DIAGNOSING VISUAL IGNORANCE IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) achieve high accuracy on many visual question-answering benchmarks, yet it remains unclear whether this accuracy reflects reliable use of the visual details a question depends on. We examine this tension through the lens of visual ignorance: cases where a model can represent task-relevant visual evidence without using it to determine the answer. Across three VLMs and twelve benchmarks, we track the answer-relevant information decodable at each decoder layer and measure when it becomes coupled to the final output. The visual evidence needed for the correct answer can already be read out from intermediate layers, yet it has little effect on the output until the final layers of the decoder. Causal residual interchange confirms that shifting the answer requires the late states, not the earlier ones. We then evaluate all twelve benchmarks under progressive Gaussian blur and find that substantial fractions of predictions survive the entire degradation sequence while continuing to receive benchmark credit. These results distinguish the availability of visual information from its influence on prediction, and show that standard accuracy can coexist with limited sensitivity to the tested visual detail. They motivate training and evaluation that make answer-relevant visual distinctions necessary to answer correctly.

1 INTRODUCTION

Vision-language models (VLMs) combine visual encoders with language decoders pretrained on large text corpora (Liu et al., 2024a; Dai et al., 2023; Bai et al., 2025; Luo et al., 2025). They can recognize familiar image content while overlooking details that determine the answer, particularly when those details conflict with learned expectations (Lin et al., 2024; Lee et al., 2025; Deng et al., 2025; Vo et al., 2025; Asadi et al., 2026). This phenomenon, called visual ignorance, has been documented behaviorally, but its internal basis and its consequences for evaluation remain unclear. We examine where task-relevant visual evidence becomes associated with and influences answer revision, and whether standard VQA benchmarks penalize insensitivity to that evidence.

We study these questions from two directions. Internally, we use matched image pairs in which a visual edit changes the correct answer while the question stays the same. Across Qwen2.5-VL-3B-Instruct (Qwen-3B) (Bai et al., 2025), Qwen2.5-VL-7B-Instruct (Qwen-7B), and LLaVA-v1.6-Mistral-7B (LLaVA-7B) (Liu et al., 2024a), we ask whether the internal update induced by the edit at each layer predicts the answer in the output. We find that in many pairs a trained probe can recover the edited answer from intermediate-layer states, yet these intermediate representations are only weakly coupled to the output; the association becomes strong only in the final layers of the decoder, where late readouts often favor the answer expected from learned priors rather than the edited visual evidence. In Qwen-3B, causal residual interchange confirms that it is the late states, not the earlier ones, that move the answer margin, and that within the sharpest transition the effect is attributable mainly to attention. The coupling pattern recurs in all three models, though the depth at which it begins varies.

Externally, we evaluate twelve VQA benchmarks under progressive Gaussian blur. For each prediction we track whether the answer stays fixed through every level of degradation, rather than changing and later returning to the original answer. On several benchmarks a large fraction of predictions survive the entire blur sequence. On RePOPE, for example, about half remain unchanged through the strongest blur, and this invariant subset stays above 90% accuracy even as overall accu-

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.

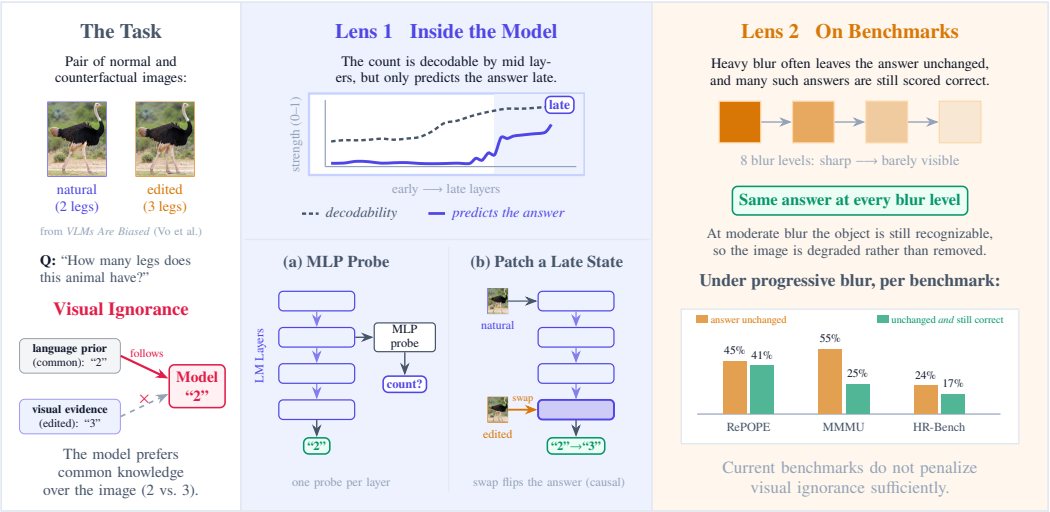


Figure 1: **Two lenses on visual ignorance in VLMs.** On paired normal/counterfactual images (Vo et al., 2025), the model returns the prior count ("2") for both. *Lens 1:* per-layer MLP probes show the count is decodable at intermediate layers but strongly coupled to output changes only late; replacing a late state shifts the answer margin (Qwen2.5-VL-3B-Instruct; replicated observationally in 7B and LLaVA). *Lens 2:* under progressive blur many answers stay unchanged and still score correct, so current benchmarks do not penalize visual ignorance sufficiently.

racy falls sharply. Inspecting these invariant answers reveals the same tendency observed inside the decoder: as the image is degraded, some explanations drift from visual description toward familiar associations, retaining the prior-consistent answer even when it is wrong. These results show that benchmark accuracy can stay high among predictions that do not respond to the degraded detail, revealing a gap between reported performance and use of the visual evidence a benchmark is intended to test.

This paper makes two contributions. First, we provide a causal, mechanistic account of visual ignorance: paired readouts and residual interchange show that answer-relevant evidence is often decodable in intermediate layers but shapes the output only late in the decoder, which tends to favor a learned prior; in Qwen-3B this late shift is driven mainly by attention, and the pattern holds across three VLMs. Second, we propose a degradation-based evaluation that blurs images progressively under a consecutive-agreement criterion across twelve benchmarks, quantifying how much credit goes to blur-insensitive predictions. Together these results separate the availability of visual information from its influence on prediction, and motivate training and evaluation that make answer-relevant visual distinctions necessary to answer correctly.

2 Related Work

Behavioral tests of visual dependence. Prior work documents VLM outputs predictable from text alone (Lin et al., 2024; Lee et al., 2025; Luo et al., 2025), preference for textual cues when text and images conflict (Deng et al., 2025), and stability after image removal (Asadi et al., 2026), showing that high accuracy need not imply dependence on instance-specific evidence. Yet an answer surviving image removal may reflect chance in a constrained answer space. We instead degrade the image gradually and require the answer to stay fixed across the whole sequence, reducing coincidental matches. Accuracy within this invariant subset then exposes benchmarks that credit predictions insensitive to the degraded detail.

Counterfactual and mechanistic analyses of visual information use. Counterfactual and conflict-based studies examine competition between visual evidence and learned priors. Visual CounterFact uses contrastive images to study pixel–prior competition and construct activation-steering directions (Golovanevsky et al., 2025); VLMBias tests whether familiar expectations over-

ride edited visual evidence (Vo et al., 2025); and related analyses show that visual detail can be preserved without guiding the answer (Rahmanzadehgervi et al., 2024; Fu et al., 2025). Mechanistic studies use vocabulary projections, trained lenses, sparse features, attention analyses, and interventions to examine representations and information flow (Nostalgebraist, 2020; Belrose et al., 2023; Makhzani & Frey, 2013; Cunningham et al., 2023; Pach et al., 2026; Jiang et al., 2025b;a). VLM studies further investigate visual-information transfer, knowledge evolution, and modality arbitration (Kaduri et al., 2025; Wang et al., 2025a; Nooralahzadeh et al., 2026). These methods establish such gaps observationally, associating representations with outputs but not testing whether the implicated states cause the answer. We instead use matched VLMBias pairs, where a normal and a counterfactual image share the question but differ in the correct count. We measure whether the internal update between the two images predicts the output change, and swap these states between paired images to reveal a sharp late-layer transition in their causal influence on the answer margin.

3 Diagnosing Internal Visual Updates

A VLM exhibits visual ignorance when an answer-relevant image change fails to produce an appropriate prediction update. We analyze this failure in three stages. First, layer replacement provides an initial localization, while per-layer semantic probes track where answer information is decodable. Second, matched normal–counterfactual pairs, in which the visual edit changes the correct answer, reveal when image-induced semantic updates become coupled to output changes, and residual interchange tests their causal influence on the output margin. Third, we test whether the coupling pattern generalizes across architectures. A consistent picture emerges: information about the answer may already be decodable from intermediate states, but image-induced updates remain weakly coupled to the output until a sharp late-layer transition. This late-coupling pattern recurs across architectures. In Qwen-3B, the sharpest transition is causally localized to attention in layer 27.

3.1 Coarse Layer Localization

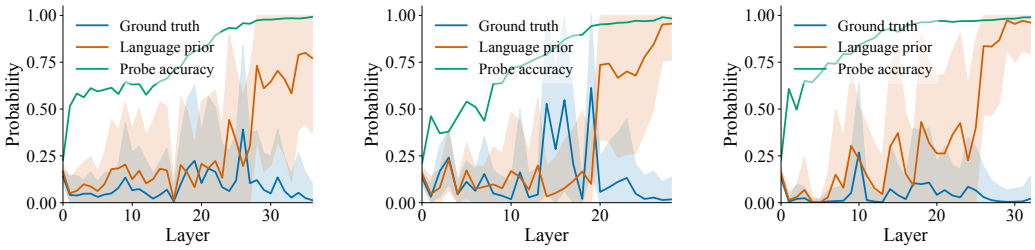
We begin with coarse layer localization in a Qwen-3B model fine-tuned on VLMBias counterfactual pairs. Each pair contains a normal image and an edited counterpart with the same question but a different correct answer. For example, a two-legged ostrich and an edited three-legged ostrich are paired with "How many legs does this animal have?" Fine-tuning increases accuracy on the analyzed subset from 11.6% to 65.2%. We replace contiguous groups of baseline decoder layers with their fine-tuned counterparts and measure the resulting improvement. Accuracy rises from 11.6% to 65.2% as replacement expands, with the largest increase across the final 14–20 layers; replacing only the middle layers raises accuracy to 24.3%. These results provide an initial localization to the middle and late decoder. Because hybrid networks can introduce representation mismatch, we next examine the frozen model layer by layer. Training details and examples appear in Appendix A.

3.2 Per-Layer Semantic Readout

Within each frozen model, we use independently trained MLP probes to read answer-related information from the final prompt-token state at every decoder layer. We focus on VLMBias examples whose correct answers are numerals from 1 to 9, and each probe predicts a distribution over these nine candidate counts; training details appear in Appendix D. For each counterfactual image, we track the probability assigned to the visual ground-truth count (GT) and the familiar count favored by the language prior (LP). Figure 2 shows that both counts are decodable across multiple layers, while late readouts increasingly favor LP. More directly, among Qwen examples whose final answers follow LP, GT receives over 50% probe probability at an intermediate layer with over 90% held-out probe accuracy in 49.8% of 3B cases and 62.5% of 7B cases, before LP becomes persistently dominant. Thus, in many failures, the visual ground-truth answer is strongly decodable before later states favor the language prior. We next use matched image pairs to test whether these internal changes are coupled to changes in the output.

3.3 Paired Counterfactual Updates

Single-image readouts show which answers are decodable at each layer, but not whether a visual edit changes those representations or reaches the output. We therefore analyze 197 matched VLMBias



(a) Qwen2.5-VL-3B-Instruct. (b) Qwen2.5-VL-7B-Instruct. (c) LLaVA-v1.6-Mistral-7B.

Figure 2: Full-softmax semantic probe readout. At every decoder layer, an independently trained nine-class MLP assigns probability to all counts in $\{1, \ldots, 9\}$; blue and orange are the mean probabilities of the GT and LP counts on the filtered VLMBias evaluation subset, with standard-deviation bands. Green is held-out probe accuracy.

pairs, each holding the question fixed while an image edit changes the correct count from the familiar count A_{LP} to the visual ground-truth count A_{GT}. Pairwise comparison removes shared preferences and isolates the response to the edit. We use two matched wordings: Q1 retains the dataset form, e.g., "How many legs does this animal have?", while Q2 uses "Count the legs of this animal." Both require a single numeral. Under Q1, Qwen-3B answers 138 normal images correctly but only 6 counterfactual images, retaining the original answer in 131 of the 138 initially correct cases. Q2 reproduces this failure in 120 of 132 cases. To link this failure to internal changes, we reuse the readout probes. For pair p at layer l, the probe margin $s_{p,l}$ is the logit assigned to A_{GT} minus that assigned to A_{LP}; the output margin t_p uses the model's logits for the same counts. We define:

$$u_{p,l} = s_{p,l}^{\mathrm{cf}} - s_{p,l}^{\mathrm{normal}}, \qquad M_p = t_p^{\mathrm{cf}} - t_p^{\mathrm{normal}},$$

where cf denotes the counterfactual image. Positive values indicate movement toward the edited count. If an internal semantic update contributes to answer revision, larger $u_{p,l}$ should predict larger M_p.

3.4 Internal–Output Update Coupling

We quantify this relationship by regressing the output update M_p on the internal update $u_{p,l}$ at each layer. The slope β measures output-margin change per standard deviation of internal update. Figure 3(a) reports the layerwise slopes under both question wordings, while panel (b) reports coupling for layer-range averages and the late-minus-middle contrast. Full experimental and statistical details appear in Appendix E.

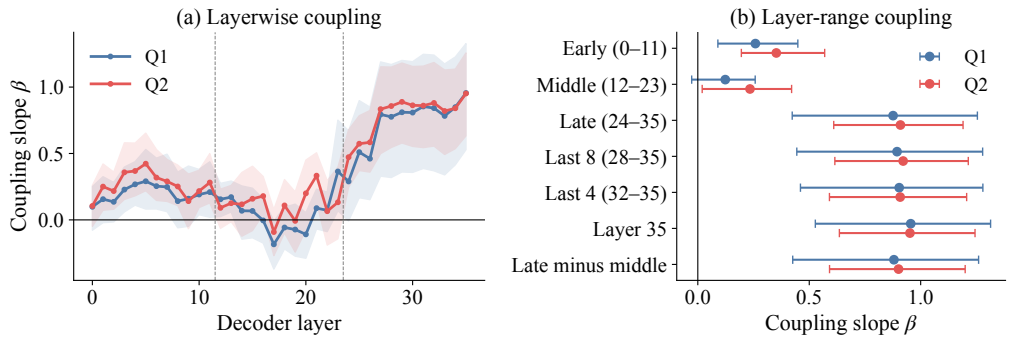


Figure 3: Internal–output update coupling in Qwen2.5-VL-3B-Instruct. (a) Layerwise coupling slopes under two prompts Q1 and Q2. (b) Coupling for layer-range averages and the late-minus-middle contrast. Shading and bars show 95% confidence intervals.

As Figure 3 shows, coupling remains weak through layers 12–23, where Q1 yields $\beta = 0.123$, and then jumps sharply after layer 24. Over layers 24–35, β reaches 0.876 for Q1 and 0.909 for Q2.

Prior work shows that visual information and multimodal knowledge evolve across decoder depth, with late layers increasingly involved in forming the output (Kaduri et al., 2025; Wang et al., 2025a; Nooralahzadeh et al., 2026). Our layerwise and range-averaged analyses reveal a clear boundary rather than a gradual progression. After this boundary, image-induced semantic updates become tightly linked to the corresponding answer revision.

3.5 Causal Residual Interchange

The clear late-coupling boundary motivates a causal test of the corresponding states in Qwen-3B. We intervene at one layer at a time, replacing only the normal run's final prompt-token state at that layer with its counterfactual counterpart. Using the output margin t_p defined above, we compute the induced logit-margin update Δt as the intervened margin minus the original normal-run margin. Positive values indicate movement toward GT. More details appear in Appendix F.

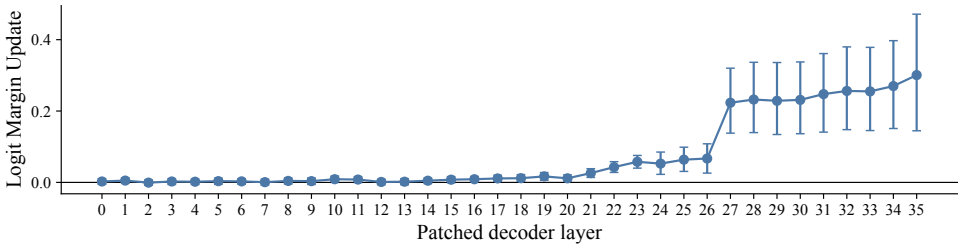


Figure 4: Causal residual interchange in Qwen2.5-VL-3B-Instruct. Each point is the mean logit-margin update Δt after replacing the normal run's final prompt-token state with its counterfactual counterpart at a given layer. Positive values indicate movement toward GT; bars are 95% confidence intervals.

As Figure 4 shows, residual interchange in Qwen-3B has little effect through the middle decoder and begins to shift the output margin around layers 21–23. Mean Δt then more than triples at layer 27, rising from 0.067 to 0.223, and remains large through the final layers, reaching 0.301 at the output. Range-level interventions reproduce this separation: layers 25–31 and 32–35 yield $\Delta t = 0.247$ and 0.301, respectively, while layers 4–10 have negligible influence. Replacing the normal run's late states with their counterfactual counterparts therefore causally shifts the answer margin toward GT.

3.6 Module-Level Localization

The abrupt jump at layer 27 motivates a finer localization within the layer. A Transformer layer contains self-attention and MLP submodules, each combined with the residual stream through a residual update (Vaswani et al., 2017). In the pre-norm Qwen architecture, each submodule receives an RMS-normalized residual state. We therefore repeat residual interchange at three boundaries in layers 26, 27, and 28: the layer input, the post-attention state, and the layer output. For all 197 Q1 pairs, we replace the normal run's final prompt-token state at each boundary with its counterfactual counterpart and allow the remaining computation to proceed from the replaced state. We compute the same logit-margin update Δt as in Section 3.5; paired differences in Δt between adjacent boundaries measure the gain associated with the attention and MLP steps.

Table 1 localizes the jump to the attention step of layer 27. Mean Δt jumps sharply from 0.067 at the layer input to 0.227 after attention, then remains nearly unchanged at the layer output (0.223). The corresponding attention gain is $+0.160$, whereas the MLP gain is -0.003. The neighboring layers show no comparable attention gain, and only the layer-27 attention gain survives correction over the nine layer-by-step tests. The sharp causal transition in Qwen-3B therefore occurs primarily during attention in layer 27. Appendix G provides the technical details.

3.7 Cross-Architecture Generalization

We test whether the observational internal–output coupling pattern found in Qwen-3B generalizes to Qwen-7B (28 layers) and LLaVA-7B (32 layers). For each model, we train model-specific probes

Table 1: Module-level residual interchange in layers 26–28 of Qwen2.5-VL-3B-Instruct, evaluated on all 197 Q1 pairs. At each boundary, the normal run's final prompt-token state is replaced with its counterfactual counterpart. Values are mean logit-margin updates Δt, with positive values indicating movement toward GT. Gains are paired differences in Δt between adjacent boundaries. Only the layer-27 attention gain, shown in bold, survives correction over the nine layer-by-step tests.

Layer	Input	After attn.	Output	**Attn. gain** (95% CI)	**MLP gain** (95% CI)
26	0.064	0.073	0.067	$+0.009$ [-0.007, 0.026]	-0.006 [-0.014, 0.002]
27	**0.067**	**0.227**	**0.223**	$+0.160$ [0.096, 0.233]	-0.003 [-0.011, 0.004]
28	0.223	0.231	0.232	$+0.008$ [-0.015, 0.032]	$+0.001$ [-0.007, 0.009]

and repeat the paired regression, aligning layers by normalized decoder depth. Because the slope β depends on each model's output scale, we compare models using the signed square root of partial R^2. This analysis tests observational coupling in all three models; causal residual interchange and module-level localization were evaluated in Qwen-3B.

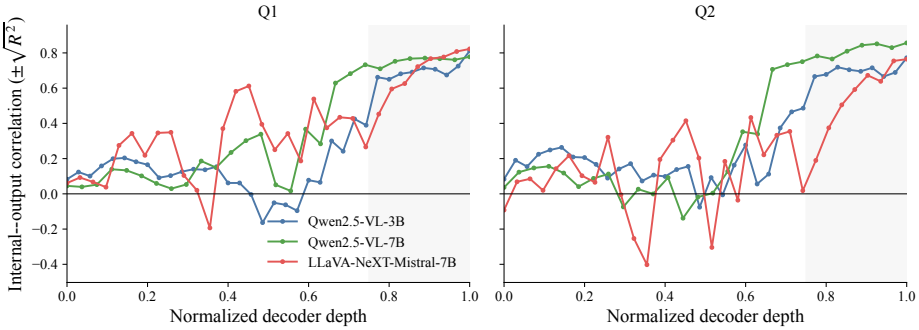


Figure 5: Cross-architecture internal–output update coupling for Qwen-3B, Qwen-7B, and LLaVA-7B. The left and right panels show Q1 (dataset wording) and Q2 (matched command paraphrase), respectively, with layers aligned by normalized decoder depth. Each curve reports the signed square root of partial R^2, which is a correlation-like measure of the direction and strength of internal-output coupling. Shading marks the final quarter of the decoder.

Figure 5 shows weak and variable coupling across early and middle layers, followed by a sustained positive late band in all three models and under both question wordings. At the final layer, the Q1 values reach approximately $+0.81$, $+0.78$, and $+0.82$ for Qwen-3B, Qwen-7B, and LLaVA-7B, respectively. The sustained positive late-layer band begins at normalized decoder depths 0.71, 0.59, and 0.77, respectively. The band therefore begins at different normalized depths across architectures, while the late concentration of positive coupling is consistent across models and prompts. Model-specific training, tokenization, and statistical details appear in Appendix H.

Together, these analyses separate the availability of answer-related information from its influence on prediction: the relevant answer may be decodable from intermediate states, while image-induced semantic updates remain weakly or variably coupled to output revision until late in the decoder. This pattern recurs across all three architectures as a sustained positive late-layer band, although the band begins at different normalized depths. In Qwen-3B, residual interchange further reveals a sharp transition in causal influence, with late states shifting the answer margin toward the visual ground truth and most of the increase occurring during the attention step of layer 27.

4 Visual Insensitivity Under Controlled Image Interventions

Answer-related visual information may be represented within the decoder yet influence prediction only late, raising a behavioral question: can a model receive benchmark credit while remaining insensitive to the visual detail relevant to its answer? We test this across twelve benchmarks and three models by progressively removing image detail and measuring the fraction of answers that remain unchanged throughout substantial detail loss, together with the accuracy retained by this invariant

Question Q: Where is the drink with the lime? A. In the middle of the two yellow drinks. B. On the right of the two yellow drinks. C. On the left of the two yellow drinks.
Ground Truth Answer $\tilde{A}$: A

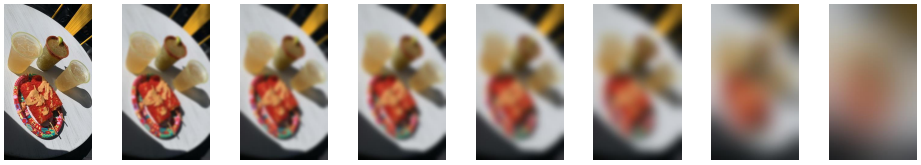


Image I								
Gaussian Blur Kernel Size k_i	1×1	3×3	5×5	7×7	11×11	15×15	31×31	61×61
	Consecutively Identical Answers							
Model's Answer A_{I,Q,k_i}	A	A	A	C	C	A	C	A
Identical Answer Indicator E_{I,Q,k_i}	1	1	1	0	0	1	0	1
Consecutively Identical Answer Indicator C_{I,Q,k_i}	1	1	1	0	0	0	0	0

Figure 6: Progressive visual decay and consecutive agreement. A return to the original answer at a later blur level is excluded after the agreement chain first breaks.

subset. Sustained invariance helps determine which visual distinctions benchmark success actually requires, although an unchanged answer can remain justified when sufficient evidence survives.

4.1 Progressive-Decay Evaluation

We consider twelve benchmarks: RePOPE (Neuhaus & Hein, 2025), HR-Bench (Wang et al., 2025b), both HallusionBench partitions (Guan et al., 2024), AI2D (Kembhavi et al., 2016), MMMU (Yue et al., 2024), V*Bench (Wu & Xie, 2024), VLMBias (Vo et al., 2025), RealworldQA (X.AI, 2024), BLINK (Fu et al., 2024), MMBench (Liu et al., 2024b), SEED-Bench (Li et al., 2023), and MMStar (Chen et al., 2024). For each image I and question Q, we progressively apply Gaussian blur with kernels $K = (1\times1, 3\times3, 5\times5, 7\times7, 11\times11, 15\times15, 31\times31, 61\times61)$. At the strongest blur setting, 61×61, object outlines are completely indistinguishable, as illustrated in Figure 6. Greedy decoding yields answer A_{I,Q,k_i} at each level, with k_1 denoting the unblurred baseline. We define

$$E_{I,Q,k_i} = \mathbb{1}[A_{I,Q,k_i} = A_{I,Q,k_1}], \qquad C_{I,Q,k_i} = \bigwedge_{j=1}^{i} E_{I,Q,k_j}.$$

E_{I,Q,k_i} indicates whether the answer at the current blur level equals the original-image answer at k_1. As the logical AND operator $\bigwedge$ requires, $C_{I,Q,k_i} = 1$ only when this equality holds at every level through k_i; once broken, the chain cannot be restored by a later return (Figure 6). Under an illustrative independent uniform-guessing baseline for a binary question, the chance of matching the original answer falls from $1/2$ for one degraded image to $(1/2)^7$ across all seven degraded levels, reducing accidental matches in constrained answer spaces. We report dataset means $\bar{E}_{\mathcal{D},k_i}$ and $\bar{C}_{\mathcal{D},k_i}$, overall accuracy $\bar{\alpha}_{\mathcal{D},k_i}$, and invariant-subset accuracy $\bar{\gamma}_{\mathcal{D},k_i}$; full definitions and settings appear in Appendix I.

The two accuracy metrics answer different evaluation questions. At each blur level k_i, overall accuracy $\bar{\alpha}_{\mathcal{D},k_i}$ is computed over the full evaluation set with every image blurred to k_i, whereas invariant-subset accuracy $\bar{\gamma}_{\mathcal{D},k_i}$ is computed only over samples whose predictions have matched the original-image answer through levels $k_1, k_2, \ldots, k_i$. Because this subset shrinks with degradation, its accuracy can exceed the unblurred full-dataset accuracy. Such an increase reflects selection of invariant responses, not improved robustness of a fixed population. A high $\bar{C}_{\mathcal{D},k_i}$ establishes persistence along the blur sequence, and a high $\bar{\gamma}_{\mathcal{D},k_i}$ establishes that those persistent responses receive benchmark credit. We therefore report invariant-subset accuracy together with the fraction of predictions that remain invariant.

4.2 Experimental Results

We evaluate progressive decay across the twelve benchmarks and three models introduced above. Figure 7 reports trajectories averaged across datasets, while Appendix Figure 15 provides the individual trajectories for each benchmark. Appendix J reports complementary visual-decay controls, including grayscale, nested spatial masking, and text-only prompts, across the same three models and twelve benchmarks.

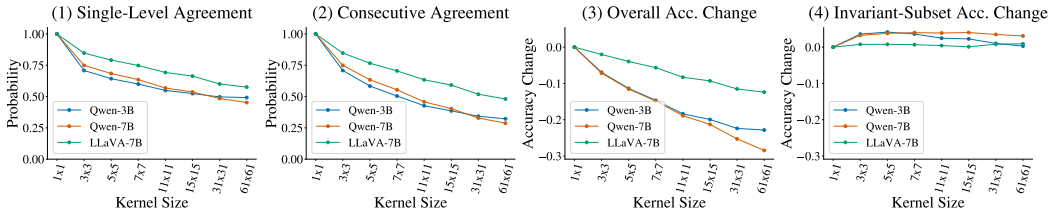


Figure 7: Dataset-averaged trajectories under increasing Gaussian blur for Qwen-3B, Qwen-7B, and LLaVA-7B. Panel 1 shows single-level agreement: the answer at the current blur level k_i equals the answer on the unblurred image k_1. Panel 2 shows consecutive agreement: the answer equals the unblurred answer at every blur level from k_1 through k_i. Panels 3 and 4 show overall accuracy change $\bar{\alpha}_{\mathcal{D},k_i} - \bar{\alpha}_{\mathcal{D},k_1}$ and invariant-subset accuracy change $\bar{\gamma}_{\mathcal{D},k_i} - \bar{\gamma}_{\mathcal{D},k_1}$, respectively. Because raw accuracies differ across benchmarks, each dataset's own k_1 accuracy is subtracted before averaging.

Many predictions persist under severe blur. As Figure 7 shows, substantial fractions of predictions remain unchanged throughout the blur sequence. At the strongest blur, dataset-averaged consecutive agreement remains 32.2% for Qwen-3B, 28.7% for Qwen-7B, and 48.1% for LLaVA-7B. Per-benchmark trajectories show particularly high persistence on RePOPE and substantial persistence on MMMU (Figure 15a; Figure 15f). Even when object outlines are indistinguishable, the models do not uniformly abstain or report that the image is unreadable. Many predictions instead retain the same scorable answer through all seven degraded levels, showing that the models do not fully use the visual detail removed by blur on a substantial subset of questions.

Consecutive agreement excludes trajectories that later return to the original answer. At the strongest blur, single-level and consecutive agreement are respectively 49.2% and 32.2% for Qwen-3B, 45.2% and 28.7% for Qwen-7B, and 57.6% and 48.1% for LLaVA-7B (Figure 7). The gaps represent answers that changed at an earlier blur level and later returned to the original answer, which endpoint-only matching would count as unchanged. By excluding these interrupted returns, consecutive agreement provides a more reliable conservative lower bound on detectable visual insensitivity.

Unchanged predictions retain high benchmark accuracy. On RePOPE, overall accuracy falls from 92.7% to 56.8% under the strongest blur, while invariant-subset accuracy remains at 93.3% (Figure 15a). The dataset average shows the same separation. From the unblurred images to the strongest blur, overall accuracy decreases by 22.8, 28.4, and 12.4 percentage points for Qwen-3B, Qwen-7B, and LLaVA-7B, respectively, whereas invariant-subset accuracy changes by $+0.3$, $+3.1$, and $+0.9$ percentage points (Figure 7). Thus, benchmark credit remains high among predictions that do not respond to the removed detail even as full-dataset performance declines. Together, these results show why accuracy alone does not establish use of the visual details affected by degradation.

Visually dependent questions lose invariant-subset accuracy faster. HallusionBench's Visual Dependent split loses invariant-subset accuracy faster than its Visual Supplement split for both Qwen models (Figure 15c; Figure 15d). This ordering is consistent with the greater importance of image content to correctness in the Visual Dependent split: unchanged predictions receive less credit as relevant evidence is degraded. The contrast supports the evaluation's sensitivity to differences in benchmark visual dependence.

Model size does not consistently predict invariance. In the dataset average (Figure 7), the strongest blur reduces LLaVA-7B's accuracy by 12.4 percentage points, compared with 22.8 and

28.4 percentage points for Qwen-3B and Qwen-7B, respectively; LLaVA-7B also retains more original answers. Yet the Qwen 3B and 7B models exchange order across datasets, so larger size does not consistently reduce invariance. Lower degradation and higher consecutive agreement may reflect more severe visual insensitivity rather than stronger use of visual evidence.

4.3 Qualitative Analysis

Figure 8 illustrates an invariant error from Qwen-7B. Regarding the cows' food, the model retains *hay* as blur increases and shifts from visual claims to the familiar association between cows and hay, although the label is *grains*. Appendix I provides a complementary correct invariant response.



Original Image



Blurred (61x61)

Question: What is the primary food source for the animals in this image?
Options: A. Grains; B. Hay; C. Fruits; D. Vegetables
Ground Truth: A. Grains

VLM Responses for Different Kernel Sizes of Gaussian Blur:

1x1 (Original Image): ... let's analyze the visual clues: ... They are standing in what looks like a barn or a feeding area... There is a large amount of what appears to be hay on the ground in front of them. Cows are known to eat hay as part of their diet. Given these observations, the most likely primary food source for these animals is hay. ANSWER: B

3x3: ...Cows are typically herbivores and are known to eat hay as a primary food source. The image shows the cows with their heads down, which is consistent with them eating hay. There is no visible evidence of grains, fruits, or vegetables being fed to the cows in the image. Given these observations, the primary food source for the animals in the image is most likely hay. ANSWER: B

5x5: (almost the same as 3x3)

7x7: ... In the image, the animals are not shown eating hay or vegetables, but they could be grazing on grass. Grains are often used as supplementary feed in dairy farming to supplement the diet with additional nutrients. Given these points, the most likely primary food source for these cows would be hay... ANSWER: B

11x11: The image shows the animals with what looks like hay in their mouths, suggesting they are eating hay... ANSWER: B

15x15: (similar to 3x3)

31x31: To determine the primary food source for the animals in the image, let's analyze the options provided: A: Grains ... are common feed for many animals. B: Hay ... often used as bedding but can also be a food source for some animals ... Given the blurred nature of the image, it is difficult to identify specific food sources. However, considering the context of the question and the options provided, grains and hay are more commonly associated with animal feeding than fruits and vegetables. Since the image does not provide clear evidence of fruits or vegetables, and both grains and hay are plausible options, we need to choose the most likely primary food source based on general knowledge ... the most likely primary food source for the animals in the image is either grains or hay. ANSWER: B

61x61: (similar to 31x31)

Figure 8: An invariant error from Qwen2.5-VL-7B-Instruct in the MMStar (Chen et al., 2024) dataset. The prediction remains *hay* by general knowledge, despite the label *grains*.

5 Discussion and Conclusion

Visual evidence availability and use. Many training distributions may allow a model to predict the answer from the question, coarse scene information, and familiar associations. When these signals are usually sufficient, there is little pressure to inspect the local details that determine the answer or preserve their influence through decoding. Our results show that answer-relevant detail can remain decodable while exerting limited influence on answer revision. This influence changes sharply across a clear late-layer boundary, including the abrupt attention-mediated jump at layer 27 in Qwen-3B, while coupling becomes strong only in late layers across models. A central direction for future work is to identify what determines this boundary and whether training can make detailed visual evidence influence predictions earlier and remain influential throughout decoding.

Implications for training. A training set remains vulnerable to this shortcut when the question and broad visual content reliably predict the target. Scaling similar data will continue to reward answers based on these coarse signals. Effective training distributions should make the target difficult to infer without the decisive visual detail, reduce the reward for relying on broad scene information, and maintain the influence of that detail as the answer is formed.

Implications for evaluation. A strong visual evaluation should ensure that the answer genuinely depends on the tested detail and cannot be reliably inferred from the text and broad scene content. Model behavior should change when the decisive evidence changes, while plausible guesses based on familiar associations should receive no advantage. Scoring should distinguish correct use of the relevant evidence from coincidental correctness, so that high performance reflects sensitivity to the visual information the task is intended to test.

Taken together, our findings underscore the importance of task-relevant visual evidence in shaping a model's output. Broad scene understanding provides useful context, while reliable vision-language reasoning also depends on the continued influence of the specific visual details that determine the appropriate response. Ultimately, an ideal VLM should accurately identify and use the visual details that matter for the task.

AI USE STATEMENT

We used generative AI tools only to assist with writing. Specifically, they were used to edit the manuscript for readability, to tighten phrasing, and to help refine the wording of the abstract, introduction, and related-work sections. All suggestions were reviewed and revised by the authors, and every scientific claim was checked against our experiments and the cited sources. We did not use generative AI tools to generate data, develop the methodology, design or implement experiments, produce proofs, or interpret results; these tasks are either not applicable or were carried out entirely by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work evaluates publicly available vision-language models on existing benchmarks, including released counterfactual image pairs. We do not collect new data or infer information about individuals. Our interventions are limited to controlled image degradations and state replacements for analysis. The goal is to improve whether models use task-relevant visual evidence. Any deployment of more reliable vision-language systems should nevertheless be evaluated for application-specific risks, including errors that may affect people.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the paired counterfactual analysis, residual interchange procedure, progressive-decay evaluation, and consecutive-agreement metric. The appendix reports the fine-tuning and layer-replacement setup, probe architecture and data splits, paired-regression details, intervention definitions, model-specific replication settings, visual-decay controls, and per-benchmark trajectories. We use publicly available model checkpoints and benchmark data; prompts, sample counts, random seeds, hyperparameters, and statistical procedures are specified in the corresponding sections and appendices.

REFERENCES

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding, 2026. URL https://arxiv.org/abs/2603.21687.

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

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3867–3876, 2025.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations. *arXiv preprint arXiv:2506.08008*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166, 2024.

Michal Golovanevsky, William Rudman, Michael A Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24848–24863, 2025.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pp. 14375–14385, 2024.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025a.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025b.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14549–14558, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29914–29934. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lin24c.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. In *International Conference on Machine Learning*, pp. 41120–41156. PMLR, 2025.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Yannic Neuhaus and Matthias Hein. Repope: Impact of annotation errors on the pope benchmark. *arXiv preprint arXiv:2504.15707*, 2025.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026.

Nostalgebraist. Interpreting gpt: The logit lens. https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, Aug 2020.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *Advances in Neural Information Processing Systems*, 38:95706–95742, 2026.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Proceedings of the Asian Conference on Computer Vision*, pp. 18–34, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://papers.nips.cc/paper/7181-attention-is-all-you-need.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased, 2025. URL https://arxiv.org/abs/2505.23941.

Sudong Wang, Yunjian Zhang, Yao Zhu, Jianing Li, Zizhe Wang, Yanwei Liu, and Xiangyang Ji. Towards understanding how knowledge evolves in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29858–29868, 2025a.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

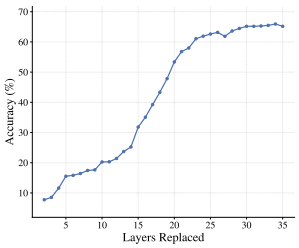
Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024.

X.AI. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

A GRPO Fine-Tuning and Layer Replacement

We fine-tuned a debiased Qwen2.5-VL-3B-Instruct with GRPO and LoRA on VLMBias (Appendix B) and replaced suffixes of the baseline decoder with the corresponding fine-tuned layers. Figure 9 shows that accuracy rises from 11.6% to 65.2% as more layers are replaced, with the largest increase when replacing the final 14–20 layers and a smaller increase for the final 3–5 layers; its two examples illustrate cases corrected by five and twenty layers. Replacing only the middle range improves accuracy to 24.3%, suggesting that neither middle nor late layers alone explain the failure. Because hybrid networks can suffer representation mismatch, we use this as coarse causal localization rather than precise layer attribution.



(a) Accuracy versus replaced suffix.



Question: How many chess pieces are there on this board?

Ground Truth: 31 **Expected Bias:** 32

(b) Corrected by five layers.



Question: How many legs does this animal have?

Ground Truth: 5 **Expected Bias:** 4

(c) Corrected by twenty layers.

Figure 9: Coarse layer replacement and representative VLMBias examples. Additional examples appear in Appendix C.

B Training Setup

This section details the GRPO fine-tuning run whose layers are transplanted in Appendix A. We use the Qwen2.5-VL-3B-Instruct (Bai et al., 2025) model as our base vision-language model. The training is conducted using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a reinforcement learning algorithm that optimizes the model policy based on relative rewards within sampled groups. The experiments are executed on a cluster of 4 NVIDIA RTX 5880 Ada Generation GPUs using the Accelerate (Gugger et al., 2022) framework for distributed processing. We employ LoRA (Low-Rank Adaptation) (Hu et al., 2022) for parameter-efficient fine-tuning, targeting all major projections in the language model backbone (including the QKV projection layers in the attention modules and the MLP layers) with a rank $r = 32$ and $\alpha = 32$.

The model is trained for 8 epochs with a global batch size of 64 and a learning rate of 1×10^{-5} following a cosine decay schedule. To ensure training stability, we include a warmup phase of 100 steps. During the GRPO rollout, we set the temperature to 1.0 and top_p to 0.9, generating $G = 8$ completions per prompt to calculate relative advantages.

Our reward function is a composite metric designed to enforce both structural adherence and factual accuracy:

- Format Reward: Points are awarded for the inclusion of required markers (1. through 5.) and the specific curly brace delimiters.
- Accuracy Reward: A significant weight (75% of the total potential reward) is assigned to the correctness of the final extracted answer. Correctness is determined by exact string matching or numerical equivalence, ensuring the model's reasoning culminates in an accurate conclusion.

For the training dataset, we identify matching normal-counterfactual image pairs in the VLMBias dataset (Vo et al., 2025), e.g. a 2-legged ostrich image and a synthetic 3-legged counterpart. We include these images and the corresponding questions (e.g. How many legs does this animal have?) in

the training dataset. To enhance the diversity of the training distribution and improve model robustness, we apply perceptually-conservative augmentations to each image. Specifically, we generate 30 augmented variations per original image using a stochastic pipeline of color jitter, ISO noise, and JPEG compression. This process results in a curated dataset of 9,300 samples, where the semantic integrity and ground-truth labels of the visual reasoning tasks remain preserved despite the introduced low-level image perturbations.

To induce structured chain-of-thought (CoT) reasoning, we modify the system prompt to require a five-part response:

1. Object identification.
2. Visual cue description.
3. Evidence-to-question mapping.
4. Visual reasoning derivation.
5. Final answer enclosed in curly braces (e.g., {answer}).

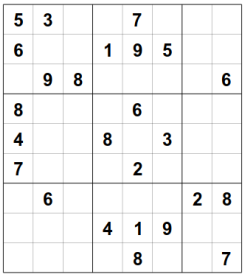
C More Examples for Layer Replacement

This section supplements the coarse layer-replacement study in Appendix A with more examples, grouped by how many final decoder layers must be replaced before the prediction becomes correct. The baseline Qwen2.5-VL-3B-Instruct model initially generates incorrect answers for all examples presented in Figure 10 and Figure 11. For the instances in Figure 10, substituting only the final 5 decoder layers with those from the fine-tuned model is sufficient to correct the predictions. On the other hand, the examples in Figure 11 require a deeper layer replacement; substituting the final 20 layers is necessary before the model successfully recovers the correct answers.



Question: How many visible white stylized curves are there in the logo of the left shoe?

Ground Truth: 2 **Expected Bias:** 1



Question: How many columns are there on this puzzle?

Ground Truth: 8 **Expected Bias:** 9

Figure 10: Questions related to the last 5 layers.

D Per-Layer Semantic Readout Details

The readout probes are trained on the official Pixmo-Count training split (Deitke et al., 2025), using only examples the frozen model answers correctly; no validation or test examples enter probe fitting. For each of the 36 decoder layers we train an independent three-layer MLP that maps the layer's final prompt-token state to a nine-class softmax distribution over the counts in $\{1, \ldots, 9\}$. Splits are grouped by image hash to prevent near-duplicate leakage, probe outputs are temperature-calibrated, and results are averaged across five initialization seeds for Qwen-3B and three for each replication model. The retained correct-only pools contain 19,753, 21,705, and 19,317 examples for Qwen-3B, Qwen-7B, and LLaVA, respectively. Qwen numerals are single tokens; for LLaVA/Mistral we use the shared numeral-prefix position followed by the distinct digit token. The green curve reported in



Question: How many visible white stripes are there in the logo of the left shoe?

Ground Truth: 4 **Expected Bias:** 3



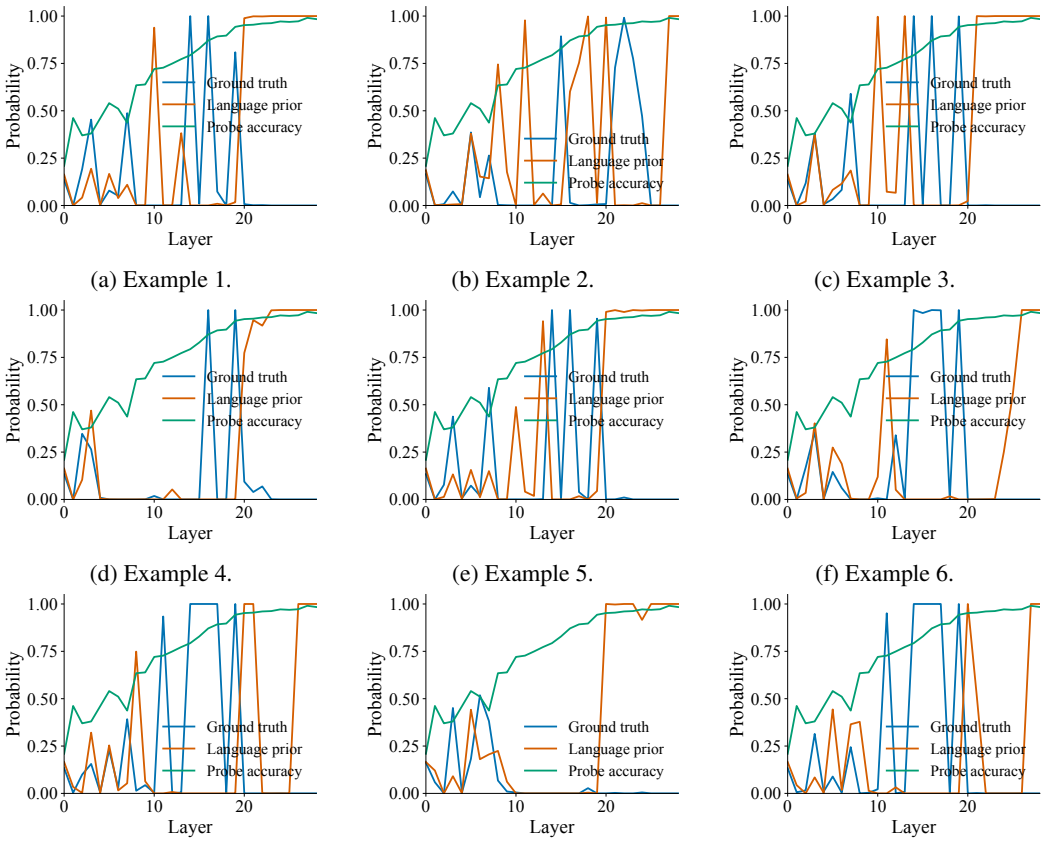
Question: How many stripes are there in this flag?

Ground Truth: 4 **Expected Bias:** 5

Figure 11: Questions related to the last 20 layers.

the main text is held-out probe accuracy on this split. Full optimizer, class-weighting, and calibration settings are shared with the paired analysis and are listed in Appendices E and H.

Beyond the population averages shown in the main text, Figure 12 shows the readout on individual Qwen2.5-VL-7B-Instruct VLMBias examples. In most cases the ground-truth score temporarily exceeds the language-prior score at an intermediate layer before the late layers again favor the prior; one example instead ends on a third count. These cases illustrate highly non-monotonic probe trajectories, but they are qualitative examples rather than a population-level estimate of representational instability.



(a) Example 1. (b) Example 2. (c) Example 3.

(d) Example 4. (e) Example 5. (f) Example 6.

(g) Example 7. (h) Example 8. (i) Example 9.

Figure 12: Per-layer semantic readout on selected Qwen2.5-VL-7B-Instruct VLMBias examples.

E Paired-Bridge Experimental Details

This appendix gives the full setup behind the paired analysis in Section 3: how the probes are trained, how the internal and output updates are defined, which reliability checks an effect must pass, and the diagnostic figures that support the main result.

Data and probes. Both Pixmo-Count and paired VLMBias images use the model-native processor with visual-token bounds corresponding to 224^2 minimum and 512^2 maximum pixels. The official Pixmo-Count training split has 36,916 rows, of which 30,547 have counts in $\{1, \ldots, 9\}$. Of these, 29,383 rows have locally recoverable images and 1,164 are unavailable; after removing 386 repeated image–label example IDs, we extract 28,997 unique examples. The frozen Qwen2.5-VL-3B-Instruct model answers 19,753 correctly, and only these examples enter probe fitting; no validation or test split from Pixmo-Count is used. Probe features are extracted with the prompt "How many {objects} are there in the image? Answer with a single Arabic numeral, without any additional text." For paired VLMBias evaluation, Q1 preserves the source question form (for example, "How many legs does this animal have?") and Q2 uses its matched command paraphrase ("Count the legs of this animal."); both receive the same single-numeral suffix. Analogous templates cover chess pieces, flag features, logo elements, and board rows or columns; optical-illusion pairs are excluded from the 197 numeric-count subset.

We train an independent three-layer MLP probe for each of the 36 decoder layers, with hidden widths 128 and 32, GELU activations, AdamW at learning rate 10^{-3} and weight decay 10^{-4}, inverse-frequency class weights, and batch size 256. A single 70:15:15 split grouped by image hash is fixed with seed 2027; only initialization and minibatch order vary across seeds 2027–2031. We early-stop on the validation split and fit one temperature per layer there to calibrate the probe outputs. The

retained counts for numerals 1 through 9 are 4,505, 3,921, 3,574, 3,386, 2,616, 930, 471, 297, and 53, so the largest counts are comparatively rare.

Internal and output updates. For each pair p, A_{GT} is the counterfactual target count and A_{LP} is the normal-image reference count; both labels remain fixed across the two image conditions. In either condition, the probe margin $s_{p,l}$ is the temperature-calibrated probe logit for A_{GT} minus that for A_{LP}, while the output margin t_p uses the model's own output-head logits. The internal and output updates are $u_{p,l} = s^{\mathrm{cf}}_{p,l} - s^{\mathrm{normal}}_{p,l}$ and $M_p = t^{\mathrm{cf}}_p - t^{\mathrm{normal}}_p$, respectively; positive values indicate movement toward the counterfactual target. For each model and prompt, we fit

$$\widehat{u}_{p,l} = \frac{u_{p,l} - \mu_l}{\sigma_l}, \qquad M_p = a_{g(p),l} + \beta_l \widehat{u}_{p,l} + \varepsilon_{p,l},$$

where μ_l and σ_l are the mean and population standard deviation of $u_{p,l}$ across the evaluated pairs, $g(p)$ denotes the pair's topic, $a_{g(p),l}$ is a topic-specific intercept, and $\varepsilon_{p,l}$ is the residual. The output update M_p remains in logit units, so β_l measures output-margin change per standard deviation of internal update. The main text uses β for the corresponding layer or range coefficient.

Reliability checks. We treat a layer's effect as reliable only if it passes all of the following: significance after multiple-comparison correction (Benjamini–Hochberg across layers, Holm across the predefined layer ranges), replication under the Q2 wording, agreement in four of five probe seeds, agreement in four of five leave-one-topic-out fits, adequate probe accuracy at that layer, a nonzero 95% interval from a 10,000-sample topic-stratified bootstrap, and significance against a 10,000-permutation null that shuffles the output update within each topic. Every test uses the image pair as the sampling unit.

Layer ranges and sliding windows. Panel (b) of Figure 3 reports coupling for mean internal updates over layers 0–11, 12–23, 24–35, 28–35, and 32–35, for layer 35 alone, and for the late-minus-middle contrast (the mean over layers 24–35 minus that over 12–23). As a further check that the late effect is not an artifact of one particular range, we also slide a fixed-width window across depth and refit at every start layer, using widths of 2, 4, 6, and 12 layers. For each range, contrast, or window, we first form each pair's aggregated internal update, then standardize it across pairs and fit the same regression.

Diagnostic figures. Figure 13a plots each layer's probe accuracy on the held-out test split (examples never seen during probe training); accuracy rises sharply after layer 16. Figure 13b splits the 197 pairs into LP update failures and other pairs, and plots the mean internal update across layers for each group; the two groups separate in the late layers. Figure 14 shows the slope β for every sliding window; the late layers stay strong at all four window widths, so the result does not depend on one range definition.

F Activation Interchange Details

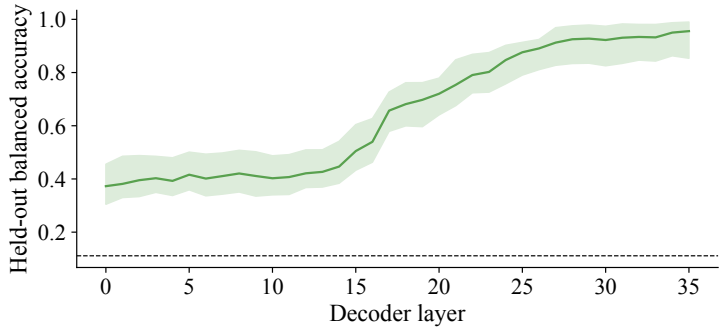
This appendix specifies the causal interchange experiment in Section 3.5. We use all 197 Q1 pairs and run Qwen2.5-VL-3B-Instruct on both the normal and counterfactual images under the same prompt. The paired images use identical 768×768 canvases, yielding the same input length, attention mask, and final prompt-token position. All image tokens and preceding text tokens occur before this final prompt token in the causal sequence. This alignment lets us replace its state without changing the remaining sequence.

Residual interchange. We index the L decoder blocks by $l = 0, \ldots, L-1$. For pair p, let $H^{\mathrm{normal}}_{p,l} \in \mathbb{R}^{T \times d}$ be the input sequence to block l for the normal image, where T is the sequence length and d is the hidden width. Writing F_l for decoder block l, its output is

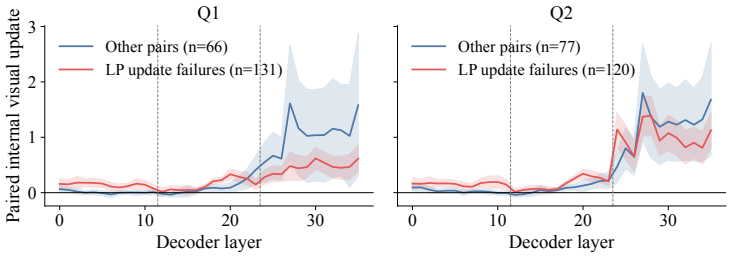
$$H^{\mathrm{normal}}_{p,l+1} = F_l\big(H^{\mathrm{normal}}_{p,l}\big).$$

An output intervention labeled l replaces row f, the final prompt-token state, in $H^{\mathrm{normal}}_{p,l+1}$ with its counterfactual counterpart:

$$\widetilde{H}^{\,\mathrm{normal}\leftarrow\mathrm{cf}}_{p,l+1}[j] = \begin{cases} H^{\mathrm{cf}}_{p,l+1}[f], & j = f, \\ H^{\mathrm{normal}}_{p,l+1}[j], & j \neq f. \end{cases}$$



(a) Probe accuracy on held-out examples, per layer.



(b) Mean internal update for LP update failures and other pairs.

Figure 13: Diagnostics for the paired analysis. (a) Shaded regions span bootstrap bounds across probe seeds. (b) Bands are 95% confidence intervals around each group mean.

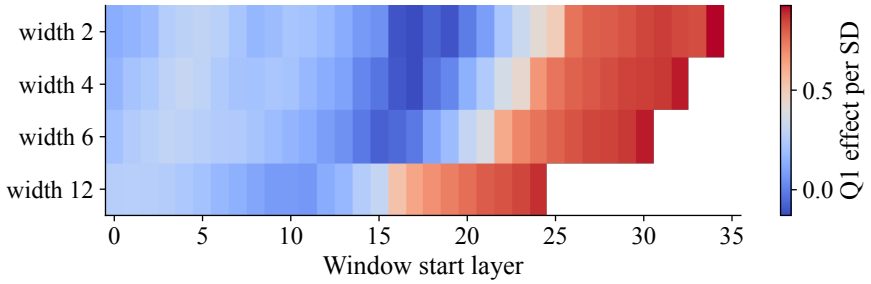


Figure 14: Coupling slope β for sliding windows of widths 2, 4, 6, and 12 layers, plotted against the window's start layer. The late layers remain strong at every width, so the effect is not tied to one range definition.

All other token states remain from the normal run at the intervention point. Because the image and preceding text tokens occur before position f, causal attention prevents the replacement at f from changing their subsequent states. The modified final prompt token can still attend to these preceding states as the remaining decoder layers operate normally:

$$\widetilde{H}^{\text{normal}\leftarrow\text{cf}}_{p,k+1} = F_k\left(\widetilde{H}^{\text{normal}\leftarrow\text{cf}}_{p,k}\right), \qquad k = l+1, \ldots, L-1.$$

We repeat this intervention independently at each of the 36 layers and read the resulting logits for counts 1–9. We use the same GT-minus-LP output margin t_p as in Section 3.3. Let $\widetilde{t}^{\text{normal}\leftarrow\text{cf}}_{p,l}$ denote this margin after replacing $H^{\text{normal}}_{p,l+1}[f]$ with its counterfactual counterpart. The logit-margin update is

$$\Delta t_{p,l} = \widetilde{t}^{\text{normal}\leftarrow\text{cf}}_{p,l} - t^{\text{normal}}_{p}.$$

Positive values indicate movement toward GT. As an implementation check, replacing a state with itself produces exactly zero update at every layer.

For significance, we form 95% confidence intervals with a 10,000-sample bootstrap over pairs and test each layer with a Wilcoxon signed-rank test, correcting across all 36 layers with the Holm method. Besides single layers, we also replace the normal states with their counterfactual counterparts across four fixed layer ranges (4–10, 12–18, 25–31, and 32–35) to summarize several layers at once. Finally, as a sensitivity check, we repeat the intervention using an unrelated image matched on prompt, topic, and both counts, for 12 pairs at six representative layers.

G Locating the Layer-27 Jump Within the Layer

This appendix gives the details behind Section 3: how we test which step inside a decoder layer produces the sharp jump at layer 27, and the checks that confirm the interventions are implemented correctly.

Qwen-3B uses a pre-norm Transformer decoder. For a fixed pair and image condition, let $x = H_{p,l}$ denote the full input sequence to block l, y its post-attention sequence, and $z = H_{p,l+1}$ its output sequence. The two residual updates are

$$y = x + \mathrm{Attn}(\mathrm{RMSNorm}_{\mathrm{attn}}(x)), \qquad z = y + \mathrm{MLP}(\mathrm{RMSNorm}_{\mathrm{mlp}}(y)),$$

where $\mathrm{RMSNorm}_{\mathrm{attn}}$ and $\mathrm{RMSNorm}_{\mathrm{mlp}}$ denote the separate normalization operations before the attention and MLP submodules. The states x, y, and z are the three replacement points reported in Table 1. Because layers are indexed from zero, the increase between the preceding layer output and the layer-27 output is produced within the layer labeled 27.

At each boundary, we replace only row f, the normal run's final prompt-token state, with its counterfactual counterpart and recompute the remaining operations. We apply this intervention to all 197 Q1 pairs in layers 26, 27, and 28, using the same logit-margin update Δt as in Section 3.5. For each pair, attention gain is the post-attention Δt minus the input Δt, and MLP gain is the output Δt minus the post-attention Δt. Replacing the normal state with itself yields zero update, and the layer-input and layer-output replacements reproduce the corresponding points from the layerwise scan. Confidence intervals use the image pair as the sampling unit, and Holm correction is applied across the nine layer-by-step tests.

H Cross-Architecture Paired Analysis

The Qwen-7B and LLaVA replications use the same 28,997 unique Pixmo-Count training examples and all 788 paired records (197 pairs, two prompt variants, and two image conditions). Full artifact checks require unique IDs, finite tensors, and feature shapes $[28, 3584]$ for Qwen2.5-VL-7B-Instruct and $[32, 4096]$ for LLaVA-v1.6-Mistral-7B. The correct-only probe pools contain 21,705 and 19,317 examples, respectively. Three initialization seeds (2027–2029) share an image-hash-grouped 70:15:15 split, normalization statistics, held-out IDs, and calibration protocol; only initialization and minibatch order vary.

Qwen numerals are represented by nine distinct single tokens. Mistral tokenizes a bare numeral as a shared prefix token (ID 28705) followed by one of nine digit tokens. We append the shared prefix to each LLaVA prompt and extract the final-token representation and logits for the following digit decision. The LLaVA paired input contains 2,928 image tokens and has identical sequence geometry across pair members. For both architectures, a hook-based extractor that stores only the final-token state at every layer is compared against the full `output_hidden_states=True` path; features and output logits match bit-for-bit on paired smoke tests.

At each layer, we fit the topic-fixed-effect regression using the calibrated probe margins and predictor standardization defined in Appendix E. Q1 is primary and Q2 is a wording replication. Layerwise Q1 tests use Benjamini–Hochberg correction; 10,000-sample topic-stratified bootstrap intervals quantify uncertainty, and all three seeds must agree in direction. For L decoder blocks indexed by $l = 0, \ldots, L-1$, normalized depth is $\delta_l = l/(L-1)$, with $L = 36, 28, 32$ for Qwen-3B, Qwen-7B, and LLaVA, respectively. We report this depth and predefined early, middle, late, final-quarter, and final-eighth intervals. The terminal contiguous band begins at the first layer after which every

remaining layer meets the cross-prompt criteria, while the complete curve retains any earlier isolated effects. Partial R^2 is the fraction of residual variance explained by the internal update after controlling for topic. Its square root, signed by β, is correlation-like.

I Behavioral Metric Details and Representative Results

This appendix gives the exact definitions behind the four metrics reported in Section 4 and shows the per-benchmark trajectories omitted from the main text. For each sample $(I, Q, \tilde{A})$, where $\tilde{A}$ is the ground-truth answer, and each blur level k_i, E_{I,Q,k_i} indicates single-level agreement, meaning that the current answer matches the original-image answer. The indicator C_{I,Q,k_i} records consecutive agreement, meaning that this match has held at every level from k_1 through k_i. Averaging these indicators over a dataset gives

$$\bar{E}_{\mathcal{D},k_i} = \mathbb{E}_{\mathcal{D}}[E_{I,Q,k_i}], \qquad \bar{C}_{\mathcal{D},k_i} = \mathbb{E}_{\mathcal{D}}[C_{I,Q,k_i}],$$

the single-level and consecutive agreement rates, respectively.

Overall accuracy at blur level k_i is $\bar{\alpha}_{\mathcal{D},k_i} = \mathbb{E}_{\mathcal{D}}[\mathbb{1}(A_{I,Q,k_i} = \tilde{A})]$. We also report invariant-subset accuracy over $\mathcal{D}_{k_i} = \{(I, Q, \tilde{A}) \in \mathcal{D} : C_{I,Q,k_i} = 1\}$, the subset whose answers have remained unchanged through k_i. For a nonempty subset,

$$\bar{\gamma}_{\mathcal{D},k_i} = \mathbb{E}_{(I,Q,\tilde{A})\sim\mathcal{D}_{k_i}} \left[\mathbb{1}(A_{I,Q,k_i} = \tilde{A})\right],$$

where the expectation is the uniform empirical mean over $\mathcal{D}_{k_i}$. Because the invariant subset becomes smaller as blur increases, its accuracy can exceed the unblurred full-dataset accuracy. A high invariant-subset accuracy under severe blur means that the benchmark continues to reward many predictions that do not respond to the removed visual detail.

All blur conditions within a benchmark use the same evaluation samples, prompt format, greedy decoding, and answer-extraction rule; only the image changes. Greedy decoding ensures that changes along the blur path reflect the image intervention rather than sampling variability. Consecutive agreement therefore measures the stability of each deterministic prediction across the tested blur sequence. Dataset-specific prompting and answer normalization follow each benchmark's evaluator.

J Visual-Decay Controls Across Benchmarks

We complement progressive Gaussian blur with visual-decay controls across the same twelve benchmarks and three models. Grayscale removes color while preserving luminance and geometry. For nested spatial masking, we divide each image into an 8×8 grid and replace a deterministic nested subset of 25%, 50%, or 75% of cells with neutral gray. Thus, cells masked at 25% remain masked at 50% and 75%. The text-only control removes image content from the prompt. These controls test whether the blur results also appear when visual evidence is reduced through color removal, spatial occlusion, or complete removal.

The masking results show the same broad pattern as blur. The fractions of predictions unchanged from the original image through all three mask levels are 48.1%, 53.4%, and 65.0% for Qwen-3B, Qwen-7B, and LLaVA-7B, respectively. Their accuracy within these unchanged subsets is 74.4%, 76.6%, and 57.6%. Text-only accuracy remains 34.8%, 37.8%, and 34.2%, indicating that benchmark credit can persist without image input. These controls do not show that every unchanged answer ignores all visual evidence: grayscale and masks retain some cues, and a stable answer can remain justified when sufficient evidence survives. Instead, together with progressive blur, they show that benchmark accuracy alone is incomplete evidence that a model uses the visual detail affected by the tested visual decay.

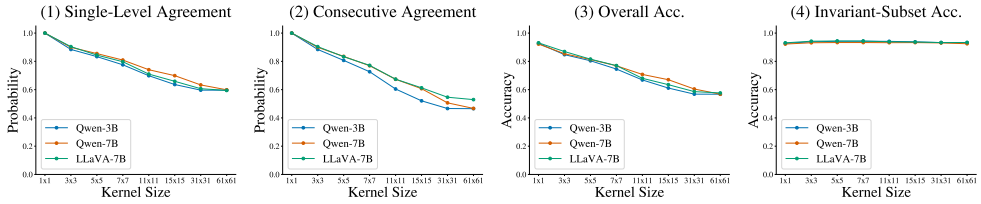
K Per-Benchmark Blur Trajectories and More Qualitative Examples

Figure 15 collects the per-benchmark trajectories, with one row for each benchmark or diagnostic partition and one curve per model. The dataset-averaged trajectories in Figure 7 are unweighted

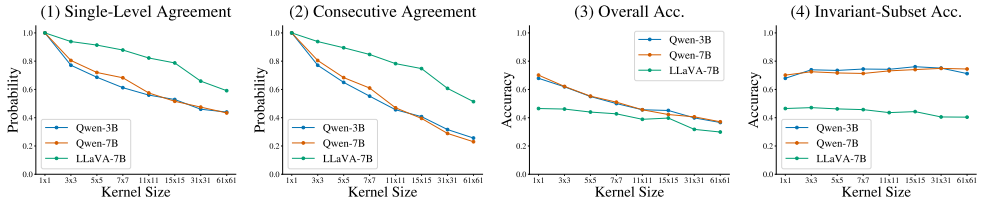
Table 2: Benchmark-scale visual-decay controls across Qwen-3B, Qwen-7B, and LLaVA-7B. All values are percentages and micro-averages over the same 42,078 examples from 12 benchmarks. "Same" is exact agreement with the corresponding clean-image answer. Nested masks divide each image into an 8×8 grid and replace nested 25%, 50%, and 75% subsets of cells with neutral gray; "all-level same" requires agreement under the original image and all three mask levels, while its paired accuracy is measured within that unchanged subset. Blur values use the historical blur-run clean answer; grayscale, masks, and text-only use the fresh clean answer generated with the same inference stack.

Model	**Clean acc.**	**Blur** 61×61	**Grayscale**	**Nested masks**	**Text-only**
		acc. / same	acc. / same	all-level same / acc.	acc. / same
Qwen-3B	65.0	37.2 / 46.9	61.1 / 79.6	48.1 / 74.4	34.8 / 43.8
Qwen-7B	68.1	34.1 / 42.6	63.7 / 83.4	53.4 / 76.6	37.8 / 45.1
LLaVA-7B	54.7	34.8 / 53.5	51.8 / 89.6	65.0 / 57.6	34.2 / 46.9

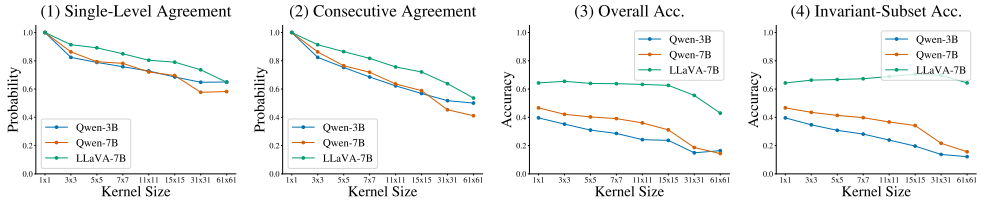
means over the twelve benchmarks listed in Section 4. HallusionBench Visual Dependent and Visual Supplement appear below as diagnostic partitions and do not enter the average as additional benchmarks. A MMStar example in which Qwen-7B gives the correct answer and remains invariant through all blur levels follows the complete figure series in Figure 16.



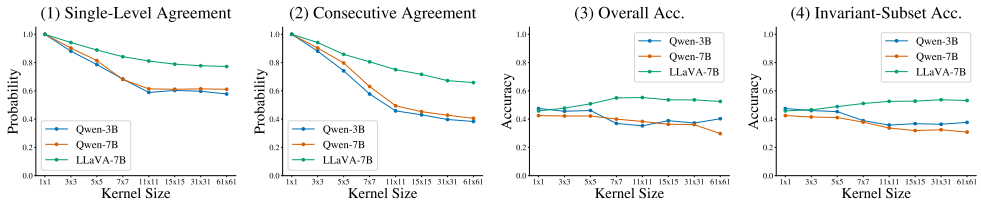
(a) RePOPE (Neuhaus & Hein, 2025).



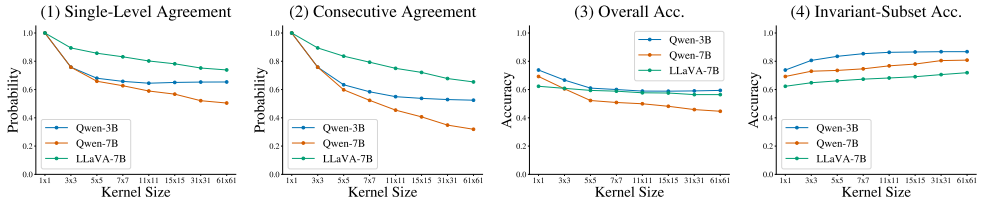
(b) HR-Bench (Wang et al., 2025b).



(c) HallusionBench Visual Dependent (Guan et al., 2024).

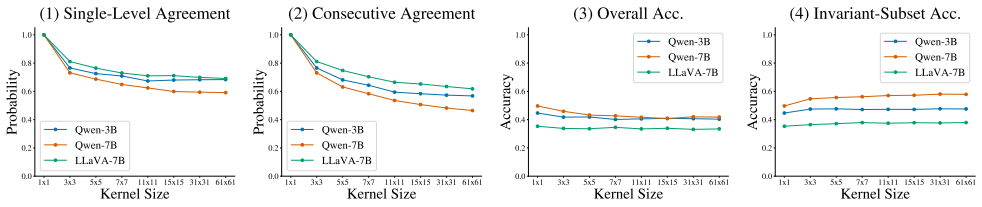


(d) HallusionBench Visual Supplement (Guan et al., 2024).

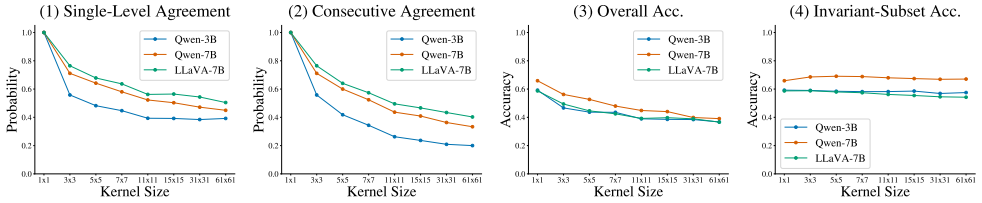


(e) AI2D (Kembhavi et al., 2016).

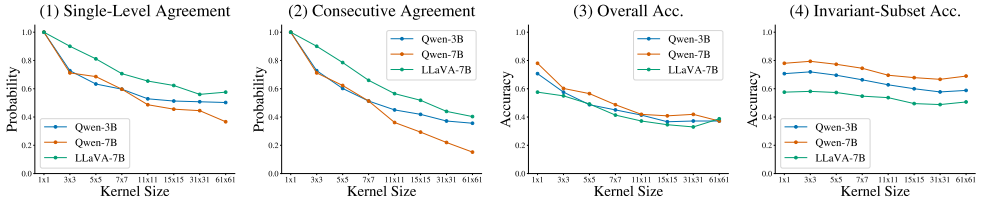
Figure 15: Per-benchmark trajectories under increasing Gaussian blur for Qwen-3B, Qwen-7B, and LLaVA-7B. From left to right, the panels show single-level agreement $\bar{E}_{\mathcal{D},k_i}$, where the current answer matches the original-image answer; consecutive agreement $\bar{C}_{\mathcal{D},k_i}$, where that match holds at every level through the current level; overall accuracy $\bar{\alpha}_{\mathcal{D},k_i}$; and invariant-subset accuracy $\bar{\gamma}_{\mathcal{D},k_i}$. Panels 3 and 4 report raw accuracies, and each curve represents one model.



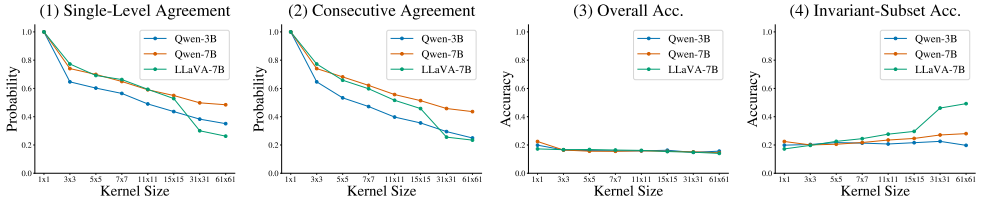
(f) MMMU (Yue et al., 2024).



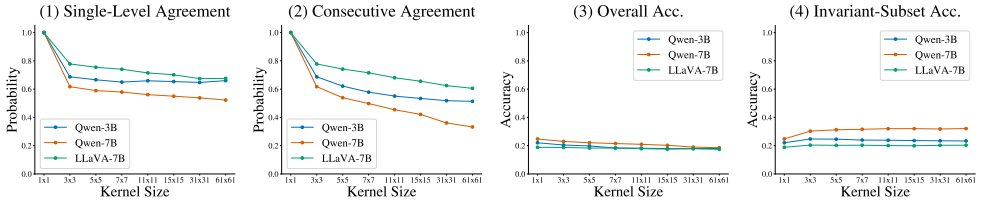
(g) RealworldQA (X.AI, 2024).



(h) V*Bench (Wu & Xie, 2024).

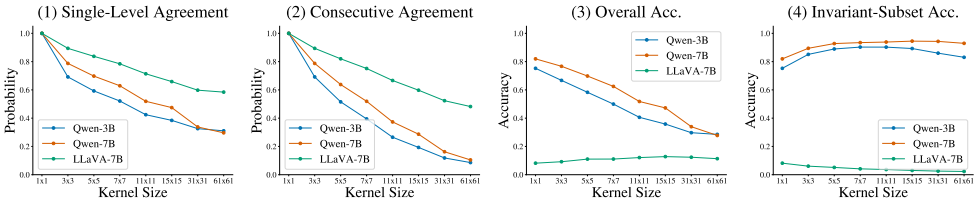


(i) VLMBias (Vo et al., 2025).

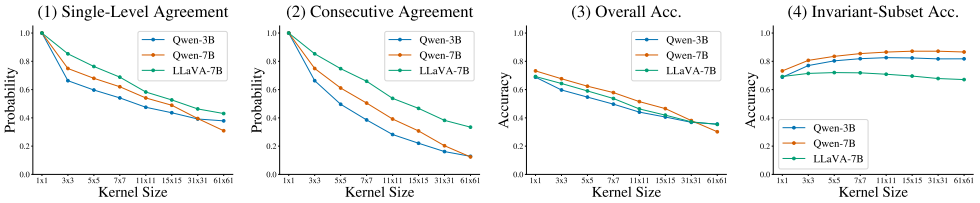


(j) BLINK (Fu et al., 2024).

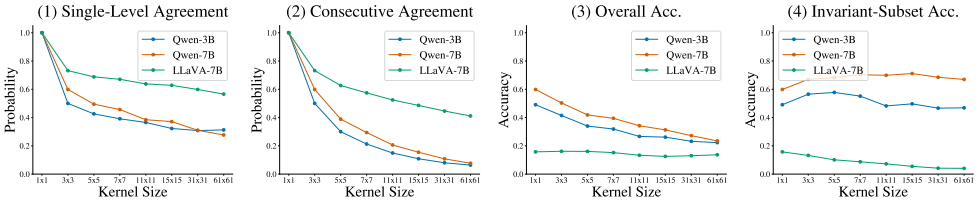
Figure 15: Per-benchmark trajectories under increasing Gaussian blur for Qwen-3B, Qwen-7B, and LLaVA-7B (continued). From left to right, the panels show single-level agreement $\bar{E}_{\mathcal{D},k_i}$, consecutive agreement $\bar{C}_{\mathcal{D},k_i}$, raw overall accuracy $\bar{\alpha}_{\mathcal{D},k_i}$, and raw invariant-subset accuracy $\bar{\gamma}_{\mathcal{D},k_i}$. Each curve represents one model.



(k) MMBench (Liu et al., 2024b).



(l) SEED-Bench (Li et al., 2023).



(m) MMStar (Chen et al., 2024).

Figure 15: Per-benchmark trajectories under increasing Gaussian blur for Qwen-3B, Qwen-7B, and LLaVA-7B (continued). From left to right, the panels show single-level agreement $\bar{E}_{\mathcal{D},k_i}$, consecutive agreement $\bar{C}_{\mathcal{D},k_i}$, raw overall accuracy $\bar{\alpha}_{\mathcal{D},k_i}$, and raw invariant-subset accuracy $\bar{\gamma}_{\mathcal{D},k_i}$. Each curve represents one model.

Ferry fares

Ferry	Car	Bicycle
Ocracoke	$15	$3
Mukilteu-Clinton	$7	$5
Seattle-Bremerton	$12	$8
Southport-Fort Fisher	$5	$2

Original Image



Blurred (61x61)

Question: How much higher is the fare for a car on the Mukilteu-Clinton ferry than on the Southport-Fort Fisher ferry? (Unit: $)
Options: A.0; B.1; C.2; D.4
Ground Truth: C. 2

VLM Responses for Different Kernel Sizes of Gaussian Blur:

1x1 (Original Image): To determine how much higher the fare for a car on the Mukilteu-Clinton ferry is compared to the Southport-Fort Fisher ferry, we need to follow these steps:
1. Identify the fare for a car on the Mukilteu-Clinton ferry. The fare for a car on the Mukilteu-Clinton ferry is $7.
2. Identify the fare for a car on the Southport-Fort Fisher ferry. The fare for a car on the Southport-Fort Fisher ferry is $5.
3. Calculate the difference between the two fares. Difference = $7 - $5 = $2.
Therefore, the fare for a car on the Mukilteu-Clinton ferry is $2 higher than on the Southport-Fort Fisher ferry. ANSWER: C

3x3: (similar to 1x1)

5x5: To determine how much higher the fare for a car on the Mukilteo-Clinton ferry is compared to the Southport-Fort Fisher ferry, we need to look at the fares for cars on both ferries from the table provided: Mukilteo-Clinton ferry: $5; Southport-Fort Fisher ferry: $3. Now, we calculate the difference between these two fares: $5 (Mukilteo-Clinton) - $3 (Southport-Fort Fisher) = $2. So, the fare for a car on the Mukilteo-Clinton ferry is $2 higher than on the Southport-Fort Fisher ferry. ANSWER: C

7x7: (similar to 5x5)

11x11: … From the image provided: The fare for a car on the Mukilteo-Clinton ferry is $4. The fare for a car on the Southport-Fort Fisher ferry is $2. Now, calculate the difference: Difference} = Fare on Mukilteo-Clinton - Fare on Southport-Fort Fisher. Difference = 4 - 2 = 2. So, the fare for a car on the Mukilteo-Clinton ferry is $2 higher than on the Southport-Fort Fisher ferry. ANSWER: C

15x15: (similar to 11x11)

31x31: … Let's assume the fares are as follows: Fare for a car on the Mukilteo-Clinton ferry: $5. Fare for a car on the Southport-Fort Fisher ferry: $3. Now, calculate the difference: Difference = $5 - $3 = $2. Therefore, the fare for a car on the Mukilteo-Clinton ferry is $2 higher than on the Southport-Fort Fisher ferry. ANSWER: C

61x61: (similar to 31x31)

(a) Invariant and correct.

Figure 16: A correct invariant MMStar response from Qwen2.5-VL-7B-Instruct. As blur increases, changing intermediate OCR values are rationalized into the same correct difference. The complementary incorrect example appears in Figure 8.