أكاديمية كاوست
KAUST ACADEMY

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

AnswerMap: Faithful Black-Box Spatial Interpretability for Vision Language Models

Fardows Adam, Duaa M. Tahir, Lama Alamoudi,
Sana Ammar, Atheer A. Alboloshi, Jory Albluey,
Mohamed Eltahir, Tanveer Hussain

fardoos.ahmad@kaust.edu.sa, duaa.tahir@kaust.edu.sa, lama.alamoudi@kaust.edu.sa,
sana.taffour@kaust.edu.sa, atheer.alboloshi@kaust.edu.sa, jory.albluey@kaust.edu.sa,
mohamed.hamid@kaust.edu.sa, hussaint@edgehill.ac.uk

Abstract

When a vision language model answers a question about an image, the answer arrives with no indication of which region produced it, or whether the image was used at all. This is the central question of spatial interpretability, and the standard tools are poorly matched to it. Attention and gradient read-outs require internals that deployed models do not expose. AnswerMap extracts the explanation from the model's answers instead of its machinery. The image is partitioned into K row and K column bands, the frozen model is asked one binary question per band, and the yes posterior is read from the first answer token. The outer product of the row and column posteriors is a query-conditioned spatial map, its expectation is a continuous location read-out, and its maximum grades how strongly the model commits to the region. Because the operator touches nothing but answer logits, it is training-free, applies to black-box models behind an API, and produces the map before any answer is generated. Because a self-report can be confabulated, we validate the map with two ground-truth-free tests that any spatial explanation should pass, agreement with the model's own pointing and causal necessity under matched-area deletion. The map passes both across three query distributions, four models, and two model families. The validated read-outs then transfer directly, auditing object hallucinations before generation, granting pointing to models that cannot emit coordinates, and letting a model fix its own failures by re-reading where its map points. Spatial interpretability for VLMs can be faithful, measured, and black-box at once.

Date: August 24, 2026
Code: https://github.com/sanataff/AnswerMap

1 Introduction

Vision language models answer questions about images fluently and confidently, whether or not the image played any role in the answer. For hallucination auditing, for grounding-sensitive applications such as medical imaging, and for basic trust, the question that matters is simple. *Which part of the image did this answer come from, if any.* Answering it for a given model is the core problem of VLM spatial interpretability.

The standard tools are poorly matched to the question. Attention read-outs are the default lens, yet they require access to internals, and recent work shows that VLM attention is a distorted signal. Image tokens receive a vanishing share of attention in deep layers [1], high attention concentrates on visual sink tokens whose removal does not change the output [2], and averaged text-to-image attention maps are noisy enough that grounding methods must first search for the few heads that localize [3]. Gradient attribution needs backpropagation through billions of parameters and transfers poorly to VLM architectures [4]. Black-box perturbation methods avoid internals but pay for it with thousands of masked queries per map [5], and the one VLM heatmap method validated by deletion requires gradients and the generated answer before it can explain anything [4]. Across all of these, faithfulness is presumed rather than tested.

Our approach.. We ask the model itself. AnswerMap lays K row bands and K column bands over the image and asks the frozen model $2K$ yes/no questions of the form "is the {query} in this band". The yes posterior is read from first-token logits, never from generated text. The outer product of the row and column marginals gives a $K \times K$ query-conditioned map, its normalized expectation gives a continuous sub-cell location, and its maximum gives a scalar confidence. The read-out is training-free, uses no weights, gradients, or attention, works before any answer is generated, and costs 16 queries for an 8×8 map.

A self-report, however, can be confabulated. Language models produce fluent reasoning that systematically misrepresents the true cause of their predictions [6], so a map built from the model's own answers inherits that suspicion and must be tested, not trusted. We therefore evaluate AnswerMap with two tests that any spatial explanation for a VLM should pass, neither of which needs ground truth.

1. **Agreement.** If the map reads the model's spatial belief, the model's own generated point should land where the map is high. We score the full map at the generated point with the standard saliency metrics NSS and AUC [7].
2. **Causal necessity.** If the map marks the evidence, deleting the map's region should change the answer, and deleting a matched random region should not. This is the deletion principle [5] applied to free-form VQA.

Findings.. The probe map passes both tests and the standard tools do not. On RefCOCOg it scores NSS 1.36 at the model's generated point while attention scores below chance at every layer, including the layer a per-layer sweep hands it. Deleting the map's region flips 53.4% of TextVQA answers against 3.3% for the matched random control. The ordering holds across three query distributions, four models, and two model families (§4), and the strongest baseline, gradient-weighted relevancy, stays behind on both tests while needing white-box gradients.

Contributions.. Our contributions are methodological.

- **The AnswerMap operator.** Query-conditioned spatial maps read from first-token answer posteriors of a frozen VLM. Black-box, training-free, answer-free, $2K$ queries for a $K \times K$ map, a continuous expectation read-out, and a multigrid product that lifts the rank-1 limitation of a

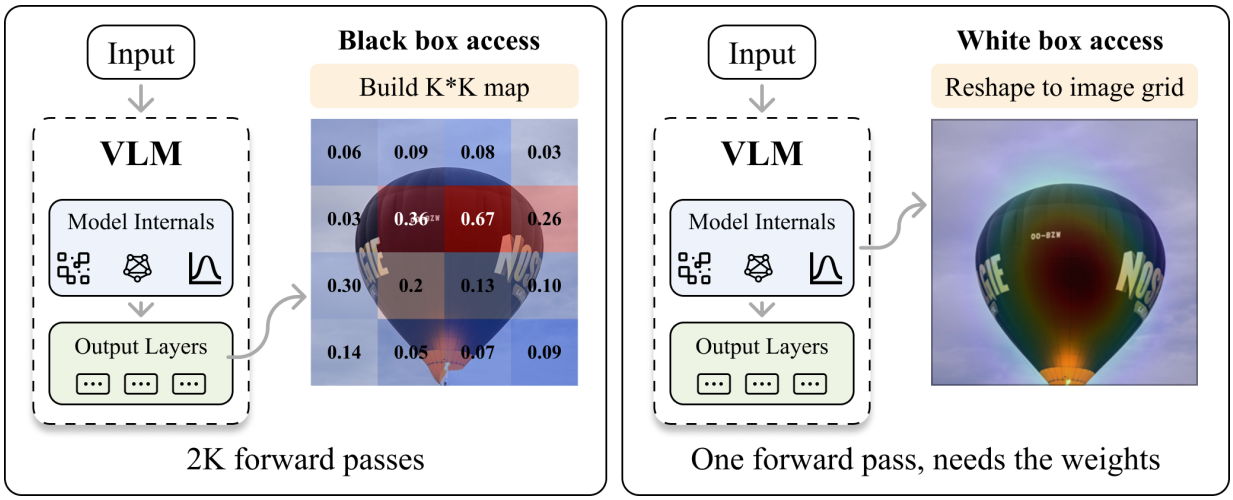


Figure 1 Two ways to ask a frozen VLM where it looked, judged at the model's own generated point (red pin). Dissecting internals yields a below-chance attention map. Asking the sealed model 2K band questions yields a map that passes both faithfulness tests (Tables 2 and 3).

single grid (§3).

- **A two-test faithfulness protocol** for VLM spatial explanations. Agreement with the model's own pointing, scored with saliency canon metrics, and causal necessity under matched-area deletion. Both are ground-truth-free, so they apply to any image, any query, and any explanation method (§3.2).
- **A graded grounding signal and an image-reliance label.** The map maximum as a scalar grounding-strength read-out, and whole-image deletion within the same protocol as a per-question label of whether the model used the image at all (§4). (§2).

2 Related Work

Attention as explanation.. Attention weights are the most common lens on where a transformer looks, both raw and propagated through the stack by rollout [8]. For VLMs the signal is known to be distorted. Attention to image tokens collapses after the first layers, with image tokens receiving a fraction of a percent of the attention given to system prompt tokens in deep layers [1]. High attention concentrates on visual sink tokens that are irrelevant to the text, and removing those tokens does not change the output [2]. Averaged text-to-image maps are sparse and noisy, and grounding from attention requires first finding the few heads that localize [3]. In the text domain, intervention studies reached the same verdict earlier, attention is at best a noisy indicator of importance [9, 10]. What is missing for VLMs is a direct faithfulness test of the attention read-out as an explanation. We run that test and attention fails it.

Gradient and relevancy attribution.. Grad-CAM [11], integrated gradients [12], and transformer relevancy propagation [13] need white-box access and backpropagation, fail basic sanity checks in some settings [14], and transfer poorly to modern VLM architectures, where relevancy baselines score poorly on VLM heatmap benchmarks [4]. The 2025 state of the art in this family conditions on the generated answer and reads internals, TAM combines activations with causal filtering [15] and GLIMPSE fuses gradient-weighted attention across heads and layers [16]. Both explain tokens the model already produced. Our read-out is forward-only, touches nothing but answer logits, and exists before generation. We carry the family's standard representative, T-MM relevancy [13],

Table 1 Spatial explanation methods for VLMs along deployment axes. Queries are forward passes for one 8×8 map.

Method	Black box	Calib.-free	Answer-free	Queries	Continuous	Tested
Attention (any layer)	×	✓	✓	1	×	×
Attention rollout [8]	×	✓	✓	1	×	×
Localization heads [3]	×	×	✓	1	×	×
Grad-CAM, IG [11, 12]	×	✓	×	1 to 50	×	×
Occlusion [17]	✓	✓	×	65	×	partial
RISE [5]	✓	✓	×	4000+	×	×
Mask optimization [4]	×	✓	×	optim.	✓	✓
F-LMM [20]	×	×	✓	1	✓	×
GridProbe [19]	✓	✓	✓	16	×	×
AnswerMap (ours)	✓	✓	✓	**16**	✓	✓

as a baseline in both faithfulness tests.

Black-box perturbation saliency.. Occlusion [17] and RISE [5] estimate importance by masking inputs and watching outputs, which makes them black-box but expensive. RISE uses 4000 to 8000 masked forward passes for a map on a 7×7 mask grid. The closest method to our setting optimizes a deletion-validated heatmap for VLM answers [4], but the optimization needs gradients, needs the generated answer before it starts, and takes seconds per sample. Our map needs 16 closed-form queries, no gradients, and no answer.

Faithfulness of stated reasoning.. Deletion and insertion curves [5], ROAR [18], and sanity checks [14] established that explanations must be tested against interventions. In parallel, chain-of-thought studies showed that a model's stated reasoning can systematically misrepresent the true cause of its prediction [6]. A spatial self-report elicited by questioning the model inherits exactly that suspicion. Our two-test protocol is the answer, a stated spatial belief is only called an explanation after it predicts the model's own pointing and survives a deletion test.

Probing answer space and VLM pointing.. GridProbe uses answer-space probes over a grid to select video frames by the peak of the response [19]. We differ in object and in claim, we read a continuous map with an expectation read-out, we validate it as an explanation with agreement and deletion tests, and we use its maximum as a grounding-strength signal. F-LMM shows that word-pixel correspondences exist in frozen LMM attention but require trained decoding layers to extract [20], and localization heads require a calibration corpus to identify [3]. Both support our framing, the grounding signal in attention is real but is not available as a zero-shot read-out.

Positioning.. Table 1 compares along the axes that matter for a deployable explanation. We evaluate inside the class of zero-shot, calibration-free, answer-free methods. Methods that need a calibration corpus [3] or trained components [20] solve a different problem and appear in the table for completeness.

3 Method

AnswerMap separates two concerns that spatial interpretability usually conflates, eliciting a spatial self-report from a frozen model, and deciding whether that self-report deserves to be called an explanation. The operator (§3.1) does the first with $2K$ answer-space queries. The two-test protocol (§3.2) does the second with interventions on the model's own behavior.

3.1 The AnswerMap operator

Let x be an image and q a free-text query. We partition x into K horizontal bands $\{r_i\}_{i=1}^{K}$ and K vertical bands $\{c_j\}_{j=1}^{K}$. For each band we reassemble its tiles into a single strip image, show the frozen model that strip alone, and ask one binary question, "is $\{q\}$ in this band". The answer is never read from generated text. We take the first answer-token logits and compute the yes posterior over the closed label set,

$$p_i^{\text{row}} = \frac{\exp z_{\text{yes}}}{\exp z_{\text{yes}} + \exp z_{\text{no}}}, \qquad i = 1, \ldots, K, \tag{1}$$

and likewise p_j^{col} for columns. The two marginals are combined into a $K \times K$ map by the outer product

$$M_{ij} = p_i^{\text{row}}\, p_j^{\text{col}}. \tag{2}$$

Two functionals of M serve two distinct roles, and we validate each role separately (§4).

- **Expectation (location).** The normalized expectation $\mathbb{E}_M[\ell] = \left(\sum_{ij} \tilde{M}_{ij}\, x_j,\ \sum_{ij} \tilde{M}_{ij}\, y_i\right)$ with $\tilde{M} = M / \sum M$ is a continuous sub-cell location, unlike an argmax over cells.
- **Maximum (strength).** $\max_{ij} M_{ij}$ is a scalar confidence. High values mean the model commits to a region.

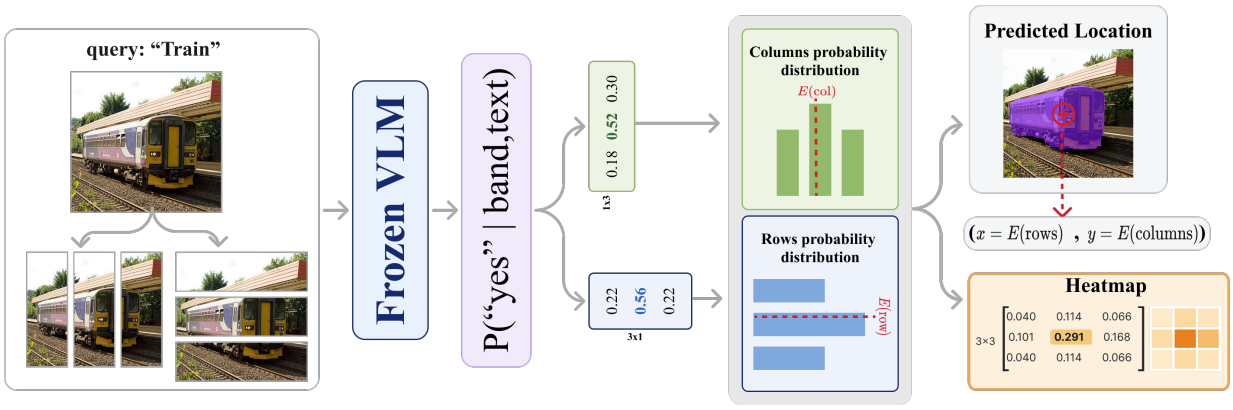
Multigrid product.. A single grid makes M a rank-1 outer product, which carries no more information than its two marginals and cannot represent two disjoint modes. We therefore probe several coarse grids $K \in \{2, 3, 5\}$ and multiply the upsampled maps elementwise. Each grid is an independent coarse view of the same belief, and the elementwise product is a product of experts, a soft intersection under which a location survives only if every view endorses it. Coprime grid sizes share no band boundaries, so the views stay decorrelated, whereas nested sizes such as $\{2, 4, 8\}$ re-ask implied questions and double-count evidence. The result is a higher-rank, non-separable map from purely separable probes. The single grid is the default throughout, the multigrid variant and its combination rule are evaluated in §6.

Cost and access.. An 8×8 map costs $2K$=16 forward passes. The bands are presented as a single concatenated strip image, so one forward scores one band question at full band resolution. For comparison, occlusion needs 65 passes for the same resolution and RISE uses 4000 to 8000 masks [5, 17]. Measured wall-clock and memory per map, for every method at three model scales, are in Table 9 (§6), the probe runs at 0.3 to 1.2 seconds per map on one A100. The only access requirement is the top-k first-token logprobs, which several closed APIs expose, so the operator applies to models with no visible weights, gradients, or attention. The map exists before any answer is generated, so a system can audit grounding first and generate second.

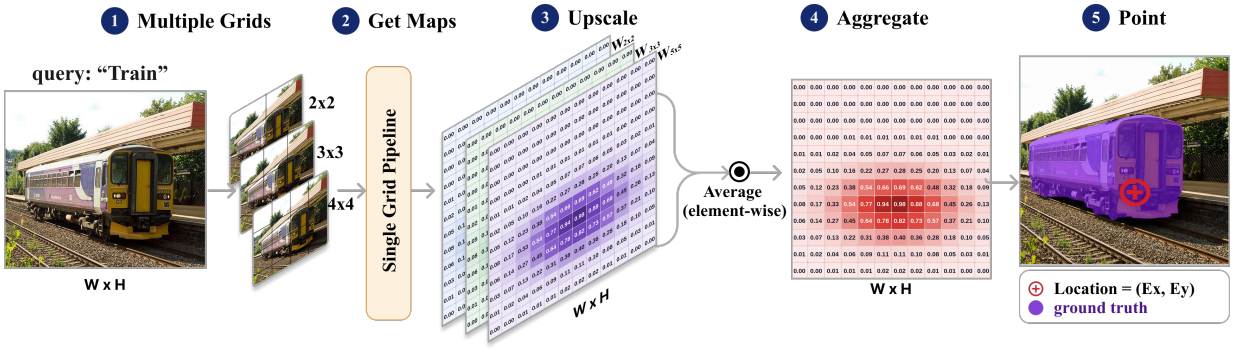
3.2 The two-test faithfulness protocol

A map elicited from the model's own answers could be confabulation [6]. We therefore treat faithfulness as an empirical property to be measured, with two ground-truth-free tests. Both apply unchanged to any spatial explanation method, which makes the protocol reusable beyond this paper.

Test 1, agreement with the model's own pointing.. We ask the same frozen model to point at the query with its native grounding prompt and parse the coordinate in its native convention. If the map reads the model's spatial belief, the generated point g is a sample from that belief and should land where the map is high. We score the full map at g with the standard saliency metrics [7].



(a) Single Grid Pipeline



(b) Multi-Grid Fusion

Figure 2 The AnswerMap operator. (a) First figure description. (b) Second figure description. K row and K column yes/no band questions fill a $K \times K$ map through the outer product. The crosshair is the expectation read-out, and the flag is the maximum. The model stays sealed, first-token logits are the only thing that crosses.

NSS, the z-scored map value at g, with chance at 0. AUC, the probability that the map ranks the true cell above a random cell, with chance at 0.5. Both score the map directly and do not reduce it to one point, so multimodal maps are treated fairly. We additionally report the distance between the map expectation and g as a fraction of the image diagonal, against an image-centre floor, and the Pearson correlation of the coordinates, which is shift-invariant so centre bias cannot inflate it.

Test 2, causal necessity by deletion.. For each question the model answers correctly with the full image, we delete the map's claimed region, the top-mass cells covering 12% of the image, by setting them to grey, and re-ask. The control deletes a random region of matched area, disjoint from every method's region. The metric is the answer change rate. A faithful map names regions whose deletion breaks the answer, and a random region should not [5]. Deleting the whole image in the same way labels each question by image reliance, which separates answers that used the image from answers produced from priors (Appendix C).

Comparison class.. All methods in the main tables are zero-shot, calibration-free, and answer-free. Attention is represented by its strongest member. We sweep every decoder layer on a held-out split, hand attention its best layer, and carry that layer into both tests (§4.1).

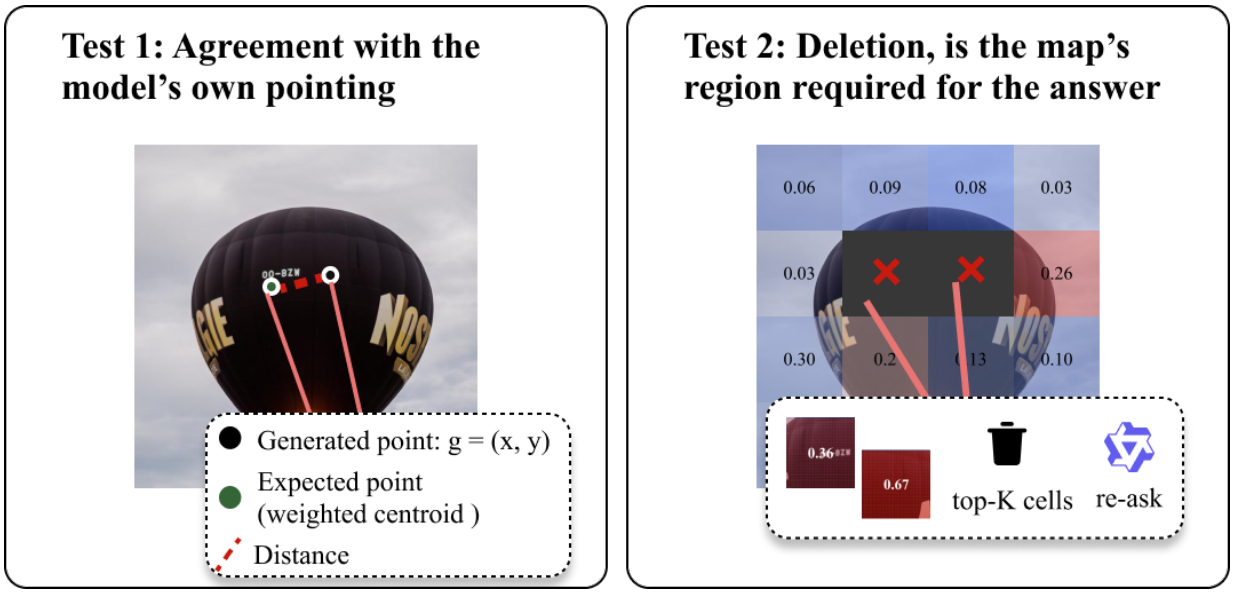


Figure 3 The two-test faithfulness protocol. Left, agreement, score the full map at the model's own generated point (Table 2). Right, deletion, remove the map's top-mass region against a matched random region and re-ask (Table 3).

4 Faithfulness Results

Setup.. All experiments run on a single NVIDIA A100 80GB in bfloat16 with HuggingFace Transformers, greedy decoding, and no training of any kind. The primary model is frozen Qwen3-VL-4B-Instruct [21], with replications on Qwen3-VL-8B, Qwen3-VL-30B-A3B, and InternVL3.5-8B [22]. Probes use K=8 unless noted, probe images are resized to a 512-pixel long side, generation and deletion use 1024. White-box baselines require eager attention, all other runs use fused attention kernels. Open-ended answers are scored by normalized string match against the annotator answers (lowercase, punctuation stripped, containment in either direction counts). In the agreement test, examples whose generated point cannot be parsed are excluded, in point-in-region evaluation a parse failure is scored as a miss. Random regions, random maps, and shuffles use fixed per-example seeds, all runs are deterministic. Exact prompt templates and per-experiment configuration are listed in Appendix A.

4.1 Test 1, agreement with the model's own pointing

Table 2 reports agreement across three query distributions, referring expressions (RefCOCOg, n=800), appearance-only expressions (RefCOCO+, n=800, location words banned by construction), and anomaly descriptions (CAVE, n=334). The ordering is identical on all three. The probe map leads, the equal-cost multigrid recipe $\{3, 5\}$ leads further, every attention read-out sits at or below chance, occlusion is mid-tier, and the random floor lands at chance every time. On RefCOCOg the probe expectation also beats the always-centre distance floor by 31% while attention at any layer does not beat it at all, and on centre-biased CAVE the distance floor equals the random map, which is exactly why the chance-anchored full-map metrics carry the evaluation. Per-metric detail including distance and coordinate correlation is in Appendix C.

Attention got its fair shot, and why it is below chance.. Reading only the last layer would invite the objection that deep-layer attention to image tokens is starved [1], so we swept all 36 decoder layers on a held-out split (n=200) and handed attention its best layer (15) in every table, the selection

Table 2 Test 1. NSS and AUC of each map at the model's own generated point (chance 0 and 0.5), across three query distributions.

Method	RefCOCOg NSS	RefCOCOg AUC	RefCOCO+ NSS	RefCOCO+ AUC	CAVE NSS	CAVE AUC
Probe map (K=8)	1.36	0.824	1.44	0.817	**1.73**	0.828
Probe map ({3, 5}, equal cost)	**1.56**	**0.853**	**1.52**	**0.840**	1.70	**0.845**
Attention (best layer)	−0.22	0.380	−0.18	0.432	−0.21	0.409
Attention (last layer)	−0.15	0.464	−0.11	0.490	−0.17	0.456
Attention rollout	−0.16	0.457	−0.15	0.475	−0.13	0.543
Relevancy T-MM [13]	0.33	0.656	–	–	–	–
Occlusion	1.04	0.649	1.14	0.654	1.12	0.644
Random map	−0.05	0.486	−0.02	0.493	−0.01	0.497

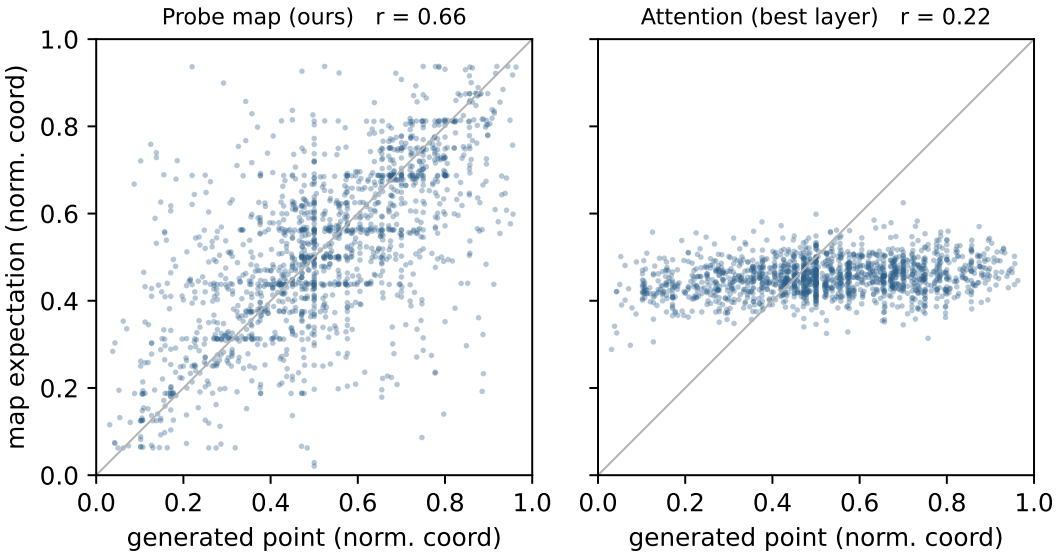


Figure 4 Map expectation against the model's generated point (RefCOCOg). The probe tracks the diagonal, best-layer attention is a flat band regardless of where the model points.

generalized to the full split (Figure 5). Below-chance AUC then has a mechanical reading, attention mass concentrates on sink tokens whose content is irrelevant to the text [2], so its expectation is nearly constant across images, the flat band in Figure 4. The finding has precedent, raw attention and rollout score near-uselessly against ground-truth masks in the current state of the art for MLLM explanation [15], our contribution is the chance-anchored quantification at the model's own output. Localization heads [3] are excluded from the comparison class because they require a calibration corpus, a per-example variant performs at noise level, consistent with their own premise (Appendix C).

The two above-chance baselines.. Occlusion, run answer-free at the same grid, is the only black-box competitor above chance and the probe beats it on both metrics on all three benchmarks at a quarter of its query cost, so the advantage is the answer-space read-out, not black-box access per se. Gradient-weighted relevancy (T-MM, all layers) is the strongest white-box baseline, its diffuse map has a well-placed centroid (coordinate correlation 0.706, above the default single grid's 0.665 though not the recipe's 0.711, Appendix C) while its mass placement stays far behind, it needs gradients through all 36 layers.

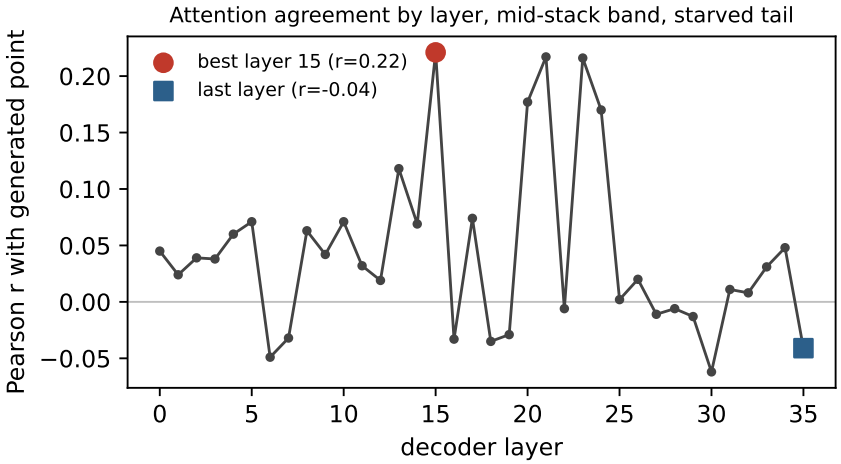


Figure 5 Agreement of each decoder layer's attention map with the model's point (held-out split). A spiky mid-stack band peaks at layer 15, the last layer is negative, reproducing Chen et al. [1] and Kang et al. [3] on a new model family.

Table 3 Test 2. Answer change rate (%) when a method's top-mass region (12% of cells) is deleted, on questions answered correctly with the full image. Right columns, corruption and question-type robustness.

Method	TextVQA	TextVQA (blur)	GQA	GQA (non-bin.)
Probe map (ours)	**53.4**	**54.0**	**22.4**	**26.2**
Relevancy T-MM (all layers)	36.8	–	17.1	–
Attention (best layer, ℓ=15)	18.8	19.3	3.0	2.2
Attention (last layer)	18.3	19.6	4.2	4.9
Attention rollout	9.5	9.5	3.4	5.3
Random region (control)	3.3	3.3	2.7	2.2

4.2 Test 2, causal necessity by deletion

Table 3 reports the deletion test on TextVQA [23], where answers are text at a location, and GQA [24], where answers are objects and relations with no reading involved. Three readings. First, the two datasets close the obvious escape, the gap is not an artifact of reading text since it persists on GQA, where every attention read-out is statistically indistinguishable from the random control. Second, the result survives both robustness columns, blur instead of the grey blank changes nothing, and restricting GQA to non-binary questions sharpens the probe while attention stays at the control. Third, the strongest baseline is again gradient-weighted relevancy, well above attention and well below the probe.

4.3 Scale and model family

Table 4 replicates both tests at 8B and 30B and on a second model family. The probe is stable everywhere with no per-model tuning. The attention baseline is not, on the 8B a mid-stack band emerges (its own sweep selects layer 24 against 15 on the 4B) while the last layer stays negative, so the below-chance verdict is a property of the 4B scale, and a read-out that must be re-swept per model and swings this much with scale is a fragile foundation for interpretability. On InternVL the black-box suite ran with no adapter work of any kind, the portability claim demonstrated rather than asserted.

Table 4 Both tests across scale and family. Agreement at the model's own point, deletion answer change rate (%), attention uses each model's own swept best layer.

	Agreement (probe)		Best baseline	Deletion	
Model	NSS	AUC	NSS	Probe	Random
Qwen3-VL-4B	1.36	0.824	occlusion 1.04	53.4	3.3
Qwen3-VL-8B	1.49	0.823	attention 0.33	51.1	2.7
Qwen3-VL-30B-A3B	1.39	0.813	occlusion 1.03	45.0	3.2
InternVL3.5-8B	1.03	0.768	occlusion 1.03	50.2	4.1

The maximum grades causal strength.. Binning examples by the map maximum, the probe-region flip rate rises monotonically from the lowest to the highest tercile in every configuration we tested, five of five (Appendix C), and as a scalar the maximum predicts the flip with AUC up to 0.677 on the 30B. The confidence the operator reports is causally meaningful, act on the map only when the maximum is high. One scope note, the gradient flattens on the non-binary GQA subset. Separately, whole-image deletion labels each question by image reliance, and the probe-region drop tracks that label on every model (Appendix C).

Why the probe works where attention does not.. The probe never inspects the machinery, it queries the function. Each band question forces the model to commit its answer distribution about one region, in its own language, so the read-out inherits whatever visual competence the model actually uses to answer. Attention is a by-product signal that serves computation, not communication. The verdict of both tests, the map is a faithful read-out of the model's spatial belief, established behaviorally and causally, without ground truth, weights, or gradients.

5 What the Map Buys You

The two tests validated two read-outs, the expectation for location and the maximum for strength. Each application below spends one of them on a task, none needs labels, training, or internals, and each inherits the faithfulness evidence of §4 instead of asserting its own.

5.1 Object hallucination auditing

The map maximum reports real spatial commitment (§4.3), and its first external use is the task where that quantity is exactly what is in question, object existence. On the full POPE benchmark [25] (9,000 probes, 4,006 yes claims by the model, 223 of them hallucinated), a low maximum flags a hallucinated object with ROC-AUC 0.833, the map for an absent object has nowhere to peak. The audit costs 16 queries, needs no labels and no second model, and runs before generation. The deletion signature (Table 5) separates the claim types from a second angle, grounded yeses die only when their own region is deleted, hallucinated yeses are fragile under any perturbation, nothing in the image was carrying them.

5.2 Coordinate-free pointing for coordinate-blind models

Several capable models cannot reliably SAY where. InternVL3.5's pointing interface fails to parse on 45% of RefCOCOg queries, and the medical Lingshu-7B [26] fails on more than half, while both carry probeable spatial beliefs (§4). The map expectation converts that belief into a point with no coordinate emission at all, giving pointing to models whose grounding interface is weak or absent. Table 7 calibrates the read-out on a strong pointer, the expectation recovers most of Qwen3-VL's own generated point (67.5 against 94.5 point-in-box), so what the expectation extracts

Table 5 Hallucination audit on POPE. Flip rate (%) of yes claims under deletion. Low-peak flags hallucination at AUC 0.833.

Claim	Map region	Random
Grounded	41.0	1.4
Hallucinated	69.5	11.2

Table 6 Self-conditioning on TextVQA. Accuracy (%) and the share of receiver failures fixed.

Input	Acc.	Fixed
Image alone	91.6	–
+ own map crop	**92.4**	**50%**
+ random crop	89.0	38%

Table 7 Point-in-region accuracy (%). On the coordinate-blind Lingshu-7B the map expectation beats the model's own generated point. CAVE n=334, RefCOCOg n=400, Lingshu on RefCOCOg n=300.

Read-out	CAVE [27]	RefCOCOg	RefCOCOg (Lingshu)
Generated point	77.8	94.5	46.7
Probe expectation (ours)	71.3	67.5	**57.7**
Attention (best layer)	57.2	40.0	40.0
Image-centre prior	58.1	40.5	39.3
Random point	24.3	24.8	24.3

is the model's belief, not an artifact. On the coordinate-blind regime the application delivers, with Lingshu-7B on RefCOCOg the probe expectation reaches 57.7 point-in-box while the model's own generation manages 46.7 (30% of its pointing attempts fail to parse at all), the model that cannot say where points better through its map than through its mouth, with attention at the centre prior (40.0 against 39.3) as usual.

5.3 Self-conditioning

The map improves the answers of the model that produced it. The recipe is probe first, then look again where your own map points, the top-mass region cropped from the full-resolution image and appended as a close-up. On TextVQA (n=500, Table 6) the gain concentrates where it matters, half of the questions the model would otherwise get wrong are fixed, against 38% for a random crop through the same mechanics. The recipe is training-free, needs no second model and no internals, and the map exists before any answer is generated. The white-box ancestors of this idea crop by attention or gradients, the probe version runs through an API. We also tested handing a larger model's map to a smaller receiver through four channels, it never beat the receiver's own map, so the practical recipe is self-use.

6 Ablations

This section isolates the operator's design choices, the read-out functional, the query budget, and the fusion rule, and closes with the measured cost per map.

Expectation versus argmax.. The expectation read-out is continuous and sub-cell, an argmax is limited to cell centres. The two functionals serve different validated roles, expectation for location (the agreement test) and maximum for strength (the deletion grading), and we keep them separate throughout.

Query budget and the shape of faithfulness.. How should a query budget be spent. Both spending strategies trace an inverted U on a fixed 296-example split (Figure 6), and the reason is the principle the operator is built on, a band question is only informative while the band retains enough context

for an easy yes/no. Refining a single grid rises to a narrow peak and collapses, NSS 0.75 at K=2, 1.46 at K=6 (12 queries), 1.14 already at K=8, a drift through 0.96 at K=28, and 0.22 at K=40 (80 queries), where the bands are slivers. Composing coprime grids rises to 1.50 at $\{3, 5\}$ (16 queries) and stays flat-topped, 1.50 at $\{2, 3, 5\}$, 1.37, 1.25, and still 1.27 with AUC 0.831 at $\{2, \ldots, 13\}$ (82 queries). The two ends tell the story, at the default 16-query budget composition beats refinement 1.50 against 1.14, and at 80 queries the contrast is 1.27 against 0.22. The peaks themselves are comparable (1.46 against 1.50), so the value of composition is *robustness*, the single-grid optimum is a spike that nothing identifies a priori, the multigrid top is flat across a five-fold budget range. The practical recipe is $\{3, 5\}$, a drop-in replacement for the single K=8 grid at identical cost, and even K=4 at 8 queries reaches NSS 1.25, which matters where queries are billed. No baseline has a comparable knob, and occlusion sits on a dominated frontier (65 queries, NSS 1.04). Two honesty notes bound the claim. Adjacent points differ by less than sampling noise at this split size, so the argument rests on the curve ends and on cross-budget consistency, and the full split confirms both headline configurations, K=6 at NSS 1.52 and $\{3, 5\}$ at NSS 1.56 with AUC 0.853, the strongest agreement numbers in the paper at the default cost (Table 2). And the gain is agreement-side only, on the deletion test the multigrid ties the single grid (52.3 against 53.4) because both select nearly the same top-mass cells.

The combination rule, and why coprime.. Both design choices behave as the product-of-experts reading predicts (§3.1), the controls are in Table 8. Intersection beats union beats single-view, so the fusion is doing real work. And the views must be independent, the nested set spends more queries for less faithfulness because nested grids re-ask implied questions and the product double-counts correlated evidence.

Table 8 Fusion and independence controls, NSS at the model's point. Intersection beats union beats single view, nested grids waste queries.

Configuration	Queries	NSS
Product, coprime $\{3, 5\}$	16	**1.50**
Product, coprime $\{2, 3, 5\}$	20	**1.50**
Mean, $\{2, 3, 5\}$	20	1.39
Max, $\{2, 3, 5\}$	20	0.89
Product, nested $\{2, 4, 8\}$	28	1.25

Cost per map.. Table 9 reports wall-clock and peak memory per explanation map. Two honest readings. First, white-box read-outs are cheaper per map when available, a single forward beats 16, and at our resolution eager attention fits in memory even at 30B, so the argument for the probe is access and faithfulness at modest cost, not raw speed. Second, the probe stays practical at every scale (0.3 to 1.2 seconds per map), costs 4.7× less than occlusion, the other black-box method, and is the only method needing nothing but answer logits. T-MM additionally carries gradient memory, up to 47% extra at 4B.

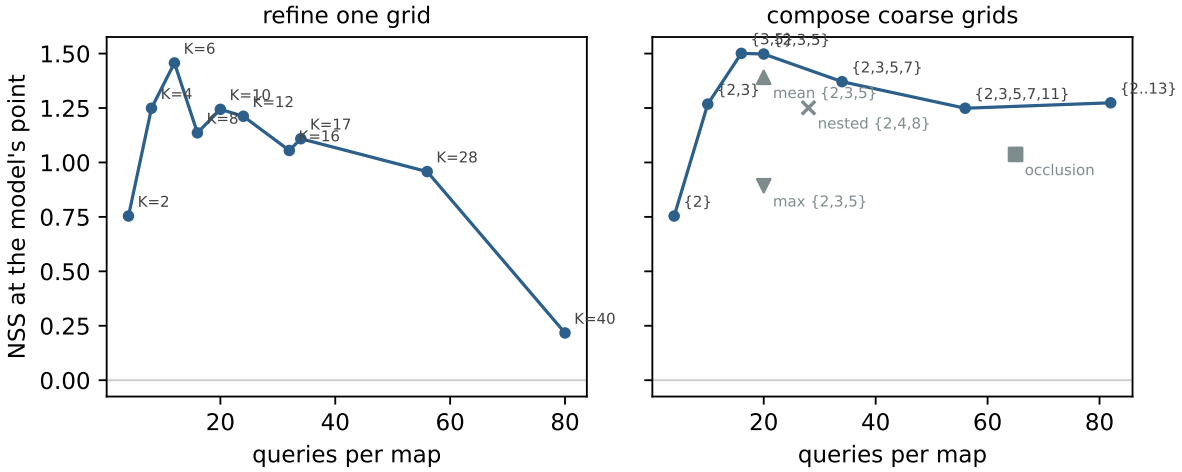


Figure 6 Faithfulness against query budget. Refining one grid (left) peaks narrowly and collapses, composing coprime grids (right) is flat-topped. Grey markers, the fusion and independence controls.

Table 9 Seconds per explanation map on one A100, 20 samples at image side 512.

Method	Queries	Access	4B	8B	30B
Probe map (ours)	16	logits only	0.41	0.34	1.25
Occlusion	65	logits only	1.80	2.81	5.90
Attention (best layer)	1	internals, eager	0.07	0.08	0.38
Attention rollout	1	internals, eager	0.59	0.22	1.15
Relevancy T-MM	1	internals, eager, gradients	0.18	0.19	1.35

7 Limitations and Future Work

Scope of the operator.. The operator asks whether a named referent is in a band, so it presumes something localizable to point at. A single grid is rank-1 and cannot represent two disjoint candidates, the multigrid product lifts the rank (§6) but genuinely multi-object and negated queries have no dedicated evaluation yet, and scene-level questions such as counting or global layout are covered only indirectly by probing their parts. Resolution is bounded by band answerability, thin bands make every question uninformative and quality collapses (the inverted U of §6), so a target small relative to the finest usable band is localized only to band scale, the sub-cell expectation interpolates rather than restores lost resolution.

Scope of the evidence.. Grey and blurred deletions are not natural images. The comparisons are matched across methods and the ordering is stable under both corruption operators, but absolute flip rates are properties of the protocol, not of deployment inputs. The raw yes posteriors are used only relatively and their calibration is unquantified, an ECE analysis over band questions with known answers would bound how literally they can be read. Coverage is four models, two families, and English benchmarks, and the API setting that motivates the black-box design is architecturally supported (only top-k first-token logprobs are required) but not yet demonstrated on a closed model. Faithfulness here is an empirical property established per model by the two tests, not a guarantee, a model trained to misreport could in principle pass agreement, though the deletion test remains causal and correspondingly harder to game.

Cost, and what the map is not.. A map costs $2K$ forward passes per image-question pair and does not amortize across questions about the same image. At 0.3 to 1.2 seconds per map (Table 9) this is practical for auditing and selective use, not for dense per-frame deployment. The probe reads belief, it does not add capability, self-conditioning gains concentrate on the model's failures but are modest at near-ceiling baselines (§5.3).

Future work.. Four directions follow directly. A closed-API demonstration, the operator needs nothing an API does not already expose. Video, where bands become temporal windows and the faithfulness protocol carries over unchanged. A multi-object and negation suite, the setting the multigrid product was built for. And probe-guided regeneration, the hallucination audit (§5.1) flags unsupported claims before they are generated, closing the loop from detection to mitigation. More broadly, the two-test protocol applies to any spatial explanation of any VLM, and a standing benchmark built on it would let explanation methods compete on faithfulness rather than plausibility.

8 Conclusion

We asked where a VLM's answer comes from and answered it without opening the model. AnswerMap reads a query-conditioned spatial map from first-token answer posteriors, in 16 queries, before any answer is generated. We refused to trust the self-report and tested it twice, agreement with the model's own pointing and causal necessity under deletion. The map passes both tests across three query distributions, four models, and two model families. The model's own attention fails both, on our primary model it scores below chance at the model's generated point at every layer, and at every scale its regions matter far less than the probe's under deletion. The practical consequence is a faithful, graded, black-box read-out of visual grounding that works through an API, and a reusable two-test protocol that any future spatial explanation for VLMs can be held to.

References

[1] Liang Chen et al. "An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models". In: *European Conference on Computer Vision (ECCV)*. 2024.

[2] Seil Kang et al. "See What You Are Told: Visual Attention Sink in Large Multimodal Models". In: *International Conference on Learning Representations (ICLR)*. 2025.

[3] Seil Kang et al. "Your Large Vision-Language Model Only Needs a Few Attention Heads for Visual Grounding". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025.

[4] Xiaoying Xing et al. *Where do Large Vision-Language Models Look at when Answering Questions?* 2025. arXiv: 2503.13891 [cs.CV].

[5] Vitali Petsiuk, Abir Das, and Kate Saenko. "RISE: Randomized Input Sampling for Explanation of Black-box Models". In: *British Machine Vision Conference (BMVC)*. 2018.

[6] Miles Turpin et al. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023.

[7] Zoya Bylinskii et al. "What Do Different Evaluation Metrics Tell Us About Saliency Models?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.3 (2019), pp. 740–757.

[8] Samira Abnar and Willem Zuidema. “Quantifying Attention Flow in Transformers”. In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2020.

[9] Sarthak Jain and Byron C. Wallace. “Attention is not Explanation”. In: *NAACL-HLT*. 2019.

[10] Sofia Serrano and Noah A. Smith. “Is Attention Interpretable?” In: *Annual Meeting of the Association for Computational Linguistics (ACL)*. 2019.

[11] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2017.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *International Conference on Machine Learning (ICML)*. 2017.

[13] Hila Chefer, Shir Gur, and Lior Wolf. “Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2021.

[14] Julius Adebayo et al. “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[15] Yi Li et al. “Token Activation Map to Visually Explain Multimodal LLMs”. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2025.

[16] Guanxi Shen. *GLIMPSE: Holistic Cross-Modal Explainability for Large Vision-Language Models*. 2025. arXiv: 2506.18985 [cs.CV].

[17] Matthew D. Zeiler and Rob Fergus. “Visualizing and Understanding Convolutional Networks”. In: *European Conference on Computer Vision (ECCV)*. 2014.

[18] Sara Hooker et al. “A Benchmark for Interpretability Methods in Deep Neural Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[19] Mohamed Eltahir et al. *GridProbe: Posterior-Probing for Adaptive Test-Time Compute in Long-Video VLMs*. 2026. arXiv: 2605.10762 [cs.CV].

[20] Size Wu et al. “F-LMM: Grounding Frozen Large Multimodal Models”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025.

[21] Qwen Team et al. *Qwen3-VL Technical Report*. 2025. arXiv: 2511.21631 [cs.CV].

[22] Weiyun Wang et al. *InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency*. 2025. arXiv: 2508.18265 [cs.CV].

[23] Amanpreet Singh et al. “Towards VQA Models That Can Read”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[24] Drew A. Hudson and Christopher D. Manning. “GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[25] Yifan Li et al. “Evaluating Object Hallucination in Large Vision-Language Models”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2023.

[26] LASA Team et al. *Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning*. 2025. arXiv: 2506.07044 [cs.CL].

[27] Rishika Bhagwatkar et al. “CAVE: Detecting and Explaining Commonsense Anomalies in Visual Environments”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2025.

A Prompt Templates

All prompts are listed verbatim. Placeholders in braces are filled at run time, `{cond}` and `{q}` with the user query exactly as it appears in the benchmark, `{hint}` with the sender's location phrase.

Band probe (every probe query).. Each of the $2K$ band questions shows the model the band content and asks:

```
You are shown one or more adjacent tiles cropped from a larger
image. Is the following present in ANY of these tiles?
"{cond}"
Answer with exactly one word: Yes or No.
```

The answer is never read from generated text. We take the first answer-token logits and compute the yes posterior over the closed label set, aggregating surface forms by logsumexp, yes forms `{Yes, yes, YES}` with and without a leading space, and likewise for no.

Pointing reference (agreement test and point-in-region).. The generated point uses each family's native grounding convention, declared per family rather than inferred, coordinates are parsed from the reply and interpreted on a 0 to 1000 scale for both families below. Qwen3-VL models, and medical derivatives built on Qwen-VL such as Lingshu:

```
Locate "{q}" in this image and output its center point as
JSON: {"point_2d": [x, y], "label": "{q}"}
```

InternVL models, the documented grounding template, the box centre is taken as the point:

```
Please provide the bounding box coordinate of the region this
sentence describes: <ref>{q}</ref>
```

Other models, a generic pixel-space prompt:

```
Point to {q} in this image. Output ONLY the pixel coordinates
of its center as (x, y). No other text.
```

Open-ended answering (deletion test, image-reliance labeling, self-conditioning)..

```
{q}
Answer with a single word or phrase.
```

Existence claims (hallucination audit)..

```
{q}
Answer yes or no.
```

Self-conditioning close-up.. The receiver sees the full image and the crop of the map's top-mass region as a second image:

```
{q}
The second image is a close-up of the region most relevant to
the question.
Answer with a single word or phrase.
```

Location hint (the words channel of the transfer comparison).. The sender is asked:

```
Where in the image is the information needed to answer the
question below? Reply with ONE short location phrase, for
example 'bottom left, on the license plate'. Do not answer
the question itself.
Question: {q}
```

and the receiver is asked:

```
{q}
Hint: look at {hint}.
Answer with a single word or phrase.
```

B Qualitative Examples



(a) KAUST Beacon **(b)** a truck number 14 on a snow bank **(c)** Carcinoma, malignant,invasive

Figure 7 Qualitative examples of the probe map. Each column shows one example, with the input image above its corresponding probe map.

C Additional Diagnostics

Localization heads, per-example variant.. The published method selects heads on a calibration corpus. A per-example adaptation with no calibration performs at noise level on both tests (agreement distance 0.375, coordinate correlation −0.051 on RefCOCOg, deletion at the random control), consistent with the authors' own finding that uncalibrated attention maps are sparse and noisy [3].

Table 10 Full agreement detail on RefCOCOg, all four metrics.

Method	NSS	AUC	Dist. ↓	Pearson r
Probe map (K=8)	1.36	0.824	0.130	0.665
Probe map ({3, 5})	**1.56**	**0.853**	**0.122**	**0.711**
Attention (best layer, ℓ=15)	−0.22	0.380	0.193	0.224
Attention (last layer)	−0.15	0.464	0.193	−0.036
Attention rollout	−0.16	0.457	0.292	0.120
Relevancy T-MM (all layers)	0.33	0.656	0.159	0.706
Relevancy T-MM (last block)	−0.15	0.420	0.213	0.048
Occlusion	1.04	0.649	0.160	0.504
Random map	−0.05	0.486	0.188	0.048
Image-centre prior	–	–	0.189	–

Table 11 Map-maximum grading, probe-region flip rate (%) by tercile.

Configuration	Low tercile	High tercile
TextVQA, 4B, blank	44.3	64.2
TextVQA, 4B, blur	49.2	61.8
GQA, 4B	16.1	28.4
TextVQA, 30B	26.9	58.5
TextVQA, InternVL3.5-8B	48.3	60.7

Table 12 Image-reliance split, probe-region flip rate (%) on image-used against image-unused rows. TextVQA unless noted.

Configuration	Image used	Image unused	Random (unused)
Qwen3-VL-4B	57.4	22.0	2.4
Qwen3-VL-8B	53.8	31.1	2.2
Qwen3-VL-30B-A3B	48.4	22.2	0.0
InternVL3.5-8B	54.5	20.6	5.9
GQA non-binary, 4B	30.0	18.7	0.0

D Configuration Details

Table 13 Per-experiment configuration. All values are fixed across models and benchmarks unless a sweep is the experiment.

Probe image long side	512 px
Answering and deletion long side	1024 px
Self-conditioning long side	1280 px (the crop inherits it)
Grid	K=8 default, multigrid product $\{3, 5\}$ or $\{2, 3, 5\}$
Band presentation	one concatenated strip image per question
Deletion region	top-mass cells, 12% of the grid (8 of 64)
Deletion fill	grey RGB (127,127,127), blur variant Gaussian
Random control	same cell count, disjoint from every method region
Crop margin (self-conditioning)	18% per side, minimum 96 px
Attention layer selection	swept on 200 held-out examples per model
Decoding	greedy, deterministic, fixed per-example seeds