PRE-GENERATION HALLUCINATION PROBES WORK FOR YES/NO ANSWERS BUT NOT FOR FREE CAPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Probes that read a vision-language model's hidden states before any token is generated have been reported to predict hallucination with AUROC up to 0.93. These results come mostly from yes/no questions, while object hallucination is usually measured in free-form captions. We test whether the same probes work in both settings by holding the model, image, queried object and sampling fixed and changing only the instruction. The endpoint is whether the model asserts one object that two annotators verified to be absent. For each query we estimate the model's propensity to hallucinate from 16 sampled decodes, score the probe on an independent decode, and report recovery, the probe's gain in AUROC over chance divided by the gain of an oracle that knows each query's propensity. Across three vision-language models and three decoding seeds, probes recover 84 to 97% under a forced yes/no format but only −14 to 6% in free captions, with upper 95% bounds of 0 to 22%. A permutation test, a sweep over layers and token positions, and nonlinear probes all support this null. Naming the object in an otherwise open prompt restores part of the signal in one of the three models. Under forced choice the signal appears soon after the object's name, before the prompt ends, and rises sharply in the middle layers, which suggests that the probe reads an answer the model has already settled on. A human audit also shows that a CHAIR-style labeler has a precision of 0.24 on the objects it flags.

1 INTRODUCTION

Vision-language models (VLMs) (Liu et al., 2023; 2024) often mention objects that are not in the image (Rohrbach et al., 2018; Li et al., 2023b). Most methods for detecting such errors look at the output after it has been written, either by comparing several sampled answers (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024) or by checking the text against the image. Both cost extra computation and come too late to prevent the error. A cheaper option is to read the model's hidden state once it has processed the prompt, and to abstain, reroute the query or decode more carefully when a hallucination looks likely. In language models, hidden states reveal whether a statement is true or an answer is hallucinated (Azaria & Mitchell, 2023; Marks & Tegmark, 2024; CH-Wang et al., 2024; Orgad et al., 2025), and signals read before generation predict whether the model knows the answer or is likely to hallucinate (Kadavath et al., 2022; Gottesman & Geva, 2024; Ji et al., 2024; Kossen et al., 2024). For VLMs, HALP (Kogilathota et al., 2026) probes eight models before generation and reports AUROC up to 0.93, including 0.915 for Qwen2.5-VL-7B and 0.896 for Llama-3.2-11B-Vision.

These numbers come mostly from short answers. About two thirds (65.7%) of the questions in the HALP evaluation have yes/no answers, results are reported by domain and hallucination type rather than by answer format, and a GPT-4 judge decides whether an answer is correct. Object hallucination, in contrast, is usually measured in free descriptions (Rohrbach et al., 2018; Wang et al., 2023; Kaul et al., 2024), and the two settings do not always agree. Kaul et al. (2024) report that hallucination in yes/no answers and hallucination in free-form generation are often anti-correlated. Because benchmarks differ at once in images, objects, labels, base rates and output format, comparing results across them cannot show whether a pre-generation signal also exists for free captions.

We therefore compare output formats within the same queries (Figure 1). The model, the image, the queried object and the sampling settings stay fixed, and only the instruction changes, from a forced

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.

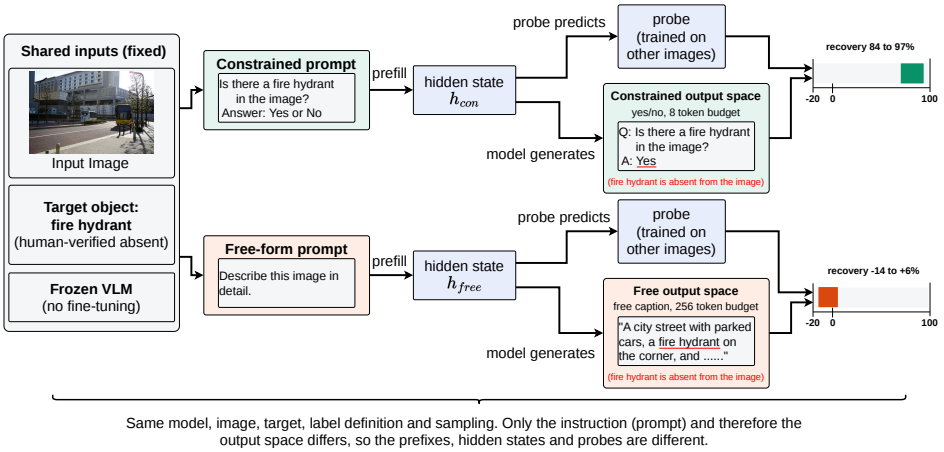


Figure 1: Within-subject design. The same frozen model sees the same image and the same verified-absent object under two instructions, a forced yes/no question and a request for a free caption. A probe trained on other images reads the hidden states at the last prefill token and predicts whether the model will assert the object. Recovery is 84 to 97% for the forced format and −14 to 6% for the free caption (Table 1).

yes/no answer through a fixed template and a single sentence to a free caption. The free caption also differs in a second way, because its prompt does not name the object, so a probe that reads the prompt cannot know which object will be scored. Two open prompts that do name the object separate these two factors. The endpoint is narrow on purpose: whether the model asserts one specific object that two annotators verified to be absent from the image. We estimate each query's propensity to produce this assertion from 16 sampled decodes, score the probe against a further independent decode, and report *recovery*, the probe's gain in AUROC over chance divided by the gain of an oracle that knows the propensity.

Under the forced format the answer is the next token, so a probe on the last prefill state should succeed. It recovers 84 to 97% of the attainable signal in all three models. In free captions the same pipeline recovers only −14 to 6%, no interval lies above zero, and the upper 95% bounds of 0 to 22% limit how much signal it could have missed. The paired difference between the two formats ranges from 92 to 110 points across models. A permutation test separates every forced-choice run from chance but no free-caption run. Refitting the probe at single layers and at earlier token positions leaves only 1 of 66 free-caption cells above chance, and nonlinear probes reach at most 9%. Naming the object in an open prompt brings back part of the signal, but in only one of the three models.

Under forced choice we can also see where the signal forms. It is absent while the model is still reading the object's name and present a few tokens later, before the prompt ends, and across depth it rises sharply in the middle layers and then stays high. For two of the models the first principal component of the hidden states, found without labels, already carries it, with a recovery of 79% for Qwen2.5-VL and −99% for Llama-3.2-V, where the minus sign marks a near-perfect ranking in reverse. We take this as evidence that a forced format makes the model settle its answer while it reads the prompt, and that the probe reads this answer. Because every queried object is absent, a settled yes is also a hallucination, so this remains an interpretation.

Free-form hallucination labels usually come from CHAIR (Rohrbach et al., 2018) or from similar matching of caption words against annotations. An earlier version of this study used such a labeler for every object in a caption. When two annotators audited it, its precision on the objects it flagged was only 0.24 [0.16, 0.33]. This is why our endpoint is a single verified-absent object rather than every object in a caption.

Contributions. (i) A within-subject protocol that separates output format from target information, with human-verified absent targets, an independent outcome decode and an oracle-normalized measure (§3). (ii) Evidence from three VLMs and three seeds that pre-generation probes recover most of the attainable signal under forced choice and none we can detect in free captions (§4). (iii) Controls for

chance, layer, token position and probe type, and an analysis of where the forced-choice signal forms (§5). (iv) A human audit showing that a CHAIR-style labeler has low precision on the objects it flags (§5.4).

2 Related Work

Probing language models for truth and hallucination. Probes trained on the activations of language models can tell whether a statement is true (Azaria & Mitchell, 2023; Marks & Tegmark, 2024). A direction that separates true from false statements can also be found without labels (Burns et al., 2023), and shifting activations along such a direction at inference time makes answers more truthful (Li et al., 2023a). Probes on the generated tokens detect hallucinated answers (CH-Wang et al., 2024), and Orgad et al. (2025) find that the signal is strongest in the tokens of the exact answer and that detectors trained on one task transfer poorly to others. Other work reads the model before it answers. Kadavath et al. (2022) train models to predict whether they know the answer to a question. Probes on the query alone predict hallucination risk (Ji et al., 2024) and the semantic entropy of answers that have not been sampled yet (Kossen et al., 2024), and a probe on the representation of an entity estimates how much the model knows about it, which correlates with the factuality of what the model later writes about that entity (Gottesman & Geva, 2024). None of these studies keeps the query fixed and changes only the output format, which is the comparison we make.

Pre-generation signals in VLMs. In VLMs, the logits of the first generated token reveal whether a question can be answered (Zhao et al., 2024), and projecting image representations onto the vocabulary separates objects that are present from hallucinated ones (Jiang et al., 2025). The closest work to ours is HALP (Kogilathota et al., 2026), which probes visual features, vision-token states and query-token states of eight VLMs before generation and reports AUROC up to 0.93 on a mix of six benchmarks. About two thirds of its questions have yes/no answers, its labels come from a GPT-4 judge, and it reports results by domain and hallucination type but not by answer format. It therefore leaves open whether such probes work when the model writes a free description.

Uncertainty from sampling. Another line of work estimates uncertainty from several sampled answers. SelfCheckGPT measures their consistency (Manakul et al., 2023), semantic entropy groups them by meaning before measuring their spread (Kuhn et al., 2023; Farquhar et al., 2024), and INSIDE uses the spread of their internal embeddings (Chen et al., 2024). We use repeated sampling for a different purpose. The fraction of 16 decodes that assert the absent object estimates the model's propensity to hallucinate it. This propensity sets a ceiling for any predictor of a single decode, and the probe is scored on a separate decode.

Measuring object hallucination. CHAIR finds the objects mentioned in a caption with a synonym list and flags those absent from both the COCO segmentation annotations and the reference captions (Rohrbach et al., 2018; Lin et al., 2014). POPE instead asks the model yes/no questions about sampled objects (Li et al., 2023b), AMBER combines generative and discriminative tasks (Wang et al., 2023), THRONE judges free-form responses with language models (Kaul et al., 2024), and ALOHa matches objects extracted by a language model to references and validates the score against human annotations (Petryk et al., 2024). These benchmarks score a model over many responses and are used to compare models and decoding methods such as visual contrastive decoding (Leng et al., 2024) and OPERA (Huang et al., 2024). Our audit (Section 5.4) asks whether CHAIR-style labels are correct for individual objects, which matters when they are used to train or evaluate a predictor for each query. Reiter (2018) drew the same line for BLEU, finding that the evidence supports its use for comparing machine translation systems but not for scoring individual texts.

Probing methodology and output format. A probe's accuracy reflects both what a representation encodes and what the probe can learn (Hewitt & Liang, 2019; Belinkov, 2022). Cross-validated estimates are noisy when test sets are small (Varoquaux, 2018), and AUROC computed on predictions pooled across folds is biased when each fold's scores are calibrated differently (Forman & Scholz, 2010). Permutation tests compare a classifier directly with chance (Ojala & Garriga, 2010). The output format changes model behavior as well. Small formatting choices shift accuracy (Sclar et al., 2024), restricting the format affects reasoning (Tam et al., 2024), and a model that commits to a wrong answer early often backs it with further false claims (Zhang et al., 2024). These studies measure

changes in the output. We ask how the format changes what a probe can read from the model before it generates anything.

3 METHOD

3.1 QUERIES AND LABELS

We use 300 images from COCO val2017 (Lin et al., 2014). For each image we draw 4 target objects from the 80 COCO classes that its instance annotations mark as absent, one from each quartile of class frequency, so that both common and rare classes are queried. An object can be missing from the annotations and still be visible, so two annotators checked every image-object pair against the image. They confirmed 1,182 of the 1,200 pairs as absent (Cohen's $\kappa = 0.94$), and we drop the other 18 (Appendix A).

A query $q = (I, o, f)$ combines an image I, a target object o and an output format f. A decode z is a hallucination event, $\ell(z, o) = 1$, if it asserts that o is present. Since every target is verified absent, any such assertion is a hallucination, and the label asks the same question in every format. Only the way the label is read differs. Yes/no answers are parsed directly. All other outputs go through a string matcher that looks for a surface form of o with no negation in the 60 characters before it (Appendix B). The matcher misses a negation that follows the name, as in "a fire hydrant is not visible", and can therefore count extra events in the formats that ask the model to comment on the object (Section 4.3). The main comparison is less exposed to this error, because forced-choice answers are parsed and the free-caption prompt gives the model no reason to discuss an absent object.

3.2 OUTPUT FORMATS

The conditions vary two factors (Table A1 and Figure A1 in Appendix C). The first is how open the output is while the prompt names the object. In *forced choice* the model answers yes or no within 8 new tokens. The *template* condition asks it to complete a fixed sentence with one word (12 tokens), and *one sentence* asks for a single sentence on whether the object appears (48). In *free caption, object named* the model writes a detailed description and is also asked to indicate whether the object is present (256). *Named late* rewords this prompt so that it ends with the object's name. This moves the name closer to the probed position, although it also changes the rest of the wording.

The second factor is whether the prompt names the object at all. The *free caption* uses the generic instruction "Describe this image in detail." and is scored against all 4 targets of the image. Its prompt, and with it the prefill state, is the same for all of them, so a probe on this state cannot know which object will be scored. Every other condition has a separate prompt for each image-object pair. Forced choice and the named free caption therefore differ mainly in how open the output is, while the named and the generic free caption differ only in the added sentence that names the object. No constrained condition leaves the object out, because a constrained format has to name what it asks about.

3.3 MODELS AND PROBE

We study three open VLMs that pass visual information to the language model in two different ways (Appendix D). Qwen2.5-VL-7B (Bai et al., 2025) and InternVL3-8B (Zhu et al., 2025) place visual tokens in the decoder's input sequence. The first encodes images at their native resolution, and the second splits them into tiles and maps them through an MLP projector. Both build on the Qwen2.5-7B language model. Llama-3.2-11B-Vision (Llama Team, AI @ Meta, 2024) instead passes visual features to the decoder through cross-attention layers. All three models stay frozen, and we refer to them as Qwen2.5-VL, Llama-3.2-V and InternVL3.

The probe reads the last prefill token, the position from which the first output token is generated. Its input is the residual stream at this token after layers $\mathrm{round}(L/4)$, $\mathrm{round}(L/2)$ and L of an L-layer decoder, concatenated into one vector. The probe is trained to predict the outcome y of Section 3.4 with 5-fold cross-fitting grouped by image. Inside each fold, the features are standardized and reduced to 10 principal components fitted on the training images only, and a logistic regression with $C = 0.1$ is fitted on these components. Every query thus receives an out-of-fold prediction $p(q)$ from a probe

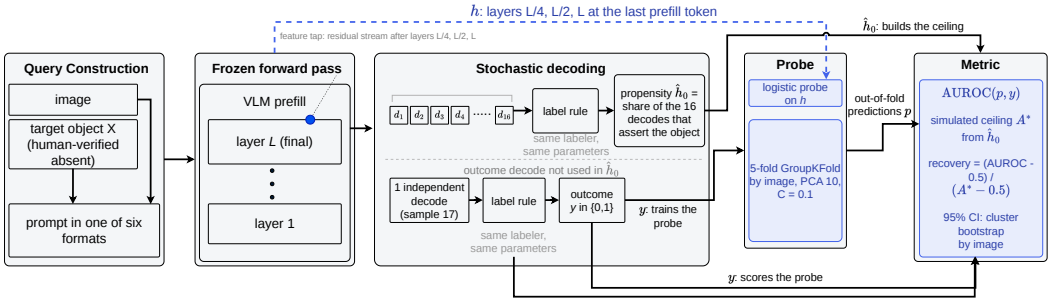


Figure 2: Measurement pipeline for one model and condition. A frozen prefill pass gives the probe's features, the residual stream at the last prefill token after layers $L/4$, $L/2$ and L. The propensity $\hat{h}_0$ comes from 16 sampled decodes, and a further independent decode gives the outcome y that trains and scores the probe. The same label rule reads all decodes. Cross-fitting grouped by image gives an out-of-fold prediction for every query, and recovery compares the probe's AUROC with the ceiling $A^\star$ simulated from $\hat{h}_0$.

that never saw its image (Figure 2 and Appendix E). These settings are fixed and the same for every model, condition and seed.

3.4 Propensity and Recovery

For every query we draw $N = 16$ decodes at temperature 1.0 and top-p 1.0, leaving all other generation settings at each model's defaults (Appendix F), and estimate the propensity

$$\hat{h}_0(q) = \frac{1}{N} \sum_{j=1}^{N} \ell(z_j, o). \tag{1}$$

A further decode z_{N+1}, sampled independently of the first N, gives the outcome $y(q) = \ell(z_{N+1}, o)$ that the probe is trained and scored on. Since y is not used in $\hat{h}_0$, the ceiling defined next does not depend on the outcomes that score the probe.

Raw AUROC is hard to compare across formats. The base rate and the spread of propensities change with the format, and so does the best AUROC that any predictor can reach. We therefore normalize by an oracle that scores each query by its propensity. Its expected AUROC $A^\star$ is estimated by simulating outcomes $y^\star(q) \sim \text{Bernoulli}(\hat{h}_0(q))$ (Appendix G), and recovery is

$$R = \frac{\text{AUROC}(p, y) - 0.5}{A^\star - 0.5}. \tag{2}$$

$R = 0$ is chance and $R = 1$ means that the probe matches the oracle. R can exceed 1, because the probe is scored on one realized outcome while $A^\star$ is an expectation. The simulation treats $\hat{h}_0$ as the true propensity, so when events are rare $A^\star$ tends to overstate the oracle and R errs on the conservative side. For this reason we also report raw AUROC (Table A3).

Each condition is run with three decoding seeds, and each seed redraws all 17 decodes of every query. We compute recovery for each seed and report the mean over seeds. The seeds share their images, so we do not treat them as independent samples. Intervals come from a cluster bootstrap over images with 2,000 draws. Each draw resamples the images once and applies the same resample to every seed and every condition of a model, which makes the differences between conditions paired. The ceilings are held at their point estimates inside the bootstrap.

4 Results

4.1 Forced Choice and Free Caption

Table 1 gives the main comparison. Under forced choice the probes recover 84% of the attainable signal for Qwen2.5-VL, 97% for Llama-3.2-V and 96% for InternVL3, and every interval lies well

Table 1: Recovery in percent of the attainable signal, mean over decoding seeds, with 95% intervals from a cluster bootstrap over images (2,000 draws). Each draw resamples images once and applies the draw to every seed and every condition of a model, so the seeds are not treated as independent and the differences in the lower block are paired. Differences are computed before rounding. Budget is the maximum number of new tokens. Per-cell AUROC, ceilings and event counts are in Table A3. *Per-query predictions unavailable, so the interval is the mean of the per-seed intervals.

Condition	Budget	Qwen2.5-VL-7B	Llama-3.2-11B-V	InternVL3-8B
forced choice	8	+84 [65, 95]	+97 [94, 99]	+96 [92, 99]
template	12	+92 [83, 99]	+83 [77, 89]	+75 [63, 85]
one sentence	48	+31 [18, 44]	+71 [61, 81]	+78 [67, 87]
free caption	256	−12 [−24, 0]	+6 [−11, 22]	−14 [−34, 8]
free caption, object named	256	−1 [−13, 10]	+41 [31, 52]	+9* [−7, 24]
free caption, named late	256	−7 [−17, 3]	+27 [12, 41]	+1* [−17, 20]
forced choice minus free caption		+96 [74, 113]	+92 [75, 108]	+110 [88, 131]
forced choice minus object named		+86 [64, 102]	+56 [45, 67]	+87 not paired
object named minus free caption		+11 [−5, 27]	+36 [18, 53]	+23 not paired

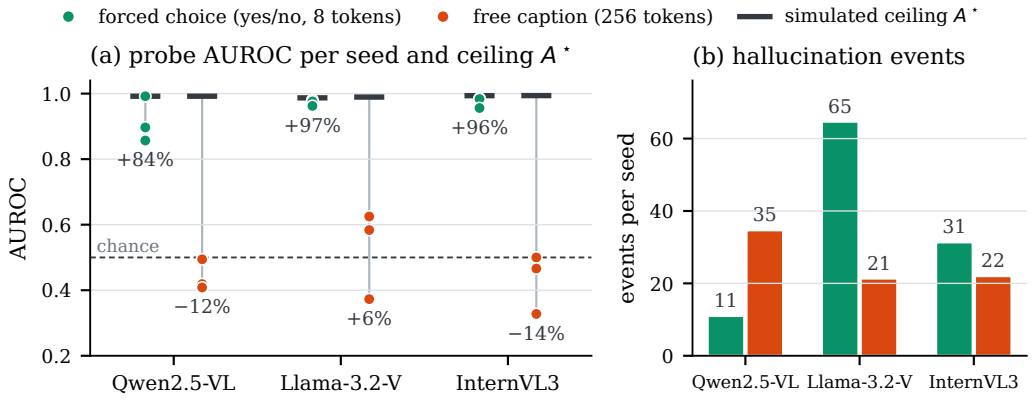


Figure 3: Forced choice and free caption for the same images and objects. (a) Probe AUROC for each decoding seed (dots) and the simulated ceiling $A^\star$ (bars), with the mean recovery printed below the dots. The ceilings are almost the same in the two formats, and only the probe changes. (b) Hallucination events per seed. Qwen2.5-VL has more events in the free caption than under forced choice, but only its forced answers are predictable.

above zero. With the same images, objects and label definition, the free caption leaves them at −12%, 6% and −14%. No interval lies above zero, and the upper bounds of 0 to 22% limit how much signal the probes could have missed. The negative point estimates fall within the range that shuffled labels produce (Section 5.1). The paired difference between the two formats is 92 to 110 points, and its interval starts at 74 points or higher in every model (lower block of Table 1).

The drop does not come from a change in the ceiling (Figure 3a). In both formats most queries are never hallucinated in 16 decodes and the events concentrate on a few queries. The outcome is then easy to predict for anyone who knows the propensity, and the ceilings lie between 0.987 and 0.994 in both formats. The probe's AUROC, in contrast, is 0.915 to 0.975 under forced choice and 0.431 to 0.527 for the free caption, so the gap is just as clear in raw AUROC. The number of events does not explain it either (Figure 3b). Qwen2.5-VL produces more events per seed in the free caption (35) than under forced choice (11), and only its forced answers are predictable. The free-caption probes also stay close to the base rate, with no prediction above 0.054, while forced-choice predictions reach 0.238.

4.2 Intermediate Formats

The template and one-sentence formats show where along the way the signal is lost (Table 1 and Figure A3 in Appendix H). If predictability simply shrank with the length of the permitted answer, recovery would fall at every step from 8 to 12 to 48 to 256 tokens. Within the constrained formats the point estimates do not follow the budget. Qwen2.5-VL recovers 84% under forced choice and 92% with the template, and InternVL3 recovers 75% with the template and 78% with one sentence, although these differences are within the intervals.

Where the signal is lost depends on the model. Qwen2.5-VL loses most of it at the single sentence (31%). InternVL3 keeps it through the sentence (78%) and loses it only in the open description, where it recovers 9% even with the object named. Llama-3.2-V declines in steps and still recovers 41% in the named open description. Without the name, all three end near zero.

4.3 Naming the Object

The free caption differs from forced choice in two ways. Its output is open, and its prompt does not name the object, so the probe sees one prefill state per image and has to score all 4 targets of the image with it. We first ask how much a score that is constant within an image could reach. Scoring each query by the mean propensity of its image gives AUROC 0.90 to 0.91 in the free caption, so an image-level score could capture most of the signal. Against this lower ceiling the probe still recovers only -16 to 7%. The missing name alone cannot explain the failure.

The named conditions give a second test. They keep the open output but name the object, so each query has its own prefill state. With the object named, recovery is -1%, 41% and 9%, far below the forced-choice values, and its interval lies above zero only for Llama-3.2-V. Ending the prompt on the name does not help (-7%, 27% and 1%). The lower block of Table 1 splits the forced-to-free gap along this path. The output-space step (forced choice minus object named) is 86, 56 and 87 points, and the naming step (object named minus free caption) is 11, 36 and 23 points, so the output-space step is the larger one in every model. This split depends on the path, because no constrained condition leaves the object unnamed, and the output-space step also changes the wording, the label rule and the base rate.

Naming the object also makes the model talk about it. The event rate rises to 10.3 to 28.4% with the object named, from 1.8 to 3.0% in the free caption. Recovery corrects for the change in the ceiling that this brings, but not for label errors. The matcher misses negations placed after the name (Section 3.1), so some counted events in the one-sentence and named formats may be denials such as "a fire hydrant is not visible". A probe that reads the model's intended answer would then be scored against the wrong label. The naming step is therefore the less certain part of the split, whereas the image-level comparison above does not use the named formats at all.

5 Ablations and Controls

5.1 Permutation Test

With as few as 7 events in a single seed, a probe can reach a sizable AUROC by luck, and a single run with shuffled labels is a weak control. For every model and seed of the forced-choice and free-caption formats we therefore fit a separate linear probe (PCA to 64 components, then logistic regression, folds grouped by image) on the true labels and on 200 permutations of them, and compute $p = (1 + \#\{\text{null} \geq \text{observed}\})/(200 + 1)$ (Ojala & Garriga, 2010). Under forced choice the true labels beat every permutation in all 9 runs, so $p = 0.005$, the smallest value the test can give (Table 2). In the free caption no run differs from its null, with p between 0.13 and 0.87.

The null distributions also show how large chance recoveries become with so few events. Their 95% bands extend 20 to 55 points on either side of means between -5 and 0%. The slightly negative means reflect the known downward bias of AUROC pooled over cross-validation folds (Forman & Scholz, 2010), which also helps explain why some free-caption recoveries in Table 1 fall below zero. The single-seed sweep below is correspondingly noisy.

Table 2: Controls for forced choice (forced) and the free caption (free). Values are recovery in percent, averaged over three seeds, with the standard deviation across seeds in small type. The permutation rows use a separate linear probe with 200 permutations per seed (Section 5.1), and 0.005 is the smallest attainable p. PC-1 is the first principal component of the hidden states, computed without labels. Its sign is arbitrary, so -99 for Llama-3.2-V is a near-perfect ranking with the sign inverted. The MLP (hidden layers of 128 and 64 units) and the boosted trees are trained on the same hidden states, with folds grouped by image.

	Qwen2.5-VL-7B		Llama-3.2-11B-V		InternVL3-8B	
Control	forced	free	forced	free	forced	free
trained probe	$+84$ ± 14	-12 ± 10	$+97$ ± 2	$+6$ ± 28	$+96$ ± 3	-14 ± 19
permutation null, mean	-3	-2	-2	-1	-2	-3
permutation p, all seeds	≤ 0.005	≥ 0.13	≤ 0.005	≥ 0.13	≤ 0.005	≥ 0.13
unsupervised PC-1	$+79$ ± 11	$+9$ ± 6	-99 ± 1	$+9$ ± 2	$+37$ ± 3	-7 ± 2
MLP probe	-16	$+6$	$+76$	$+9$	$+39$	$+9$
boosted trees	$+80$	-10	$+97$	$+2$	$+97$	-4

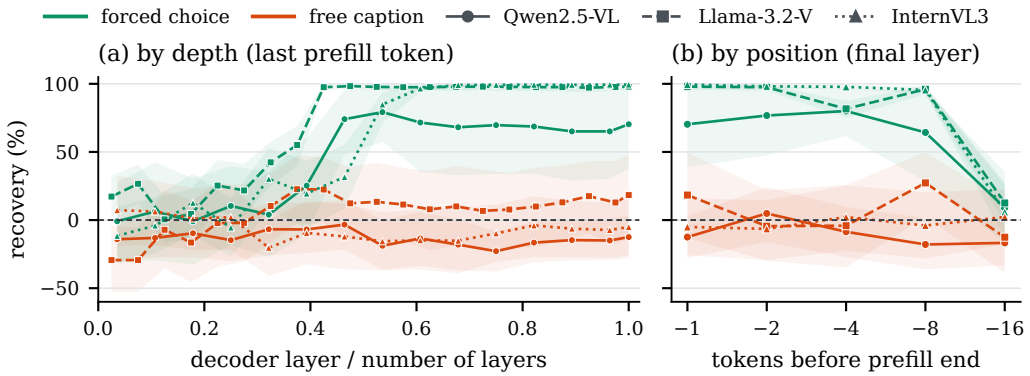


Figure 4: Recovery of a linear probe refit on single layers and earlier token positions (first seed, bands are 95% intervals). (a) Single layers at the last prefill token, every second layer, with depth divided by the number of layers. (b) The final layer at five positions before the end of the prefill, drawn at equal spacing. Under forced choice the signal appears in the middle layers and between positions -16 and -8. In the free caption no layer or position carries it. Per-model panels are in Figure A4.

5.2 Layers, Positions and Probe Types

The signal could sit at a layer or token position that the main probe does not read, or it could need a more flexible reader. For the first seed we refit a linear probe on single layers at the last prefill token, taking every second layer, and on the final layer at five positions before the end of the prefill (Figure 4 and Appendix I). In the free caption, 1 of 66 layer and position cells lies above chance (Llama-3.2-V, position -8, 27% [+2, +50]), about as many as chance alone produces across this many cells, and 5 lie below it. Under forced choice 50 of 66 cells lie above chance, and every exception is a layer at or below 11 or position -16.

More flexible probes on the same hidden states do not change the picture (Table 2). Boosted trees recover 80 to 97% under forced choice and at most 2% in the free caption. An MLP reaches at most 9% in the free caption, but it is unstable with so few events and fails even under forced choice for Qwen2.5-VL (-16%).

5.3 Forced-Choice Signal

The sweep also shows where the forced-choice signal forms. It is absent at position -16 (6 to 13%, every interval including zero) and present at position -8 (64 to 96%). In the forced-choice prompt, position -16 falls just before the end of the object's name and position -8 falls inside the instruction "Answer yes or no.", so the answer becomes readable a few tokens after the name, before the prompt ends. Positions in between were not probed. Across depth the signal rises sharply between two neighboring sampled layers in the middle of the network and then stays high, from 55 to 98% between layers 15 and 17 of 40 in Llama-3.2-V, from 32 to 85% between layers 13 and 15 of 28 in InternVL3 and from 25 to 74% between layers 11 and 13 of 28 in Qwen2.5-VL.

For two of the models the answer is also the dominant feature of the prefill state. Without labels, the first principal component recovers 79% for Qwen2.5-VL and orders Llama-3.2-V's queries almost perfectly in reverse (AUROC 0.014 to 0.023), while it recovers 37% for InternVL3 and -7 to 9% in the free caption (Table 2). These results suggest that the model settles its answer while it reads a forced-format prompt. Because every target is absent, a settled yes and a hallucination coincide here, so the results cannot separate a committed answer from a disposition to err on the object. In a free caption, whether an absent object is mentioned may instead depend on what the model has already written, since early commitments and attention patterns that build up during generation are associated with later errors (Zhang et al., 2024; Huang et al., 2024).

5.4 Label Audit

An earlier version of this study labeled every object mentioned in a caption with a CHAIR-style labeler (Section 2), and we audited it before relying on it. Two annotators independently judged 222 caption-object items against the image, and their 52 disagreements were adjudicated. The sample oversampled flagged items, so the numbers describe this sample and not rates over all captions.

Against the adjudicated labels, the labeler's precision on flagged items was 0.24 [0.16, 0.33] and its recall 0.56 [0.40, 0.70] (Table A4 in Appendix J). Three of every four flagged items were not hallucinations. The largest class among the false positives was "dining table", with 26 of 79, the same example that Rohrbach et al. (2018) give for why reference captions are needed. The annotators agreed on 77% of items before adjudication ($\kappa = 0.40$). Such errors may average out when models are ranked, which we did not test, but they matter when labels train or evaluate a predictor for each query, since a probe trained on these positives would learn mostly from mislabeled examples. This is why the study uses one human-verified absent object per query.

6 Conclusion

With the model, image, object and sampling held fixed, pre-generation probes recover most of the attainable signal when the model must answer yes or no, and none we can detect when it writes a free caption. The gap holds in all three VLMs, survives a permutation test, sweeps over layers and positions, and nonlinear probes, and is not explained by the ceiling, the event count or the missing object name. Under forced choice the probe seems to read an answer that the model settles soon after reading the object's name. Such probes can help systems that ask constrained questions, but on our evidence they cannot replace checking free descriptions. Their evaluation should use the output format of the intended application and report a ceiling next to AUROC.

Our endpoint is the assertion of one verified-absent object per query, so caption-level risk, attributes, relations and scene-typical objects remain untested. We study three open models with 7B to 11B parameters, two of which share a language model, on COCO images with English prompts. Events are rare, so the free-caption result bounds recovery from above rather than proving it is zero. The label rule misses negations after the object's name, mainly in the one-sentence and named formats. We did not compare the forced-choice probe with the model's probability of answering yes, which may do as well. Other readouts, such as vocabulary projections of image tokens (Jiang et al., 2025), might find signal we miss.

AI USE STATEMENT

We used generative AI tools, specifically a large language model, only to improve sentence structure, grammar, clarity, and overall language presentation. These tools were not used to generate or label data, and all image-object pairs and the audit sample were evaluated by human annotators. Mathematical claims, proofs, and translation are not applicable to this work. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

The images come from the public COCO dataset (Lin et al., 2014), and the three models are publicly released and were used under their licenses. Members of our lab verified the image-object pairs and annotated the audit sample. The task was to judge whether an object appears in a public image, and no personal data were collected. The main practical risk we see is false confidence, since a hallucination detector that works for one output format can fail without warning on another, and our results show one such case.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code for all experiments, statistics and figures, together with the macro file through which every number in the text is taken from the result files. Section 3 describes the queries, labels, probe and statistics, and the appendix gives the prompts, checkpoints, sampling settings, probe details and compute.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976. Association for Computational Linguistics, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

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.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. Do androids know they're only dreaming of electric sheep? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4401–4420. Association for Computational Linguistics, 2024.

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

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

George Forman and Martin Scholz. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1):49–57, 2010. doi: 10.1145/1882471.1882479.

Daniela Gottesman and Mor Geva. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.232.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

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

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104, Miami, Florida, US, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.6.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. THRONE: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27228–27238, 2024.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.287.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

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

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11:1833–1863, 2010.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations*, 2025.

Suzanne Petryk, David Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. ALOHa: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-short.30.

Ehud Reiter. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, 2018. doi: 10.1162/coli_a_00322.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236. Association for Computational Linguistics, 2024.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018. doi: 10.1016/j.neuroimage.2017.06.061.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59670–59684, 2024.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pp. 127–142. Springer, 2024. doi: 10.1007/978-3-031-73195-2_8.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Target Selection and Verification

We use 300 images from COCO val2017 (Lin et al., 2014). For each image we draw 4 target objects from the COCO classes that its instance annotations mark as absent, one from each quartile of class frequency, which gives 1,200 image-object pairs.

An object can be missing from the annotations and still be visible, for example when it is small, partly hidden or simply not annotated. Two annotators from our lab therefore judged every pair against the image, independently of each other, and agreed well (Cohen's $\kappa = 0.94$). We keep the 1,182 pairs confirmed as absent and drop the other 18. Every model and condition uses the same verified pairs.

B Label Rules

In the forced-choice and template formats the answer is parsed as yes or no, and a yes is a hallucination event because the object is absent. All other formats go through a string matcher. A decode is an event when it contains a surface form of the target's class name that is not negated, and a mention counts as negated when a negation cue occurs within the 60 characters before it. The words "orange" and "oranges" are not counted as mentions of the fruit class, because they also name a color. The lists of surface forms and negation cues are part of the released code.

In the free caption, one prompt serves all 4 targets of an image, so the same decodes are scored against each target separately. The matcher misses a negation that follows the name, as discussed in Section 3.1. It does not by itself prevent a probe from succeeding, since the same matcher labels the one-sentence format, where the probes recover 31 to 78% (Table 1).

C Prompts

Table A1 lists the prompts and generation budgets, and Figure A1 places the six conditions on the two factors of Section 3.2. The image precedes the prompt in each model's chat template.

Table A1: Prompts and generation budgets (maximum new tokens). `{obj}` is the class name of the target.

Condition	Prompt	Budget
forced choice	`Is there a {obj} in the image? Answer yes or no.`	8
template	`Complete exactly this sentence with one word: "The image contains a {obj}: yes or no". Answer:`	12
one sentence	`In exactly one short sentence, describe whether a {obj} appears in this image.`	48
free caption, object named	`Describe this image in detail. In your description, also indicate naturally whether a {obj} is present.`	256
free caption, named late	`Write a detailed description of this image. Mention, where it fits naturally, whether you see a {obj}.`	256
free caption	`Describe this image in detail.`	256

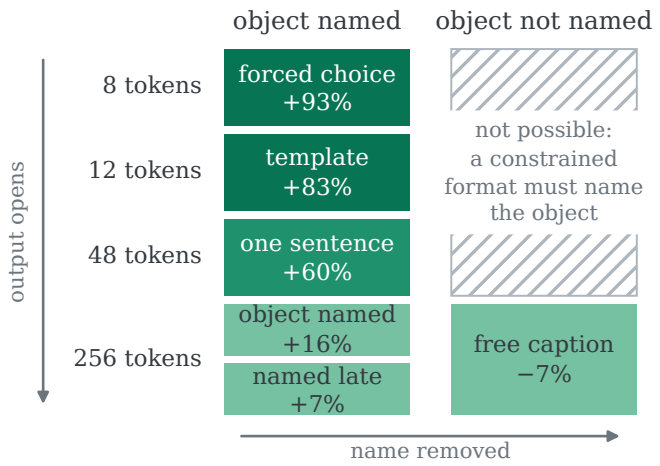


Figure A1: The six conditions on two factors. Moving down opens the output while the object stays named, and moving right removes the name while the output stays open. A constrained format has to name what it asks about, so the upper right is empty. Each box gives the mean recovery over the three models (Table 1), which also sets its shading.

D MODELS

Figure A2 shows how the three models pass the image to the language model, and Table A2 gives the checkpoints. Qwen2.5-VL and InternVL3 place visual tokens in the decoder's input sequence, while Llama-3.2-V passes visual features to the decoder through cross-attention layers. All three are used frozen, with their own chat templates.

In every model the probe reads the residual stream at the last prefill token after layers $\mathrm{round}(L/4)$, $\mathrm{round}(L/2)$ and L, where L is the number of decoder layers in Table A2. For Llama-3.2-V the count includes its cross-attention layers.

Table A2: Checkpoints and number of decoder layers L.

Model	Checkpoint	L
Qwen2.5-VL-7B	`Qwen/Qwen2.5-VL-7B-Instruct`	28
Llama-3.2-11B-Vision	`meta-llama/Llama-3.2-11B-Vision-Instruct`	40
InternVL3-8B	`OpenGVLab/InternVL3-8B-hf`	28

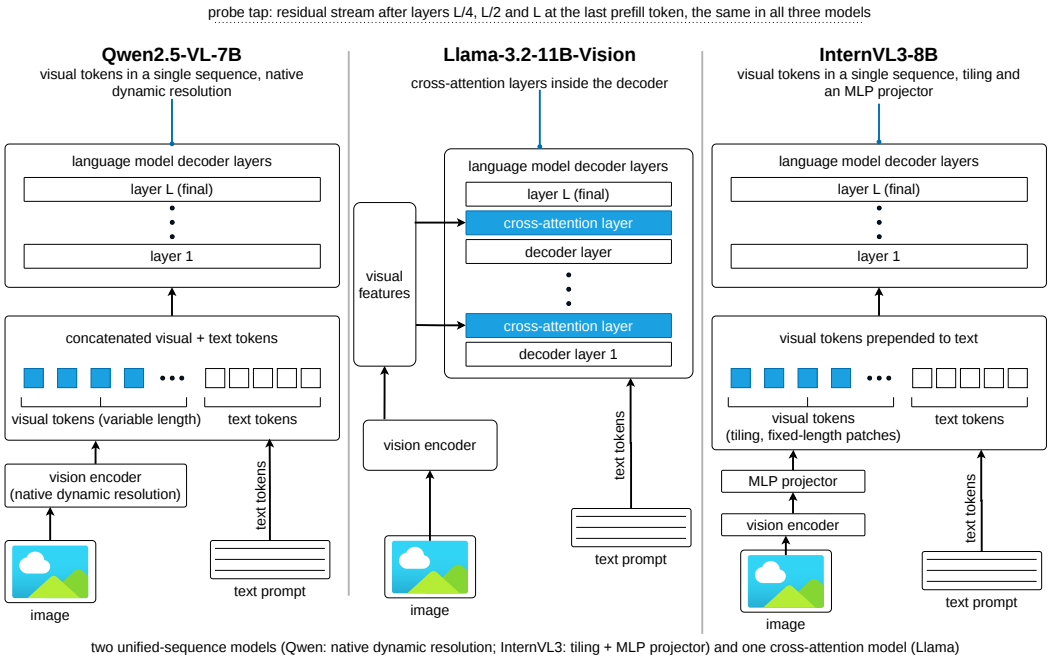


Figure A2: The three models. Qwen2.5-VL and InternVL3 place visual tokens in one sequence with the text, and Llama-3.2-V injects visual features through cross-attention layers inside the decoder. In all three, the probe reads the residual stream at the last prefill token after layers $L/4$, $L/2$ and L.

E Probe, Bootstrap and Controls

The probe's input is the residual stream at the last prefill token after the three layers of Appendix D, concatenated and stored in 16-bit floating point. The probe is fitted with 5-fold cross-fitting grouped by image, so the 4 queries of an image always fall in the same fold. In each training fold the features are standardized, reduced to 10 principal components and passed to an ℓ_2-regularized logistic regression with $C = 0.1$. These settings are fixed and the same for every model, condition and seed.

AUROC is computed on the pooled out-of-fold predictions, with ties counted as one half. Recovery is computed for each seed with that seed's ceiling and then averaged over seeds. For the intervals we draw 2,000 bootstrap samples of images. Each sample gives every image a multiplicity, every query takes the multiplicity of its image as a weight, and the weighted AUROC is recomputed for every seed and condition of a model with the same multiplicities. The 95% interval is the percentile interval of the resulting mean recovery, or of the difference between two conditions. The ceilings are held at their point estimates. For the two InternVL3 conditions marked in Table 1 the per-query predictions were not available, and we report the mean of the per-seed intervals instead.

The permutation test uses a separate linear probe, PCA to 64 components followed by logistic regression, with folds grouped by image. For every model, format and seed it is fitted on the true outcomes and on 200 random permutations of them, and $p = (1 + \#\{\text{null} \geq \text{observed}\})/(200 + 1)$. The nonlinear probes are trained on the same hidden states, with folds grouped by image. One is an MLP with hidden layers of 128 and 64 units, and the other is a histogram-based gradient-boosted tree ensemble. The unsupervised baseline projects each query's features onto their first principal component, computed without labels, and uses the projection as the score. Its sign is arbitrary, so a recovery close to -100% is a near-perfect ranking with the sign inverted.

F Sampling and Compute

Each query receives 16 decodes for $\hat{h}_0$ and one further decode for the outcome in every seed, with the budgets of Table A1. We set the budget, temperature 1.0 and top-p 1.0 in the Hugging Face

`generate` call and leave all other settings at their defaults. The library's default top-k of 50 therefore applies to all three models, and the generation configuration of Qwen2.5-VL adds a repetition penalty of 1.05. The seeds share prompts, images and features and differ only in the sampled decodes.

Most runs used an NVIDIA DGX Spark (GB10), and the remaining runs used one MIG slice of an NVIDIA H100. No model is fine-tuned. The code for all runs, statistics and figures is in the supplementary material.

G CEILING

The ceiling is the AUROC that an oracle would reach by scoring each query with its propensity. We estimate it by Monte Carlo from the 16-decode estimates. We draw 400,000 simulated queries by resampling the observed values of $\hat{h}_0$ with replacement, give each an outcome $y^\star \sim \mathrm{Bernoulli}(\hat{h}_0)$, and compute the AUROC of $\hat{h}_0$ against $y^\star$ on this sample. $A^\star$ is computed separately for each model, condition and seed.

Because $\hat{h}_0$ is itself an estimate, the simulation treats its sampling noise as real. Two queries with the same true propensity can differ in $\hat{h}_0$, for example when one is hallucinated once in 16 decodes and the other never, and the simulated oracle is rewarded for this difference. $A^\star$ therefore tends to exceed the ceiling of an oracle that knows the true propensity, most of all when events are rare. This shrinks recovery toward zero in both directions, which is why we also report raw AUROC (Table A3). The raw AUROC of the free-caption probes, 0.431 to 0.527, is close to chance whatever the ceiling.

In the free caption the probe sees one prefill state per image, so any score it produces is constant within an image. The natural score of this kind is the mean $\hat{h}_0$ of the image's targets. Against the outcome decode it reaches AUROC 0.90 to 0.91, and relative to this value the probe recovers -15% for Qwen2.5-VL, 7% for Llama-3.2-V and -16% for InternVL3.

H PER-CELL RESULTS

Table A3 lists the quantities behind Table 1 for every model and condition, and Figure A3 plots recovery across the formats. Queries whose propensity or outcome could not be computed are dropped, so some cells hold slightly fewer than 1,182 queries.

Events are rare under forced choice (0.9 to 5.6% of queries) and in the free caption (1.8 to 3.0%), and reach 10.3 to 28.4% when the open prompt names the object. The ceilings are lowest in that format (0.742 to 0.833), because more of its queries have propensities between zero and one, which makes single outcomes harder to predict.

Table A3: Per-cell diagnostics, averaged over decoding seeds: probe AUROC, the simulated ceiling $A^\star$, the event rate in the outcome decode, the number of events per seed and the number of seeds. Recovery and its intervals are in Table 1.

Model	Condition	AUROC	$A^\star$	events (%)	events per seed	seeds
Qwen2.5-VL-7B	forced choice	0.915	0.992	0.9	11	3
	template	0.956	0.994	1.7	20	3
	one sentence	0.604	0.833	8.3	98	3
	free caption	0.441	0.992	3.0	35	3
	free caption, object named	0.497	0.742	20.9	244	3
	free caption, named late	0.480	0.780	13.1	153	3
Llama-3.2-11B-V	forced choice	0.973	0.987	5.6	65	3
	template	0.836	0.904	13.7	159	3
	one sentence	0.783	0.897	8.4	98	3
	free caption	0.527	0.989	1.8	21	3
	free caption, object named	0.637	0.833	10.3	121	3
	free caption, named late	0.622	0.954	3.5	41	3
InternVL3-8B	forced choice	0.975	0.994	2.7	31	3
	template	0.856	0.977	4.1	48	3
	one sentence	0.865	0.970	4.6	54	3
	free caption	0.431	0.994	1.9	22	3
	free caption, object named	0.522	0.753	28.4	332	3
	free caption, named late	0.503	0.872	6.0	70	3

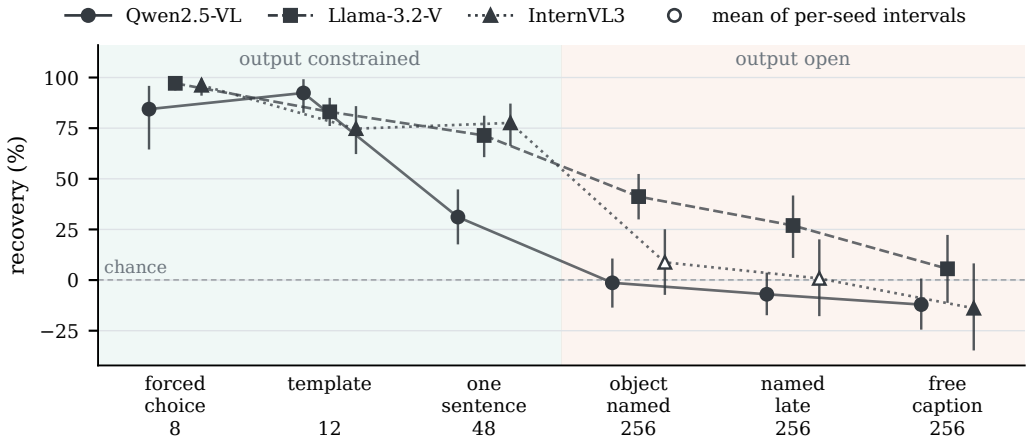


Figure A3: Recovery by condition for each model, from forced choice to the free caption. Points are means over decoding seeds, and bars are the 95% intervals of Table 1. Hollow markers are the two InternVL3 cells marked in Table 1, whose bars are the mean of the per-seed intervals.

I Layer and Position Sweep

The sweep uses the first seed. At the last prefill token it probes every odd-numbered decoder layer and the final layer, and at the final layer it probes positions -1, -2, -4, -8 and -16, counted back from the last prefill token at -1. Each cell gets its own linear probe and is scored with the seed's ceiling, with 95% intervals from a cluster bootstrap over images. Reading every layer with its own trained map is also how the tuned lens follows the formation of predictions across depth (Belrose et al., 2023).

Figure A4 shows the three models separately. Under forced choice, Llama-3.2-V carries some signal already in its first layers, and all three models reach a plateau in the middle of the network. In the free caption, most of the few cells that differ from zero lie below it, in line with the downward bias of pooled AUROC (Section 5.1).

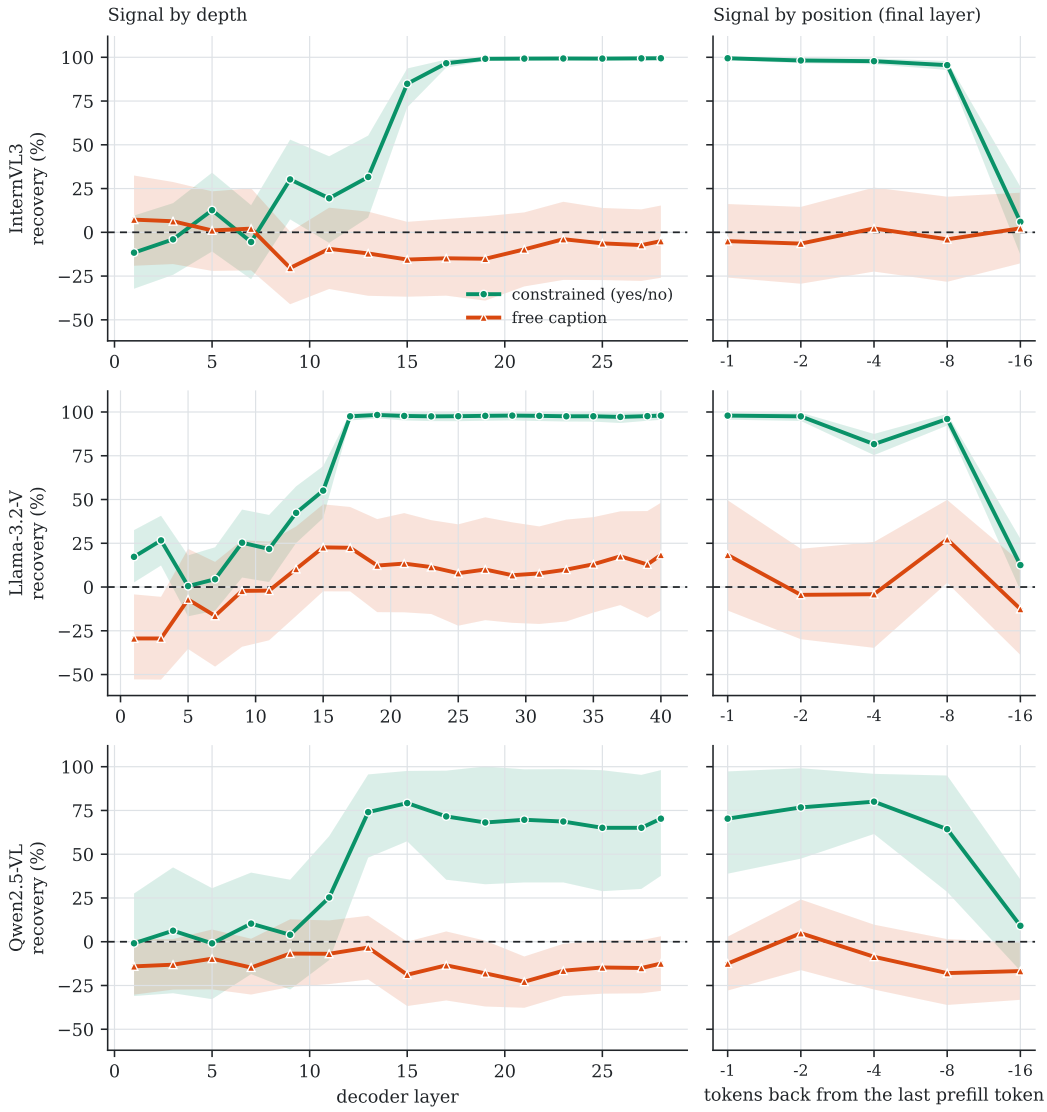


Figure A4: Layer and position sweep for each model (first seed, bands are 95% intervals). Left, single layers at the last prefill token. Right, the final layer at five positions before the end of the prefill.

J Label Audit

The audited labeler is the CHAIR-style matcher of an earlier version of this study, which labeled every object mentioned in a caption. The audit sample oversampled items that the labeler flagged. Two annotators judged every caption-object item against the image independently, and their 52 disagreements were adjudicated.

Table A4 gives the counts. The most frequent false-positive class, “dining table”, accounts for 33% of the false positives on its own.

Table A4: Human audit of the CHAIR-style labeler on 222 caption-object items, 104 of which the labeler flagged as hallucinated. Precision and recall are computed against the adjudicated labels, with exact 95% Clopper-Pearson intervals. They describe this sample, which oversampled flagged items, and not rates over all captions. κ and raw agreement compare the two annotators before adjudication.

Counts		Rates and agreement	
true positives	25	precision	0.24 [0.16, 0.33]
false positives	79	recall	0.56 [0.40, 0.70]
false negatives	20	inter-annotator κ	0.40
true negatives	98	raw agreement	77%
most frequent false-positive classes: dining table (26), sports ball (7), tv (7), snowboard (6), bowl (6)			