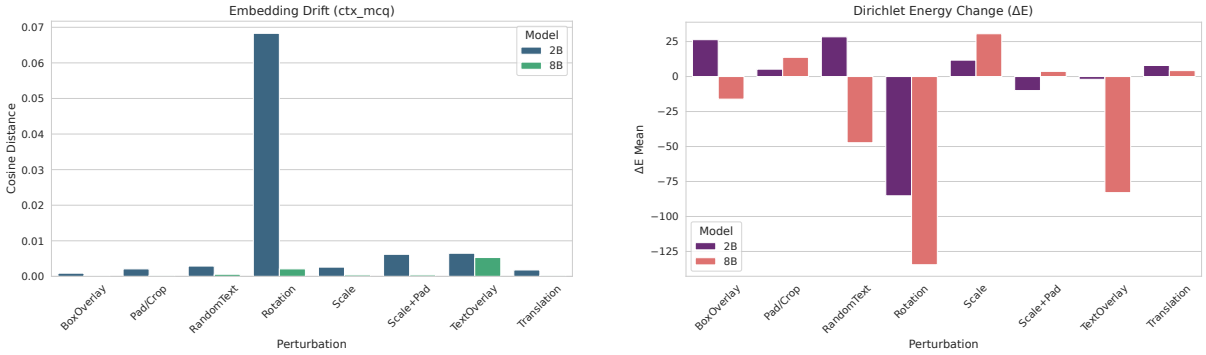

Figure 11: Representation-level effects on POPE. **Left**: embedding drift. **Right**: change in Dirichlet energy. Semantic instability aligns with representation drift and structural disruption.

components contribute to decision-making, and robustness failures arise from *frequency-space drift* that misalign vision tokens with language-conditioned decision boundaries.

Unless stated otherwise, experiments in this section are conducted on a representative model (`Qwen3-VL-2B`) and aggregated over 400 **SEED-Bench** samples.

F.1 Random Band-Limited Noise

Given an image x, we sample i.i.d. Gaussian noise η and project it into a frequency subspace using a radial Fourier mask: low-pass (L), high-pass (H), or full-spectrum (A). We construct perturbed inputs

$$x' = \mathrm{clip}\big(x + \epsilon \cdot \eta_{\text{band}}\big), \tag{7}$$

where $\epsilon \in \{1/255, 2/255, 4/255, 8/255, 16/255\}$. For each ϵ, multiple trials are performed per image and statistics are averaged across all samples.

Figure 17 reports mean flip rates as a function of ϵ. Flip rates remain high (approximately 50%) across all frequency bands, including low-frequency-only noise. This directly contradicts a purely low-frequency dominance explanation and indicates sensitivity across the spectrum.

Figure 16 shows per-sample flip behaviour and corresponding margin evolution. Even when predictions do not flip, increasing ϵ consistently erodes classification margins, revealing representational instability that precedes output-level failure.

F.2 Frequency Ablation (Reliance Test)

To assess frequency reliance directly, we perform controlled frequency ablations. For a cutoff $c \in (0, 1)$, we define a radial Fourier mask M_c and construct:

- **Low-pass keep:** retain frequencies below c (remove high frequencies),
- **High-pass keep:** retain frequencies above c (remove low frequencies).

For each image, we record the smallest cutoff c at which the predicted answer flips, yielding per-sample flip thresholds.

Figure 18 shows threshold distributions for low-pass and high-pass ablations. Low-pass ablation induces flips at smaller cutoffs, indicating sensitivity

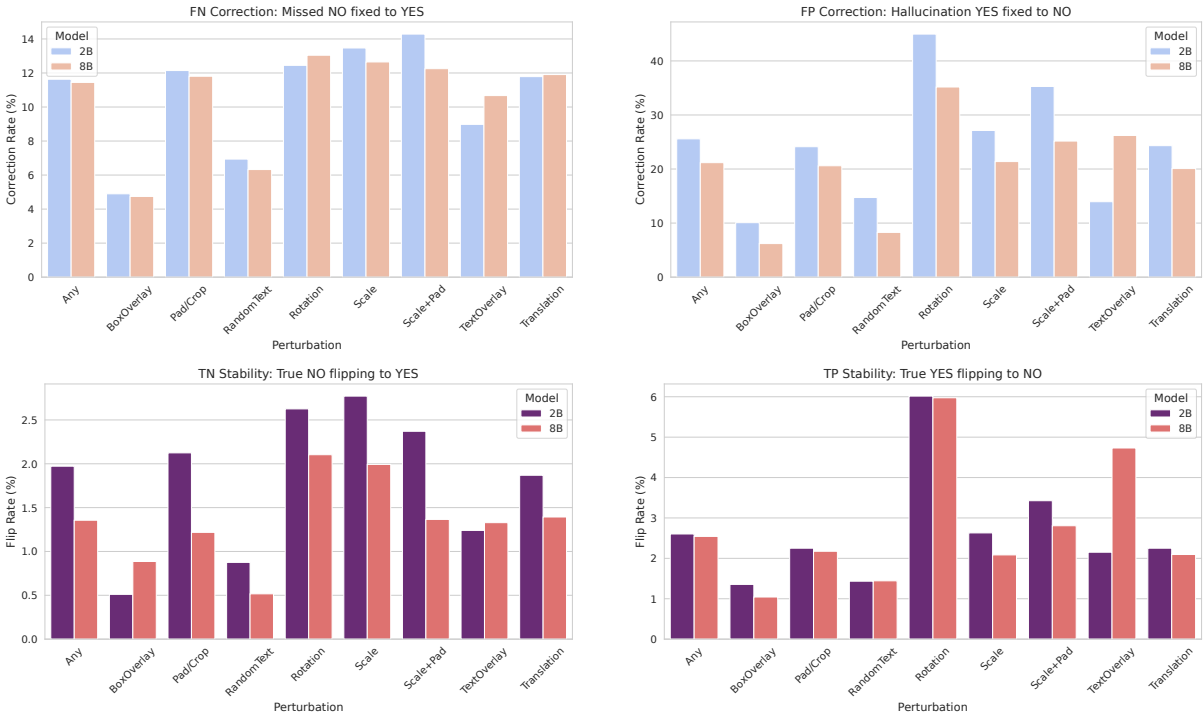

Figure 12: Asymmetric semantic error dynamics on POPE under natural perturbations. Top: instability of correct predictions. Bottom: perturbation-induced correction of semantic errors.

to high-frequency removal. However, high-pass ablation also produces failures across a broad range of cutoffs, demonstrating that low-frequency content alone is insufficient for robust decision-making.

Figure 19 visualises margin evolution for representative samples. Across both ablation types, margins degrade smoothly well before label flips occur, indicating gradual confidence erosion rather than abrupt single-band failure.

F.3 Frequency-Constrained Adversarial Attacks

We extend the analysis to adversarial perturbations using projected gradient descent (PGD) under an ℓ_∞ constraint. At each iteration, perturbations are projected into a target frequency band:

$$\delta_{t+1} = \Pi_{\|\delta\|_\infty \le \epsilon}\Big(\mathrm{Proj}_{\mathrm{band}}\big(\delta_t + \alpha \cdot \mathrm{sign}(\nabla_\delta \mathcal{L}(x+\delta_t))\big)\Big) \tag{8}$$

following (Madry et al., 2018).

Table 8 reports attack success rates and prediction flip rates for pixel-space, low-frequency, and high-frequency PGD attacks. All three modes achieve high effectiveness, confirming that adversarial vulnerability is not confined to a single frequency band.

Figure 20 visualises the frequency-domain structure of the resulting perturbations. Radial energy profiles confirm that frequency-constrained PGD successfully isolates distinct spectral bands while remaining effective.

F.4 Natural Perturbations as Induced Spectral Drift

Taken together, these controlled experiments support a unifying interpretation: many perturbations commonly regarded as meaning preserving induce robustness failures not by corrupting semantics, but by causing *spectral and phase drift*.

Translations primarily alter phase and interact with patch discretization; crop and scale operations induce resampling and aliasing; text overlays inject broadband, high-contrast edges. The frequency-aware suite reproduces the same failure patterns observed under natural perturbations, providing direct evidence that robustness failures cannot be attributed to single-band dominance, but instead arise from sensitivity across frequency components and phase-induced drift.

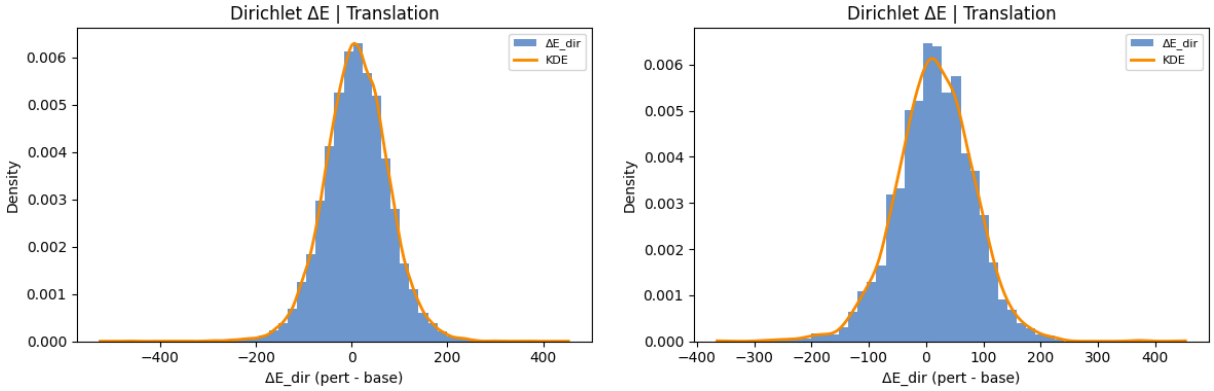

Figure 13: Dirichlet energy change ΔE_{dir} under Translation. Left: all perturbation instances. Right: instances that induce label flips.

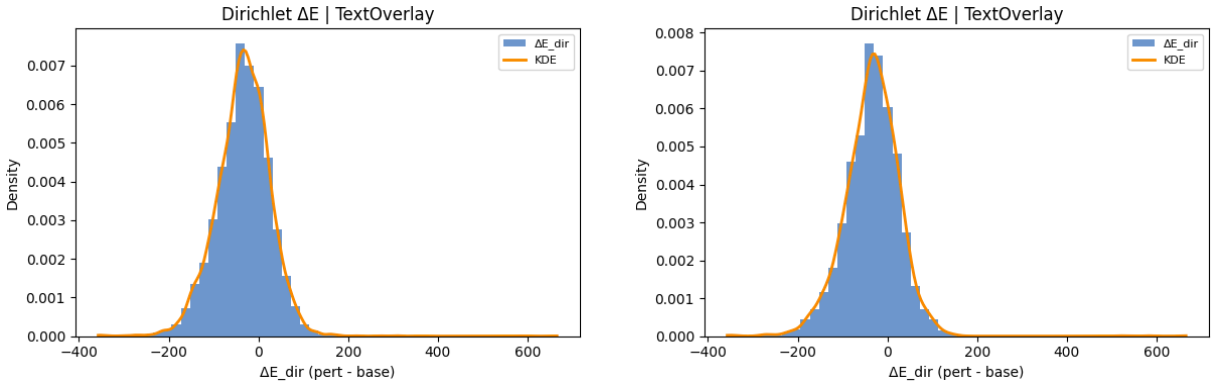

Figure 14: Dirichlet energy change ΔE_{dir} under TextOverlay. Left: all perturbation instances. Right: instances that induce label flips.

Attack mode	ASR ↑	PFR ↑
Pixel PGD	0.8125	0.9525
Low-frequency PGD	0.7900	0.9550
High-frequency PGD	0.8225	0.9475

Table 8: Dataset-level adversarial robustness under frequency-constrained PGD ($\epsilon = 8/255$, 400 samples). Both frequency bands support effective attacks.

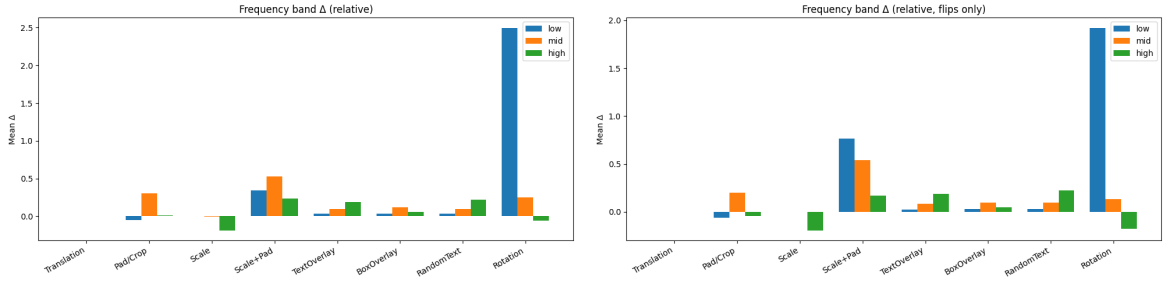

Figure 15: Frequency-band drift induced by perturbation variants. **Left:** Average over all perturbations instances. **Right:** average over flip-inducing perturbation instances only.

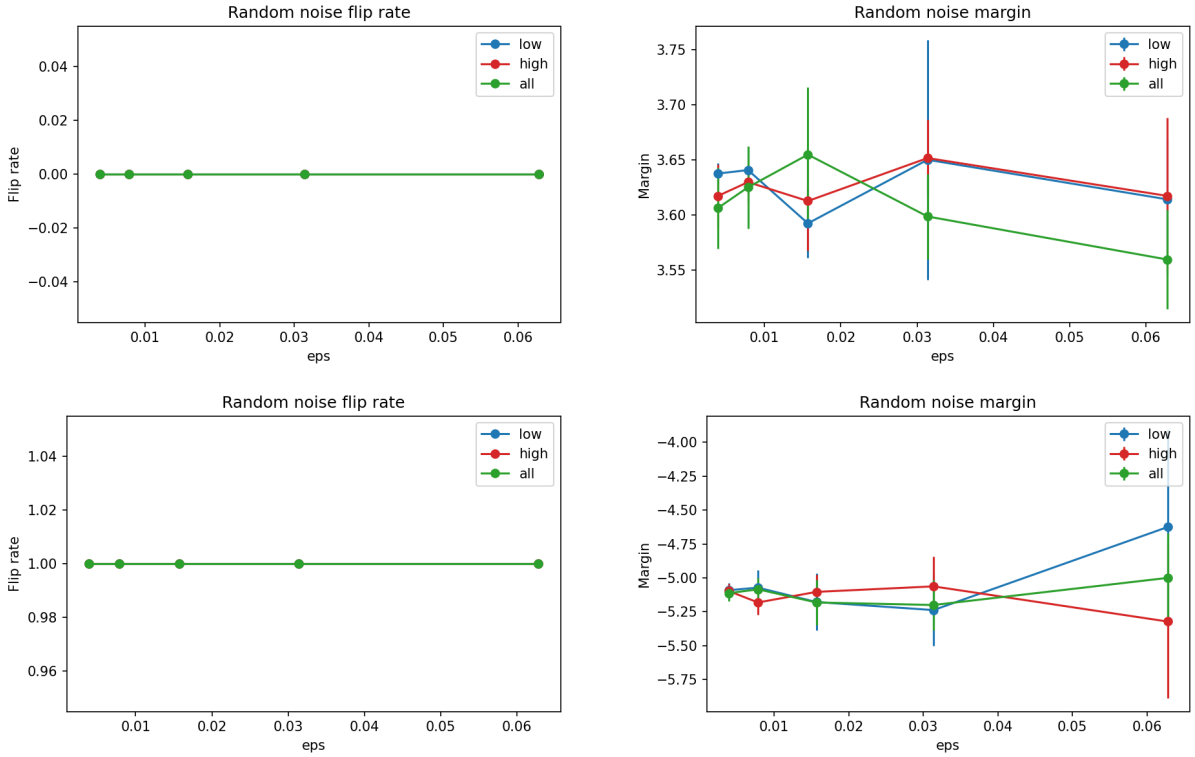

Figure 16: **Per-sample behaviour under random band-limited noise.** Each row corresponds to one SEEDBench sample. **Left column:** flip rate as a function of noise magnitude ϵ. **Right column:** corresponding margin evolution (base option minus strongest competitor). Across both samples, margins degrade steadily with increasing ϵ even when predictions do not immediately flip, revealing hidden representational instability that precedes output-level failures.

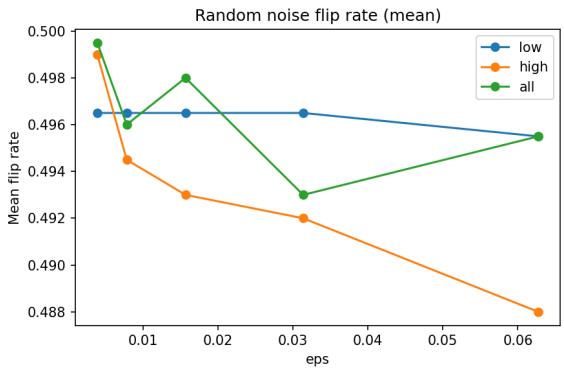

Figure 17: Mean flip rate under random band-limited noise, averaged over 400 samples. Low-frequency (L), high-frequency (H), and full-spectrum (A) noise all induce comparable instability, rejecting a purely low-frequency bias explanation.

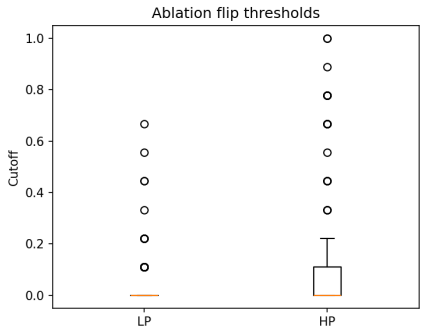

Figure 18: Distributions of flip thresholds under frequency ablation. Both low-pass and high-pass removal induce failures, supporting a cross-frequency reliance rather than a single-band bias.

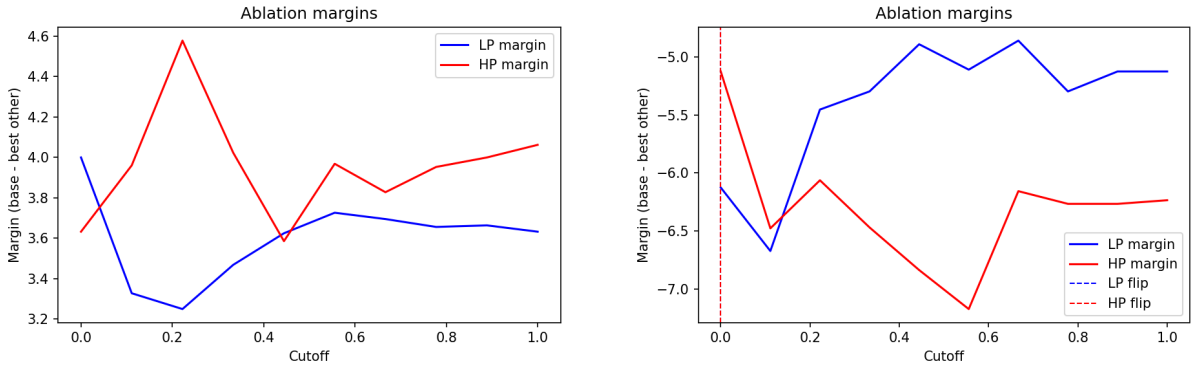

Figure 19: Frequency ablation margins for two representative samples. For each sample, we report the classification margin (base option minus strongest alternative) as a function of the frequency cutoff under low-pass keep (blue) and high-pass keep (red). In both cases, margins degrade substantially before any prediction flip occurs, revealing confidence erosion and representational drift induced by frequency removal.

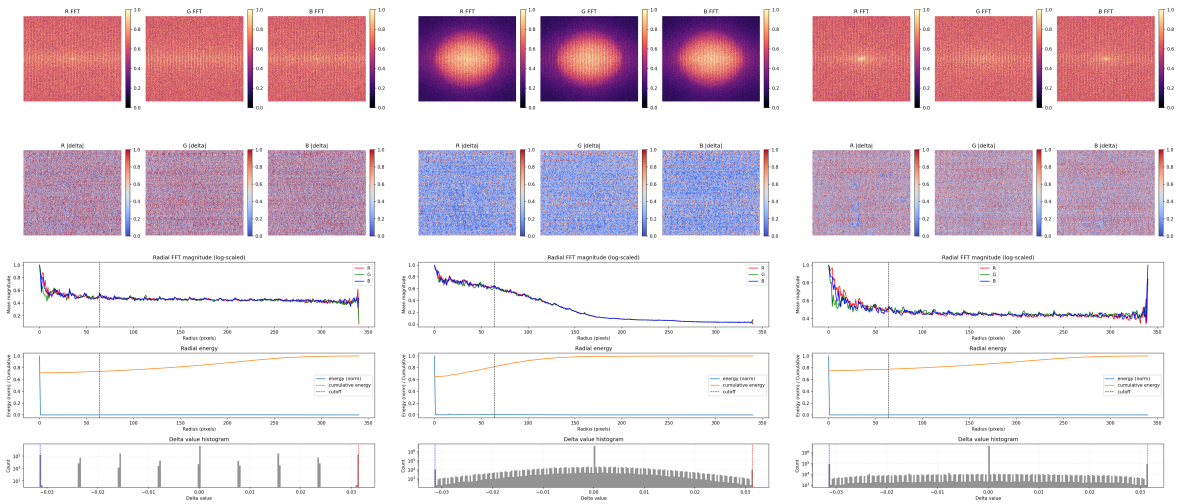

Figure 20: Frequency-domain structure of PGD perturbations. **Left**: unconstrained pixel-space PGD. **Middle**: low-frequency PGD. **Right**: high-frequency PGD. Radial energy profiles and masks confirm effective frequency isolation.