KNOWING WHERE WITHOUT READING WHAT: SALIENCY-INDUCED ACCESS FAILURE IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual question answering (VQA) requires models not only to locate task-relevant evidence but also to use it despite more salient competitors. Although multimodal large language models (MLLMs) excel at general VQA, their ability to exploit correctly localized evidence against salient distractors remains largely unexplored. To address this gap, we present **Decoy**, a benchmark designed to systematically evaluate MLLMs under conflicts between visual saliency and task relevance. DECOY contains 2,163 images from four eye-tracking datasets and 9,596 human-written question-answer pairs across five task types, each graded by a fixation-based saliency gap. Separate evaluations of Saliency Grounding, Target Grounding, and Answer Accuracy isolate suppression failures after correct grounding. Experiments on 24 MLLMs reveal that models know what is salient and where the evidence is, yet let the salient distractor dominate the answer: the lowest suppression failure rate is 37.7%, versus 4.9% for humans, and answering errors increase monotonically with the saliency gap. Beyond benchmarking, we uncover a gap between evidence localization and utilization. Restricting the input to target evidence is substantially more effective than indicating its location, and answer-relevant information remains decodable even as the causal influence of visual-region states fades with depth. These findings establish DECOY as a testbed for improving evidence utilization. The benchmark and code will be released upon publication.

1 INTRODUCTION

Natural images contain objects and regions with varying visual saliency. Visual question answering (VQA) requires models to locate task-relevant evidence and use it to derive the answer. However, task-relevant evidence may appear in less salient regions and be visually overshadowed by more salient but task-irrelevant content. Although human attention can shift from stimulus-driven salience to goal-directed search (Itti et al., 1998; Wolfe, 1994), whether multimodal large language models (MLLMs) can reliably exploit task-relevant evidence under such competition remains largely unexplored. As illustrated in Figure 1, the model correctly identifies the platform sign as the most salient object and localizes the relevant evidence on the train, yet still omits the train marking from its final answer, exposing the gap between evidence localization and utilization in current MLLMs.

Existing MLLM benchmarks have not systematically examined this problem. Hallucination benchmarks (Rohrbach et al., 2018; Li et al., 2023; Guan et al., 2024) evaluate final answer-image consistency and cannot distinguish failed localization from ineffective use of correctly localized evidence. Distraction-robustness benchmarks introduce irrelevant images or text into the input (Yang et al., 2026; Liu et al., 2025b), measuring robustness to artificial noise rather than competition among objects within an image. Moreover, existing benchmarks lack a model-independent measure of conflict strength, as task difficulty is commonly inferred post hoc from model errors. Human eye-tracking data provide a natural measure of visual competition, but have not been systematically used to evaluate how MLLMs utilize task-relevant evidence under salient distraction. These limitations prevent existing benchmarks from separating failures of evidence localization from failures of utilization under salient distraction.

To address these gaps, we introduce **Decoy**, a benchmark designed to diagnose conflicts between visual saliency and task relevance in MLLMs. DECOY contains 2,163 natural images from four eye-tracking datasets and 9,596 human-written question-answer pairs across five task types. Each

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.

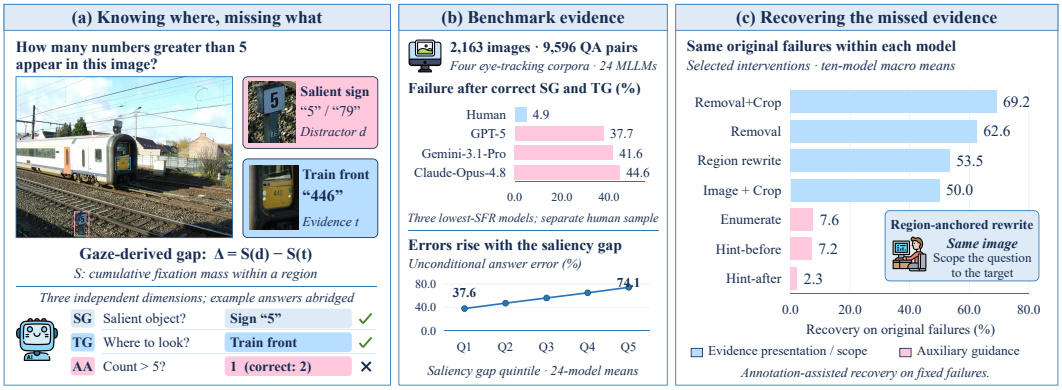


Figure 1: **Overview of DECOY. (a)** Despite correct SG and TG, the model misses the train marking when counting numbers greater than five. **(b)** Conditional SFR for the three lowest-SFR models and a separate human sample; unconditional answer error across global saliency-gap quintiles. **(c)** Annotation-assisted recovery on fixed per-model failure sets, averaged equally over ten models.

question pairs a task-relevant target with a salient distractor in the same photograph, and a saliency gap computed from human fixations quantifies their conflict prior to model evaluation. Three annotators label every image independently, with disagreements resolved by adjudication. DECOY also provides 400 congruent controls, in which the salient object supplies the answer.

Unlike previous benchmarks that score only the final answer, DECOY evaluates three independent dimensions: Saliency Grounding (SG), Target Grounding (TG), and Answer Accuracy (AA). An incorrect AA response despite correct SG and TG responses constitutes a *suppression failure*, in which the model locates the task-relevant evidence yet fails to utilize it. This conditional design separates the use of evidence from its localization, and the fixation-based gap further allows failures to be analyzed as a function of conflict strength. We benchmark 24 representative proprietary and open-source MLLMs on DECOY. Current MLLMs still struggle to utilize evidence they have correctly grounded, with the lowest suppression failure rate reaching 37.7% compared with 4.9% for humans. Controlled interventions and mechanistic analyses trace failures to a gap between evidence localization and utilization.

Our main contributions are summarized as follows:

- We introduce DECOY, a benchmark for diagnosing conflicts between visual saliency and task relevance in MLLMs, comprising 2,163 natural images and 9,596 human-written question-answer pairs with fixation-based saliency gaps and three diagnostic dimensions.
- We benchmark 24 MLLMs on DECOY and reveal substantial suppression failures relative to humans, with answering errors increasing monotonically with the saliency gap and decreasing when the distractor is attenuated within the same image.
- We compare eleven interventions on identical failure sets, showing that removing the distractor or restricting the question to the target region recovers over half of the suppression failures, whereas location hints recover under one tenth. Mechanistic analyses further show that answer-relevant information remains decodable, while the causal influence of visual-region states diminishes with decoder depth.

2 RELATED WORK

2.1 GENERAL AND HALLUCINATION BENCHMARKS

General benchmarks such as MME (Fu et al., 2023) and MMBench (Liu et al., 2024) evaluate MLLMs across taxonomies of perception and reasoning abilities. Hallucination benchmarks narrow the focus to visual grounding: CHAIR (Rohrbach et al., 2018) counts caption objects absent from the image annotations, POPE (Li et al., 2023) polls object existence through binary questions, AMBER (Wang et al., 2023) extends this probing to attributes and relations, and HallusionBench (Guan et al., 2024) pairs unedited and edited images to separate language priors from visual illusion. These benchmarks

judge the final answer alone, leaving saliency grounding and target grounding unmeasured. In contrast, DECOY evaluates SG, TG, and AA independently and measures answer errors conditional on correct saliency grounding and target grounding.

2.2 DISTRACTION ROBUSTNESS BENCHMARKS

Distraction benchmarks evaluate whether MLLMs remain reliable when the input contains task-irrelevant content. Yang et al. (2026) add irrelevant images and text to multimodal tasks, Liu et al. (2025b) perturb inputs with visual and textual distractions, and Cai et al. (2025) diagnose interference from signals in a competing modality. In each case the distractor is constructed by the benchmark designer, and its strength is fixed by the chosen perturbation. DECOY instead draws the salient distractor and the task-relevant target from the same unedited photograph and grades their competition using the fixation-based saliency gap.

2.3 HUMAN GAZE AND MLLM REPRESENTATIONS

Eye-tracking corpora such as MIT1003 (Judd et al., 2009) and OSIE (Xu et al., 2014) record free-viewing fixations and have mainly supported the training and evaluation of saliency prediction models. Christian et al. (2026) compare human and model attention maps, assessing the alignment of attention itself rather than whether answers are drawn from the attended region. Other studies examine internal representations: Fu et al. (2025) find that VLMs can underperform linear readouts of their own visual features, and Liu et al. (2025a) trace visual information inside the language model. DECOY is the first to derive conflict strength for MLLM evaluation from recorded human fixations, and further probes and patches visual-token states on the resulting suppression failures.

3 THE DECOY BENCHMARK

3.1 OVERVIEW OF DECOY

In this section, we introduce DECOY, a benchmark for diagnosing how MLLMs utilize task-relevant evidence in the presence of a salient distractor. DECOY consists of $2{,}163$ natural images from four eye-tracking datasets and $9{,}596$ human-written question-answer pairs across five task types (Figure 2, top; per-corpus counts in Appendix A). Of these questions, $9{,}196$ are incongruent, pairing a task-relevant target with a salient distractor in the same photograph, and 400 are congruent controls, in which the target is itself the most salient object in the image.

3.2 IMAGE COLLECTION

Measuring the conflict between saliency and task relevance requires images in which human attention has been recorded independently of any model. We therefore draw candidate images from MIT1003 (Judd et al., 2009), OSIE (Xu et al., 2014), PASCAL-S (Li et al., 2014), and six natural-photograph categories of CAT2000 (Borji & Itti, 2015) (*Action*, *Indoor*, *Object*, *OutdoorManMade*, *Outdoor-Natural*, and *Social*), all of which provide raw human fixation coordinates. An image is retained when a dominant fixated object coexists with a distinct and visually resolvable target that carries task-relevant evidence (Figure 2, bottom (a)), so the distractor is an unedited part of the scene and its prominence reflects how observers actually allocate attention. Annotators then mark a box d around the salient distractor and a box t around the task-relevant target, and assign one primary task type. To turn the recorded fixations into a measure of competition between the two boxes, we reconstruct each fixation map with a Gaussian kernel density estimate of bandwidth $\sigma = 0.025$ times the image diagonal, applying the same estimator to all corpora. With $S(r)$ denoting the cumulative fixation mass inside box r, the saliency gap $\Delta = S(d) - S(t)$ grades how strongly the distractor outcompetes the target, and it is fixed before any model is evaluated.

3.3 FINE-GRAINED TASK DEFINITION

The saliency gap characterizes the stimulus, yet a final answer alone cannot reveal whether an error arises from missing the target or from failing to extract its answer content. DECOY therefore evaluates three independent dimensions with no shared conversation history. *Saliency Grounding (SG)* identifies

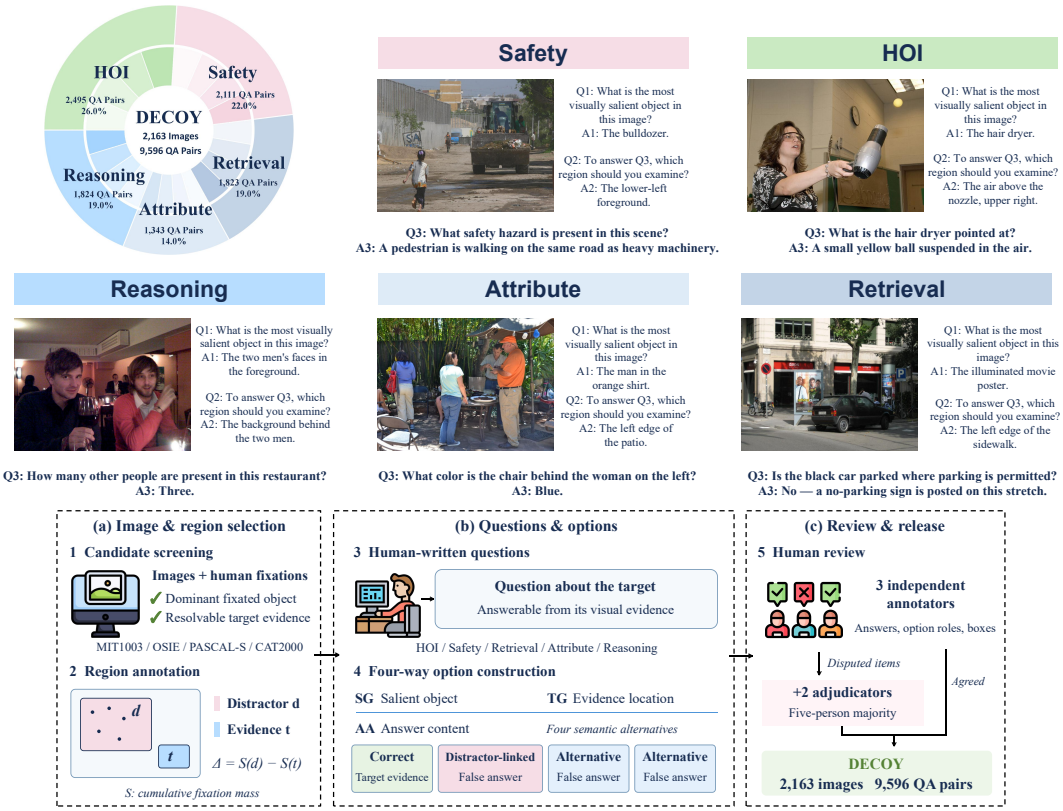


Figure 2: **Task coverage and construction of DECOY. Top:** task distribution of 9,596 questions and five examples; Q1–Q3 correspond to SG, TG, and AA. **Bottom:** (a) fixation-based image screening and distractor/target annotation; (b) human-written questions and four-way options, with location-only TG choices and a distractor-linked AA alternative; (c) independent annotation, adjudication, and the resulting benchmark.

the most salient object in the image. *Target Grounding (TG)* presents the task question and asks which region should be consulted. *Answer Accuracy (AA)* measures whether the model answers the task question correctly. All three dimensions use four-way multiple choice, with a nominal chance accuracy of 25%. Let $c_{\mathrm{SG}}, c_{\mathrm{TG}}, c_{\mathrm{AA}} \in \{0, 1\}$ indicate correct responses in the three dimensions. A *suppression failure* is the event $c_{\mathrm{SG}} = c_{\mathrm{TG}} = 1$ and $c_{\mathrm{AA}} = 0$. Its rate over incongruent questions is

$$\mathrm{SFR} = \Pr\left(c_{\mathrm{AA}} = 0 \mid c_{\mathrm{SG}} = 1,\ c_{\mathrm{TG}} = 1\right). \tag{1}$$

This conditional metric diagnoses a discrepancy between localization and answering across queries. Recognition, reasoning, option interpretation, and output-format errors can also contribute.

How a salient distractor interferes with evidence utilization depends on what the target must supply, which motivates five task types (examples in Figure 2, top). In *HOI* and *Safety* questions, the decisive detail is the object of an interaction or a hazard, while the actor, tool, or machinery that dominates the scene attracts most fixations. *Retrieval* questions ask whether an entity is present and thus require search beyond the distractor, whereas *Attribute* questions ask for a property on which the distractor carries a competing value. *Reasoning* questions spread the evidence across several regions, so no single region suffices. Task type and saliency gap are recorded as separate variables, and every type contains questions with both small and large saliency gaps.

3.4 QUESTION-ANSWERING ANNOTATION

A wrong answer is most informative when it reveals which region the model has read. Instead of drawing all false options from generic alternatives, we therefore anchor one AA option to the salient distractor. This *distractor-linked* option is the answer obtained by consulting the distractor instead of the target. In the Attribute example of Figure 2 (top), the correct answer is blue, whereas

the distractor-linked option, orange, comes from the salient distractor's shirt; the remaining two options are plausible alternatives. All questions are written by humans about the target and must be answerable from its visual evidence (Figure 2, bottom (b)). The SG and TG options separate grounding from answer content: SG false options exclude the targets of the image's questions, and TG options specify locations without revealing the answer. Their incorrect locations exclude all annotated targets and prioritize a salient distractor, the nearest non-overlapping region, and an area-matched non-target region; Appendix A.4 specifies fallback and deduplication rules. Applying the same procedure with the most salient object as the target yields the 400 congruent controls, balanced across corpora (100 each) and task types (80 each), which keep each photograph fixed while aligning saliency with relevance, analogous to congruent trials in the classic color-word Stroop paradigm (Stroop, 1935).

Because the diagnosis depends on the correctness of every option and region, each image is labeled by three independent annotators, covering the answers, option roles, and region boxes (Figure 2, bottom (c)). Disputed items are reviewed by two additional adjudicators, and the final decision is made by majority over all five annotators. Before adjudication, Fleiss' κ (Fleiss, 1971) reaches 0.84 for task type, 0.89 for the correct answer, and 0.82 for the distractor-linked option, the last of which shows that annotators consistently agree on which false option follows from the distractor (Appendix C).

4 Benchmark Results

4.1 Evaluation Setup

We conduct comprehensive experiments on 24 MLLMs on Decoy. The models fall into two groups: (1) proprietary models, including GPT-5, Gemini-3.1-Pro, and Claude-Opus-4.8, and (2) open-source models, including the Qwen3-VL, InternVL3.5, and LLaVA series (Table 1). All models are evaluated with the same prompt template provided in Appendix D. Increasing the generation budget for Enumerate and Thinking models reduces truncation and parsing failures without materially changing the main benchmark results (Appendix D.1).

Separating localization accuracy from conditional answering error requires dimension-specific diagnostics. Beyond SG, TG, and AA, we report Eligibility, the fraction of incongruent questions with correct SG and TG responses, and the SFR in Eq. 1, with uncertainty clustered by image. We collect answers from 26 participants on 350 stratified questions under the same protocol, and include a text-only run of GPT-5.6 Sol (OpenAI, 2026), a frontier model outside the main evaluation, to measure how far language priors alone can go, together with chance-level accuracy.

4.2 Main Results

4.2.1 Overall Performance

As illustrated in Table 1, proprietary models achieve the lowest suppression failure rates, led by GPT-5 at 37.7%, while the open-source MolmoPoint-8B even outperforms Gemini-2.0-Flash on this measure. Even the lowest-SFR model fails on more than one third of the questions whose evidence it has located, nearly eight times the human rate of 4.9%, and the gap persists on the questions eligible for all 24 models (Appendix E). Interestingly, higher accuracy does not imply better use of evidence, as Gemini-3.1-Pro attains the highest AA yet fails to exploit located evidence more often than GPT-5. Scaling helps only modestly: quadrupling Qwen3-VL from 2B to 8B lowers its SFR by fewer than six points. Meanwhile, GPT-5.6 Sol answers only 27.3% of the incongruent questions without the image, barely above the 25% chance rate, so language priors contribute little to the answers. Current MLLMs thus locate the evidence but still struggle to exploit it effectively.

4.2.2 Task-wise Performance

From the task-wise results (Appendix E), it is evident that difficulty grows as the salient distractor moves from competing for attention to competing for the answer. Models handle Safety and Retrieval best, since the distractor there only draws fixation away from a target whose presence or state decides the answer. The challenge becomes more pronounced for Attribute and HOI, where the distractor offers a plausible answer of its own, such as a different color or the actor of the interaction. This issue

Table 1: Diagnostic accuracy (%) from our evaluation of 24 models; citations identify model sources. SG, TG, and AA macro-average five task types, including congruent controls. Eligibility and SFR use 9,196 incongruent questions; SFR conditions on correct SG and TG. Humans answer 350 separate stratified questions. Bold/underlined values mark the best/second-best model scores per column, excluding humans.

Model	SG	TG	AA	Elig.	SFR ↓
Open-source					
XComposer2-7B (Dong et al., 2024)	75.3	78.3	44.6	59.4	56.8
VILA-8B (Lin et al., 2023)	76.8	79.8	45.8	61.2	55.3
Llama3.2-Vision-11B (Meta AI, 2024)	80.1	83.1	47.0	66.1	53.5
mplug-owl-7B (Ye et al., 2023)	68.3	71.1	35.3	48.3	64.0
MiniCPM-V-4.6 (OpenBMB, 2026)	82.1	84.8	49.6	69.2	51.2
Gemma3-4B (Gemma Team, 2025)	74.8	77.8	41.0	58.1	58.6
Phi-4-Multimodal (Abouelenin et al., 2025)	81.5	84.4	48.8	68.4	52.0
DeepSeek-VL2-Tiny (Wu et al., 2024)	68.8	71.9	36.5	49.1	62.7
LLaVA-v1.5-7B (Liu et al., 2023)	65.1	68.9	38.5	45.6	60.9
LLaVA-OneVision-2-8B (An et al., 2026)	79.2	82.3	46.4	64.8	54.4
Qwen3-VL-2B (Bai et al., 2025)	70.3	73.3	35.8	51.2	62.5
Qwen3-VL-4B (Bai et al., 2025)	73.8	77.1	38.5	56.9	59.8
Baseline					
Human	97.0	95.2	92.4	94.3	4.9

Model	SG	TG	AA	Elig.	SFR ↓
Open-source (continued)					
Qwen3-VL-8B (Bai et al., 2025)	77.6	81.5	41.0	62.7	56.8
Qwen3-VL-8B-Thinking (Bai et al., 2025)	78.4	82.1	42.9	64.1	54.9
InternVL3.5-2B (Wang et al., 2025)	72.8	76.1	41.9	55.7	57.2
InternVL3.5-4B (Wang et al., 2025)	76.3	79.8	44.6	61.0	54.8
InternVL3.5-8B (Wang et al., 2025)	79.4	83.0	47.2	65.6	52.0
Molmo2-8B (Clark et al., 2026b)	80.2	83.6	49.3	67.2	51.4
MolmoPoint-8B (Clark et al., 2026a)	82.8	86.1	52.2	71.1	48.2
Proprietary					
GPT-4o (Hurst et al., 2024)	83.0	86.3	53.9	72.0	47.0
GPT-5 (OpenAI, 2025)	84.5	87.7	57.7	74.3	**37.7**
Gemini-2.0-Flash (Google DeepMind, 2025)	82.3	85.5	50.7	70.2	49.8
Gemini-3.1-Pro (Google DeepMind, 2026)	**84.7**	**88.0**	**58.5**	**75.0**	41.6
Claude-Opus-4.8 (Anthropic, 2026)	84.0	87.2	55.8	73.1	44.6
Chance accuracy: 25% at each dimension.					

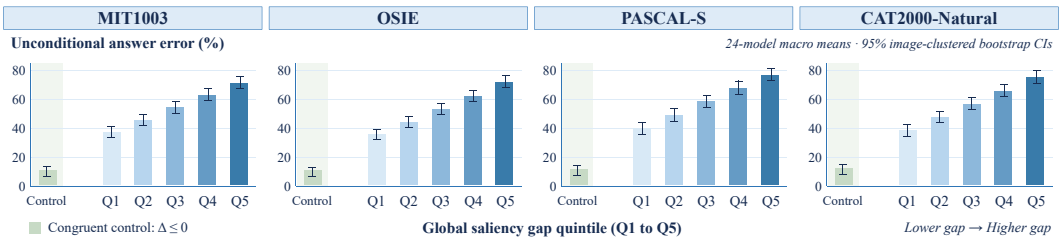


Figure 3: **Saliency-gap gradients across source corpora.** Unconditional answer error, averaged equally over 24 models. Q1–Q5 use global saliency-gap cut points, with corpus-specific counts. Congruent controls align saliency and relevance. Error bars show 95% image-clustered bootstrap confidence intervals.

is most evident in Reasoning, the hardest type for most models, where the evidence is spread across several regions and missing any one of them yields a wrong answer.

4.3 Error Analysis

We further observe that failures intensify as the salient distractor grows more dominant. As shown in Figure 4(a), the model-averaged answer error climbs steadily from 37.6% in the lowest quintile of Δ to 74.1% in the highest, and the same trend holds in every corpus (Figure 3). Humans, by contrast, show little such sensitivity: their SFR stays between 4.1% and 5.8% across quintiles, whereas the model-averaged SFR rises from 39.6% to 68.4% (Appendix C). The association also persists under joint geometric controls and when restricted to eligible or successfully parsed questions (Appendix F).

Since this trend is observational, we directly manipulate the distractor while keeping the question, target, and options fixed, on 1,297 questions over 318 images. Strikingly, merely reducing the color, contrast, and sharpness of the distractor, without removing it, lowers answer error by 13.5 percentage points on average across seven models, whereas applying the same edit to an irrelevant region has almost no effect (Figure 4(b); Table 2). The wrong answers also reveal which region models draw on: across the seven models analyzed in Section 5.2, 48.6% of incorrect answers select the distractor-linked option, compared with 33.3% if errors were spread evenly over the three false

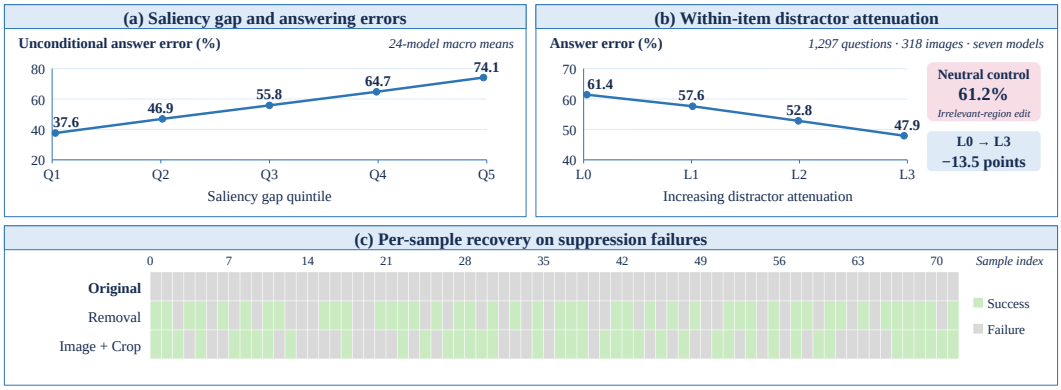


Figure 4: **Saliency competition and recovery. (a)** Answer error across global saliency-gap quintiles, averaged over 24 models. **(b)** On 1,297 questions from 318 fixed images, attenuation reduces seven-model mean error by 13.5 points; L3-strength irrelevant-region editing yields 61.2% error. **(c)** Qwen3-VL-8B on 72 OSIE suppression failures, randomly sampled (seed 23) among cases with valid outputs for all three inputs: original image, Removal, and Image + Crop; the original row fails by construction.

Table 2: **(a)** Answer error (%) under distractor attenuation: 1,297 questions from 318 fixed images per model. Neutral: L3-strength irrelevant-region editing; Preference: 24-rater check; slopes: percentage points per level. **(b)** Candidate-scoring probe (Alain & Bengio, 2016) and activation patching (Zhang & Nanda, 2024). ℓ/L: probe layer/decoder length; Acc.: content-probe accuracy; $R_{.75}$: recovery at probe depth (both %); $d_{1/2}$: interpolated half-shallow recovery depth. Probing uses 9,196 incongruent questions per model; patching uses 12,604 Removal-recovered model-question pairs across seven models.

(a) Within-item attenuation

Model	L0	Neutral	L1	L2	L3	Slope
Preference (%)	73.6	72.8	64.9	56.7	48.1	N/A
mplug-owl-7B	68.6	67.9	63.8	59.1	53.8	-4.91
Qwen3-VL-2B	67.0	66.4	62.3	56.9	52.2	-4.97
LLaVA-v1.5-7B	64.2	63.5	60.4	55.0	50.6	-4.59
Qwen3-VL-8B	61.3	61.6	57.5	52.5	47.5	-4.65
Qwen3-VL-8B-Thinking	59.1	58.5	55.7	51.3	46.5	-4.21
InternVL3.5-8B	56.6	56.9	53.1	48.4	43.4	-4.43
MolmoPoint-8B	53.1	53.8	50.3	46.2	41.2	-3.99
Macro mean	**61.4**	**61.2**	**57.6**	**52.8**	**47.9**	−4.54

(b) Decodability and causal efficacy

Model	ℓ/L	Acc.	$R_{.75}$	$d_{1/2}$
mplug-owl-7B	24/32	67.6	17.2	0.64
Qwen3-VL-2B	21/28	66.4	20.5	0.66
LLaVA-v1.5-7B	24/32	69.5	23.8	0.68
Qwen3-VL-8B	27/36	69.7	29.0	0.71
Qwen3-VL-8B-Thinking	27/36	71.1	35.0	0.74
InternVL3.5-8B	27/36	72.4	39.8	0.76
MolmoPoint-8B	27/36	76.8	47.4	0.79
Macro mean	N/A	**70.5**	**30.4**	**0.71**

options. This suggests that MLLMs do not weigh visual evidence by its relevance to the question as humans do; instead, their answers still gravitate toward whatever is most salient in the scene.

5 Intervention and Mechanistic Analysis

In this section, we investigate why MLLMs fail to use evidence they have already located, through two complementary analyses: controlled interventions that test what restores the correct answer, and probing and activation patching that trace where the failure arises inside the model.

5.1 Evidence Presentation and Scoping

5.1.1 Intervention Design

As analyzed in Section 4.3, the salient distractor pulls model answers away from evidence they have already located. If the failure stems from losing track of where the evidence lies, restating its location should repair it. If it instead stems from competition within the visual input, the answer should recover only when the input is narrowed to the evidence. We test both hypotheses on the

Table 3: Recovery (%) on each model's original suppression failures. Inpainting uses LaMa (Suvorov et al., 2022); Enumerate uses chain-of-thought-inspired verbalization (Wei et al., 2022). Crop retains the original image; Rem.+C combines removal and cropping; Rewrite scopes the question to the target. Pre/Post denote hint position, Enum. Enumerate, and TG fb. TG feedback. Summary N_{SF} sums model-question failures; percentages are unweighted model means.

Model	N_{SF}	Controls			Scoping				Guidance			
		Re-ask	Inpaint	Prefix	Removal	Crop	Rem.+C	Rewrite	Enum.	Post	Pre	TG fb.
mplug-owl-7B	2,843	0.0	1.8	2.3	55.8	42.4	62.4	40.7	5.8	1.3	5.8	2.5
Qwen3-VL-2B	2,943	0.0	2.0	2.0	56.3	43.8	63.1	43.5	5.4	1.2	5.1	2.8
LLaVA-v1.5-7B	2,554	0.0	1.6	1.8	57.0	41.6	64.0	42.3	5.2	0.9	4.7	2.2
Qwen3-VL-8B	3,275	0.0	1.9	2.3	58.0	46.0	65.4	50.9	6.1	2.0	6.6	3.4
Qwen3-VL-8B-Thinking	3,236	0.0	2.1	2.6	58.6	45.8	66.2	53.4	6.7	2.1	7.1	3.6
InternVL3.5-8B	3,137	0.0	2.2	2.8	61.5	50.1	68.7	54.3	7.4	2.4	8.2	4.0
MolmoPoint-8B	3,151	0.0	2.8	3.3	69.2	57.0	75.6	59.8	12.6	3.9	10.3	4.9
Claude-Opus-4.8	2,998	1.4	2.0	2.5	65.8	53.7	72.1	62.1	8.2	2.7	7.5	4.3
Gemini-3.1-Pro	2,869	1.6	2.4	2.9	70.4	58.4	76.8	63.5	9.6	3.2	8.8	5.0
GPT-5	2,576	1.3	2.2	2.7	73.0	61.0	77.6	64.2	9.1	3.0	8.0	4.8
Macro mean	**29,582**	**0.4**	**2.1**	**2.5**	**62.6**	**50.0**	**69.2**	**53.5**	**7.6**	**2.3**	**7.2**	**3.8**

fixed suppression-failure set of each of ten models, measuring recovery as the fraction of these failures answered correctly (Appendix H). Guidance conditions keep the input unchanged and add location information or inspection prompts (Wei et al., 2022), such as a location hint or the model's own correct TG answer. Scoping conditions narrow the input by removing the distractor through inpainting (Suvorov et al., 2022), appending a target crop (Pan et al., 2026), or rewriting the question to refer to the target region. Controls measure recovery from re-querying or editing alone.

5.1.2 Recovery Results

As shown in Table 3, scoping recovers more than half of the failures, reaching 69.2% when removal is combined with a target crop. Removing the distractor accounts for most of this gain (Figure 4(c)). Interestingly, region-anchored rewriting is purely textual yet recovers 53.5%, on par with a target crop, so the benefit comes from restricting scope rather than from extra visual input.

Guidance, by contrast, recovers less than one tenth of the failures on average, only slightly above the controls. Even returning the model's own correct localization recovers just 3.8%, although it produced that localization itself. The results favor the second hypothesis. Telling a model where evidence lies is not enough; the failure is repaired mainly when its input is confined to that evidence.

5.2 Decodability and Causal Influence

5.2.1 Probing and Patching Setup

Interventions reveal what repairs a failure, but not where it arises inside the model. We therefore examine the internal states of seven open-source models on all $9{,}196$ incongruent questions. Under causal attention, the visual tokens precede the options and carry no information about them. The content probe scores each AA candidate with a shared linear scorer over visual-candidate interaction features. Each feature combines the pooled target state at relative depth 0.75 with the candidate's representation from the same forward pass (Appendix I). Decodability, however, does not show that these states influence the answer. We therefore apply activation patching (Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2024), which needs a counterpart run that answers correctly. The model's own Removal run on the same question provides this counterpart, and patching replaces the distractor-region states of the failed run with those of the Removal run at seven relative depths (Appendix J).

5.2.2 Decodability without Use

The probe identifies the correct candidate on 70.5% of questions, far above chance. Its accuracy is the same whether the model answers correctly or not, and its success does not predict the model's own answer (Table 4(b)). The errors diverge as well. When the probe errs, it selects the distractor-linked option 34.0% of the time, close to the 33.3% expected if errors were spread evenly. The same

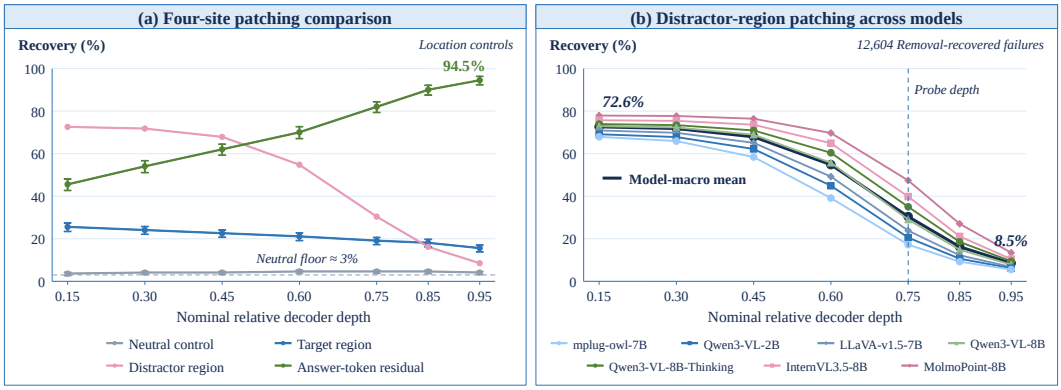


Figure 5: **Activation patching across depths.** Model-macro recovery on 12,604 Removal-recovered model-question pairs across seven models, fixed across depths. **(a)** 95% image-clustered bootstrap intervals for neutral, target-region, and answer-token residual patches; distractor point estimates. **(b)** Distractor-region curves by model; the thick mean matches (a). Nominal depths map to nearest decoder layers (Appendix J.2).

Table 4: (a) Distractor-region patching across attenuation levels. Pairs: eligible source-destination pairs, varying by level; $R_{.15}/R_{.95}$: recovery (%) at nominal depths 0.15/0.95. (b) Candidate-scoring probe versus model AA correctness on 64,372 incongruent model-question pairs across seven models. Model accuracy is 39.9% in both probe groups; probe accuracy is 70.5%.

(a) Patching across attenuation levels

Level	Pairs	$R_{.15}$	$R_{.95}$
L0	1,342	73.4	9.8
L1	1,251	71.6	10.9
L2	1,137	69.3	12.7
L3	1,024	66.8	14.1

(b) Probe / model outcomes

	Correct	Wrong
Probe correct	18,110	27,272
Probe wrong	7,574	11,416
Total	25,684	38,688

seven models, in contrast, choose that option in 48.6% of their errors (Appendix I). Answer-relevant information therefore survives inside the model, while the final answer is pulled toward the distractor.

5.2.3 Patching Efficacy Declines with Depth

Why, then, does this information fail to reach the answer? Replacing distractor-region states restores 72.6% of the failures at the shallowest depth, but only 8.5% at the deepest (Figure 5(b)), and stronger models retain this influence longer (Table 2(b)). Weaker conflict is also associated with a later decline on the attenuation ladder, although the eligible pairs vary across levels (Table 4(a)). The answer itself remains modifiable: patching the answer-token residual still recovers 94.5% late in decoding (Figure 5(a)). What fades is the route from the visual region to the answer. Models locate the evidence and retain answer-relevant information. However, the causal influence of visual-region states fades as decoding deepens, and the answer stays anchored to the salient distractor.

6 Conclusion

In this work, we introduce DECOY, a benchmark for diagnosing how MLLMs use evidence that competes with a salient distractor. DECOY pairs 2,163 images from four eye-tracking datasets with 9,596 human-written question-answer pairs, grading each conflict by human fixations and separating grounding from answering through SG, TG, and AA. Benchmarking 24 MLLMs reveals a persistent gap between evidence localization and utilization, with answering errors growing with the prominence of the distractor. Confining the input to the evidence repairs many of these failures, whereas restating its location rarely does. Inside the model, answer-relevant information remains decodable, yet the causal influence of visual-region states fades with depth. Overall, DECOY provides a rigorous yardstick for both evaluating and improving evidence utilization in MLLMs.

AI USAGE STATEMENT

Large language model tools assisted with manuscript preparation, including language refinement, sentence rephrasing, grammar correction, and improvements to readability and organization. They also assisted with LaTeX formatting, cross-reference checks, and consistency checks across the text, tables, and figure captions. These checks support manuscript preparation and do not independently validate the experimental findings. The authors retain responsibility for the research design, data, experiments, analyses, references, and all claims in the manuscript, including any AI-assisted wording. The MLLMs evaluated in this study, including the text-only baseline, are experimental systems whose outputs are assessed under the reported evaluation protocol.

REPRODUCIBILITY STATEMENT

The appendix provides the detailed protocols and results needed to inspect and reproduce the study. Appendices A–C document source-corpus selection, question and candidate construction, fixation-map estimation, annotation adjudication, and human evaluation. Appendix D records model identifiers, query accounting, decoding settings, and parsing rules. Appendices E and F provide full diagnostic results and statistical sensitivity analyses.

For the intervention experiments, Appendices G and H specify attenuation and inpainting parameters, manipulation checks, crop construction, fixed failure sets, and language templates. Appendices I and J describe candidate scoring, visual-token pooling, image-grouped splits, probe training and depth selection, patch-source construction, layer mappings, and decay-depth calculations. These details distinguish the populations and aggregation rules used by each analysis. The benchmark and code will be released to support replication.

ETHICS STATEMENT

This study uses images and previously recorded fixation data from MIT1003, OSIE, PASCAL-S, and CAT2000. Human contributions include question annotation, benchmark evaluation, and an attention-manipulation check, as detailed in Appendices C and G. The benchmark concerns model behavior on visual tasks and is not intended for identifying individuals, inferring sensitive personal attributes, or making consequential decisions about people depicted in the images. Release of benchmark materials will respect the source datasets' usage and redistribution terms and protect participant-identifying information from the human evaluations.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs, 2025.

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. *arXiv preprint arXiv:1610.01644*, 2016.

Xiang An et al. LLaVA-OneVision-2: Towards Next-Generation Perceptual Intelligence, 2026.

Anthropic. System Card: Claude Opus 4.8. https://www-cdn.anthropic.com/0b4915911bb0d19eca5b5ee635c80fef830a37ea.pdf, May 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report, 2025.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Ali Borji and Laurent Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. *arXiv preprint arXiv:1505.03581*, 2015.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019. doi: 10.1109/TPAMI.2018.2815601.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and Mitigating Modality Interference in Multimodal Large Language Models. *arXiv preprint arXiv:2505.19616*, 2025.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317.

Isaac R. Christian, Udith Haputhanthrige, Hanna Hornfeld, Declan Campbell, Samuel Nastase, Taylor Webb, and Michael Graziano. Attention Alignment Between Humans and Vision-Language Models. *arXiv preprint arXiv:2606.17410*, 2026.

Christopher Clark, Yue Yang, Jae Sung Park, Zixian Ma, Jieyu Zhang, Rohun Tripathi, Mohammadreza Salehi, Sangho Lee, Taira Anderson, Winson Han, and Ranjay Krishna. MolmoPoint: Better Pointing for VLMs with Grounding Tokens. *arXiv preprint arXiv:2603.28069*, 2026a. URL https://arxiv.org/abs/2603.28069.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding, 2026b.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model, 2024.

Joseph L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in Plain Sight: VLMs Overlook Their Visual Representations. *arXiv preprint arXiv:2506.08008*, 2025. URL https://arxiv.org/abs/2506.08008.

Gemma Team. Gemma 3 Technical Report, 2025.

Google DeepMind. Gemini 2.0 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-0-Flash-Model-Card.pdf, 2025.

Google DeepMind. Gemini 3.1 Pro Model Card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024.

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 (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o System Card, 2024.

Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11): 1254–1259, 1998. doi: 10.1109/34.730558.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to Predict Where Humans Look. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 2106–2113. IEEE, 2009. doi: 10.1109/ICCV.2009.5459462.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, 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, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20.

Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The Secrets of Salient Object Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 280–287, 2014. doi: 10.1109/CVPR.2014.43.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models, 2023.

Benlin Liu, Amita Kamath, Madeleine Grunde-McLaughlin, Winson Han, and Ranjay Krishna. Visual Representations inside the Language Model. In *Conference on Language Modeling (COLM)*, 2025a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2023.

Ming Liu, Hao Chen, Jindong Wang, and Wensheng Zhang. On the Robustness of Multimodal Language Model Towards Distractions. *arXiv preprint arXiv:2502.09818*, 2025b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision (ECCV)*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022.

Meta AI. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`, 2024.

OpenAI. GPT-5 System Card. `https://cdn.openai.com/gpt-5-system-card.pdf`, August 2025.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, July 2026.

OpenBMB. MiniCPM-V 4.6 Model Card. `https://huggingface.co/openbmb/MiniCPM-V-4.6`, 2026.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, 2007. doi: 10.1214/07-AOAS122.

Zhiyu Pan, Yizheng Wu, Jiashen Hua, Junyi Feng, Shaotian Yan, Bing Deng, Zhiguo Cao, and Jieping Ye. Through the Lens of Contrast: Self-Improving Visual Reasoning in VLMs. In *International Conference on Learning Representations (ICLR)*, 2026.

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, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437.

J. Ridley Stroop. Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology*, 18(6):643–662, 1935. doi: 10.1037/h0054651.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky.

Resolution-Robust Large Mask Inpainting with Fourier Convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, 2022.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12388–12401, 2020.

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.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837, 2022.

Jeremy M. Wolfe. Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994. doi: 10.3758/BF03200774.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding, 2024.

Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting Human Gaze Beyond Pixels. *Journal of Vision*, 14(1):28, 2014. doi: 10.1167/14.1.28.

Jinhui Yang, Ming Jiang, and Qi Zhao. Defying Distractions in Multimodal Tasks: A Novel Benchmark for Large Vision-Language Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6314–6331, 2026. doi: 10.1109/TPAMI.2026.3655641.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, 2023.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations (ICLR)*, 2024.

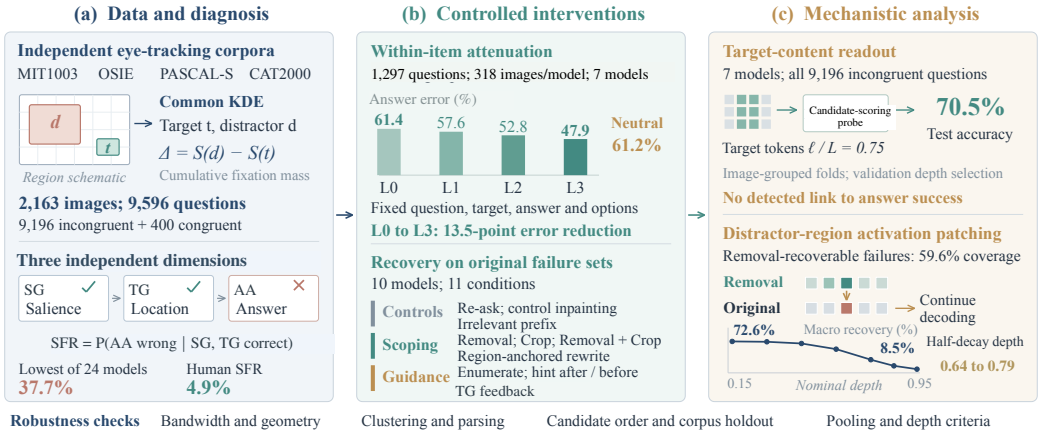


Figure 6: **Experimental and analytical overview of DECOY.** (a) Eye-tracking maps and diagnosis across three dimensions define conditional suppression failures (Appendices A to F). (b) Within-item attenuation and recovery interventions test distractor competition and reading scope (G to H). (c) Content probes measure decodability; activation patching measures causal efficacy across decoder depth (I to J). Patching covers removal-recoverable failures. Its curve averages seven models on a nominal depth grid; half-decay depths are computed per model.

A BENCHMARK COMPOSITION AND QUESTION CONSTRUCTION

A.1 IMAGES, QUESTIONS, AND SOURCE CORPORA

Figure 6 outlines the benchmark, diagnostic protocol, controlled interventions, and mechanistic analyses. DECOY distinguishes an image from a diagnostic instance. An image supplies the visual scene, its fixation map, and the salient object used in SG. An instance additionally specifies a question, the evidence needed to answer it, and its answer candidates. Several questions can therefore share a photograph without sharing a target or a correct answer. This distinction determines both the query budget and the unit at which repeated observations must be grouped in evaluation.

The benchmark contains 2,163 images and 9,596 questions. Of these questions, 9,196 are incongruent: a salient distractor competes with a distinct target needed for answering. The other 400 questions are congruent controls on images already included in the benchmark. They make the most salient object task-relevant, rather than adding a separate collection of easier photographs. Table 5 reports the image and question totals by primary task type; Table 6 gives incongruent counts by corpus; Figure 7 visualizes this composition and compares image and question weighting.

Table 5: Task composition of DECOY. Congruent questions use images already represented in the incongruent set; they add questions without increasing the image total. Each image has one primary task type and may support multiple questions about distinct task-relevant regions.

Task type	Images	Incongruent QA	Congruent QA
HOI (action target)	560	2,415	80
Safety (causal evidence)	430	2,031	80
Retrieval (identity / existence)	410	1,743	80
Attribute (attribute conflict)	330	1,263	80
Reasoning (compositional reasoning)	433	1,744	80
Total	2,163	9,196	400

The source images come from MIT1003 (Judd et al., 2009), OSIE (Xu et al., 2014), PASCAL-S (Li et al., 2014), and CAT2000 (Borji & Itti, 2015). For CAT2000, we restrict selection to six natural-photograph categories: Action, Indoor, Object, OutdoorManMade, OutdoorNatural, and Social. The restriction keeps the benchmark focused on objects in natural scenes rather than mixing photographs

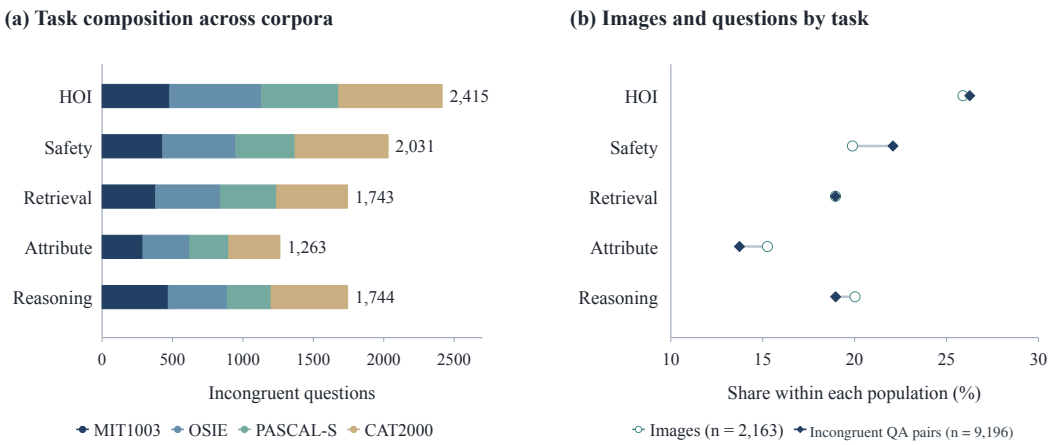


Figure 7: **Task coverage and population weighting.** (a) Incongruent questions by task and source corpus. (b) Task shares among 2,163 images and 9,196 incongruent questions, using separate denominators. The connecting segments compare these two shares within each task; they do not represent uncertainty. The 400 congruent controls use existing images and are excluded from both question summaries.

with synthetic patterns. Corpus identity is retained in the analysis because a common map estimator does not erase differences in image selection or gaze acquisition across datasets.

Table 6: Incongruent questions by task and source corpus. Counts sum to the task totals in Table 5. The additional congruent controls contribute 100 questions per corpus and are excluded from this cross-tabulation.

Task type	MIT1003	OSIE	PASCAL-S	CAT2000	Total
HOI	480	650	550	735	2415
Safety	430	520	420	661	2031
Retrieval	380	460	400	503	1743
Attribute	290	330	280	363	1263
Reasoning	470	420	310	544	1744
Total	2050	2380	1960	2806	9196

The five task labels are HOI (human-object interaction), Safety, Retrieval, Attribute, and Reasoning. Each image receives one primary task label. The labels describe the kind of evidence the target supplies: the participant or object in an interaction, a safety-relevant causal detail, an entity needed for retrieval, a distinguishing attribute, or several pieces of evidence needed jointly. They do not encode saliency levels. A reasoning question can have a small or large saliency gap, just as an attribute question can, so task and gap remain separate variables in the statistical analysis. The congruent controls contribute 80 questions per task and 100 per corpus.

A.2 Saliency-gap strata and the unit of description

The saliency gap compares the cumulative fixation mass of the distractor and the task-relevant target. For incongruent instance i, write $\Delta_i = S_i(r_d) - S_i(r_t)$, using the map definition in Appendix B. The image map is shared by questions on the same photograph; each question is written around one of its annotated targets. Consequently, gap values are attached to image-question instances; they should not be treated as independent image-level observations when several instances share a photograph.

The dose-response analysis uses global quintiles of the incongruent gap distribution. Global cut points give a common meaning to Q1 through Q5 across corpora: a high quintile represents a large gap in the pooled benchmark rather than merely a high rank within one source. Congruent controls form a separate group and do not enter the quintile construction. In particular, a congruent question is not a low-conflict incongruent question. It changes which evidence is relevant, whereas movement between incongruent quintiles changes the strength of competition between distinct regions.

There are two useful weightings of the benchmark. Question-weighted summaries describe the evaluated instance population and give more weight to images supporting more questions. Image-grouped uncertainty accounts for the dependence those questions share. Equal weighting of models serves a different purpose: it prevents a model with a larger eligible failure set from dominating an intervention average. Throughout the appendix, counts identify the relevant population, and "macro" refers to the averaging axis named in the surrounding text or caption.

A.3 Region geometry and its relation to saliency

A target is the region containing evidence sufficient for the particular question; a salient distractor is the competing object. For compositional questions, the target can comprise multiple annotated boxes. The boxes locate evidence and define the visual-token readout, but a box is not an object segmentation mask. It can contain background pixels, and its area need not coincide with the visible surface area of the object. This matters when comparing cumulative fixation mass across differently sized regions.

Let $a(r) = |r|/(WH)$ denote box area relative to an image of width W and height H. Region eccentricity measures the distance of its center from the image center relative to the image diagonal. Area and eccentricity describe different possible sources of association: a large box has more opportunity to collect fixation mass, while a centrally placed object can be prominent because of viewing geometry. Neither variable is interchangeable with the observed fixation distribution. Appendix F therefore reports them as sensitivity controls rather than silently redefining saliency to remove their contribution.

The distinction between cumulative mass and average density also changes the interpretation of a target. A small, intensely fixated patch can have high fixation density while receiving less cumulative fixation mass than a broad object. The benchmark's primary quantity measures total competing mass within the annotated regions. The area-normalized analysis asks a related question about density and provides a robustness check on that operational choice.

A.4 Human construction of questions and alternatives

All questions are written by humans. Construction starts from an image in which a dominant fixated region coexists with a distinct, visually resolvable region that can support a well-posed question. The task is then specified through evidence in that target, with the answer determined by the photograph. The salient object must be a plausible competitor for reading the scene, rather than an arbitrary object selected only because a model happens to fail. This separation keeps model responses out of the definition of the intended evidence.

The three dimensions impose different semantic constraints on their alternatives. SG asks which object is most visually salient. Its false candidates are ordinary non-salient objects, excluding the target regions of that image's questions. This keeps salient-object identification distinct from target grounding. TG presents the task question but asks only where to consult. Its correct option identifies a location; it must not include the attribute, count, identity, or relation that AA is meant to recover from that location.

TG incorrect locations are generated from non-target regions after excluding all annotated task-relevant targets. Selection prioritizes the most salient eligible distractor, the nearest non-overlapping region, and an area-matched non-target region. When a salient distractor is unavailable, the highest-saliency eligible non-target proposal is selected. Duplicate selections are replaced by the next-ranked eligible proposal; if fewer than three unique proposals remain, target-excluding background regions are added under the same spatial constraints. All options contain spatial locations without answer semantics.

Each AA query contains one correct answer and three false answers. One false answer is explicitly linked to the distractor: it is the answer suggested by consulting that competing object instead of the target. The remaining two candidates are plausible alternatives to the same question. This organization gives the errors a semantic interpretation beyond whether a letter was correct. It also requires distinguishing a distractor-linked answer from a generic incorrect option; a salient object elsewhere in the scene is insufficient unless consulting it supports the designated false response.

Four-way choice is held fixed across dimensions, so the nominal chance rate is 25%. The candidates within each dimension must answer the same kind of question: object descriptions for

salience identification, location descriptions for evidence localization, and answer content for the final task. TG's location-only constraint is especially important. If its correct alternative were to state both the evidence location and the answer, the eligibility criterion would no longer separate localization from reading. The subsequent language interventions use that location information under explicit conditions, described in Appendix H.

For a congruent control, the photograph remains fixed and the question is changed so that the most salient object supplies the required evidence. This produces a paired-image comparison, not a within-question manipulation. It controls the scene and image quality but permits differences in question content. The attenuation experiment in Appendix G supplies the complementary comparison in which the question, target, correct answer, and candidate set are all held fixed.

B Saliency Maps and Bandwidth Robustness

B.1 A common estimator from recorded fixations

We reconstruct saliency maps from the raw fixation coordinates supplied by the four corpora. Let $\mathcal{F}_I = \{f_j\}_{j=1}^{n_I}$ be the fixation locations for image I, expressed in that image's coordinate system, and let Ω_I be its pixel domain. A Gaussian kernel yields the unnormalized field $\widetilde{P}_I$, which is normalized over the image before region mass is computed:

$$\widetilde{P}_I(u) = \sum_{j=1}^{n_I} \exp\left(-\frac{\|u - f_j\|_2^2}{2\sigma_I^2}\right), \qquad P_I(u) = \frac{\widetilde{P}_I(u)}{\sum_{v\in\Omega_I} \widetilde{P}_I(v)}. \tag{2}$$

The common Gaussian normalization factor cancels in the second expression. The resulting map is a probability distribution over the image. Cumulative fixation mass, i.e., the fixation probability integrated over the box, is $S_I(r) = \sum_{u\in r\cap\Omega_I} P_I(u)$ and lies between zero and one. Region mass and the gap are consequently dimensionless. They are neither raw fixation counts nor saliency values expressed in an arbitrary image-intensity scale.

The main bandwidth is $\sigma_I = 0.025 D_I$, where $D_I = \sqrt{W_I^2 + H_I^2}$ is the image diagonal. It corresponds to 32 pixels for a 1024×768 image. Scaling by the diagonal avoids applying the same absolute pixel radius to images at different resolutions. The estimator smooths spatially discrete fixations into a field from which box-level mass can be computed, while preserving the distinction between the distractor and target annotations.

Using one estimator is preferable here to mixing distributed saliency maps that may have undergone different smoothing and normalization. It standardizes the transformation from recorded gaze to the quantity analyzed in the paper. It does not standardize the original viewing task, number of observers, or display conditions. Corpus fixed effects and the bandwidth analysis address different parts of this problem: the former allow baseline differences across sources, while the latter tests sensitivity to the smoothing scale. General distinctions between saliency maps, their normalization, and the metric applied to them are discussed by Bylinskii et al. (2019).

B.2 Sensitivity to the smoothing scale

Table 7 repeats the dose-response summaries at bandwidths $0.015D$, $0.025D$, and $0.040D$. The unconditional error gradient remains monotone under all three settings. At the primary bandwidth, error rises from 37.6% in Q1 to 74.1% in Q5. The corresponding endpoints are 37.1% and 74.8% for the narrower kernel, and 38.2% and 73.6% for the wider kernel. Thus, smoothing changes the endpoints modestly without eliminating the relation between conflict and error.

The regression coefficient has the same interpretation within each specification: the odds ratio for a one-standard-deviation increase in its gap measure, with corpus, task, and model effects included. The estimates decrease from 1.77 to 1.68 as the kernel widens, and their intervals overlap. Figure 8 juxtaposes the bandwidth-specific error gradients with the geometric-control estimates. The useful result is the persistence of a substantial positive association across a range of spatial smoothing scales.

Table 7: Bandwidth sensitivity of unconditional answer error (%). The $Q1$ to $Q5$ columns use saliency-gap quintiles, and Cong. denotes congruent questions. Odds ratios describe a one-standard-deviation increase in the corresponding gap, with image-clustered 95% intervals.

σ/D	Cong.	Q1	Q2	Q3	Q4	Q5	OR [95% CI]
0.015	10.6	37.1	46.2	55.4	65.1	74.8	1.77 [1.68, 1.88]
0.025 (main)	10.7	37.6	46.9	55.8	64.7	74.1	1.73 [1.64, 1.83]
0.040	10.9	38.2	47.4	55.6	64.1	73.6	1.68 [1.59, 1.78]

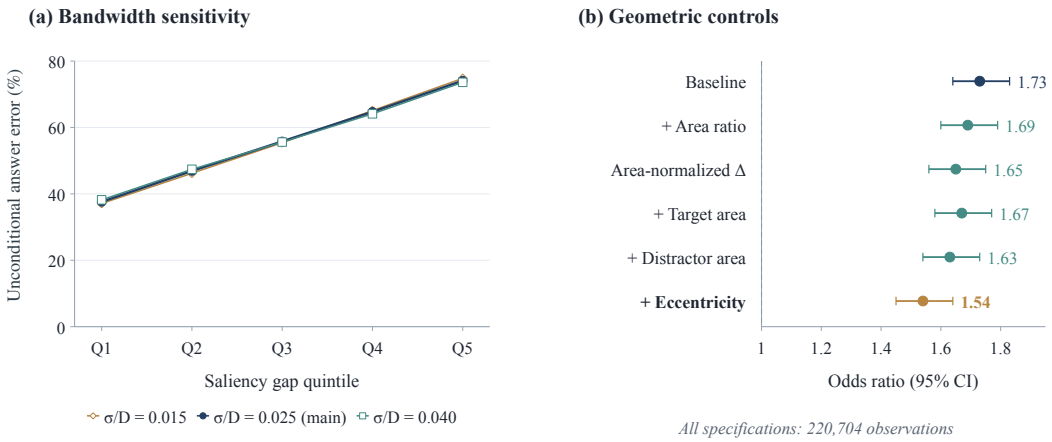


Figure 8: **Sensitivity to saliency estimation and region geometry.** (a) Unconditional answer error across gap quintiles for three KDE bandwidths, averaged equally over 24 models; D denotes image diagonal. (b) Standardized-gap odds ratios with image-clustered 95% intervals and task, corpus, and model fixed effects. All specifications use 220,704 observations. Geometric adjustments attenuate the association without eliminating it.

B.3 Cumulative fixation mass, density, and area controls

A cumulative-fixation-mass measure necessarily depends in part on the size of the region integrated. If two boxes have the same average fixation density, the larger one receives more mass. This is a feature of the chosen competition measure, but it creates a possible alternative explanation when target and distractor geometry differ systematically. A positive relation between Δ and error could then partly reflect small targets being hard to read, rather than prominence alone.

We examine this possibility by adding the area ratio, target area, or distractor area as covariates, and by replacing the mass-based gap with its area-normalized counterpart. The complete estimates appear in Table 15. None removes the association, although each changes its magnitude. Keeping these specifications visible makes the measurement assumption inspectable: the primary result concerns cumulative attention to the annotated regions, and the density-based result tests whether the conclusion depends entirely on collecting mass over a larger box.

C Annotation and Human Evaluation

C.1 Independent annotation and adjudication

Three annotators label each image independently. The annotations include its primary task type, salient distractor, task-relevant target, correct answer, and distractor-linked false answer. Object-level annotations from OSIE and PASCAL-S provide candidate regions where available; they do not themselves determine which object is relevant to a newly written question. Relevance is defined with respect to that question, while saliency is grounded in the recorded fixation data.

Disputed cases are reviewed by two additional adjudicators, with the final decision set by majority across all five annotators. This separates an initial independent assessment from the resolution of disagreements. Table 8 reports the agreement measurements for categorical labels and boxes.

Reporting both is necessary because annotators can agree on an object's identity while drawing different boundaries, or agree on a region while disagreeing about which answer it supports.

Table 8: Agreement before adjudication, separated into categorical decisions and spatial annotations. Fleiss' κ measures agreement beyond chance; box overlap uses pairwise intersection over union (IoU). The final two rows report percentages of box comparisons reaching IoU 0.5.

Measure	Annotation	Value
Fleiss' κ	Task type	0.84
Fleiss' κ	Correct answer	0.89
Fleiss' κ	Distractor-linked false option	0.82
Mean pairwise IoU	Distractor box	0.81
Mean pairwise IoU	Target box	0.76
IoU ≥ 0.5 (%)	Distractor box	92.3
IoU ≥ 0.5 (%)	Target box	88.7

Fleiss' κ (Fleiss, 1971) reaches 0.84 for task type, 0.89 for the correct answer, and 0.82 for the distractor-linked option. The last quantity directly concerns the interpretation of errors: it measures whether annotators agree that a particular false answer follows from the competing region. Mean pairwise box IoU is 0.81 for distractors and 0.76 for targets. The corresponding proportions reaching IoU 0.5 are 92.3% and 88.7%, respectively.

Target boxes have lower overlap agreement than distractor boxes, which is relevant when they define cropped inputs or pooled visual features. A compositional target can also require several regions. We therefore preserve the task-specific evidence annotations in the probing and intervention procedures instead of replacing them with a single scene-level notion of the main object.

C.2 Human task and conditional scoring

The human study includes 26 participants and 350 stratified instances, with three participant responses per instance. Of the items, 292 are incongruent and 58 are congruent controls. Participants use the same multiple-choice protocol across three dimensions as the models: Saliency Grounding (SG), Target Grounding (TG), and Answer Accuracy (AA). The comparison keeps the diagnostic decisions aligned; it does not require human gaze or computation to follow a model's internal processing path.

Scoring is defined on participant-item responses. A response is eligible when the same participant answers both SG and TG correctly for that item. A suppression failure is an eligible response with an incorrect AA answer. Responses are not first collapsed into a majority-vote label, because such aggregation can combine correct grounding responses from different people and thereby change the meaning of conditional failure. Congruent responses remain separate from the incongruent SFR calculation.

Aggregate SG, TG, and AA are 97.0%, 95.2%, and 92.4%, with eligibility 94.3%. On incongruent items, human SFR is 4.9%, with a 95% interval of $[3.0, 7.5]$. High performance supports the intended answerability of the benchmark, while the nonzero error rate makes clear that the protocol is not trivial for every individual observer.

C.3 Uncertainty and performance across quintiles

Human responses share both an image and a participant. For uncertainty, the response-level aggregation uses $10{,}000$ bootstrap replicates with image and participant identifiers resampled independently, carrying their multiplicities into the response weights. Eligibility and failure counts are recomputed together within a replicate before taking their ratio. This crossed resampling accounts for repeated observations along both axes; its general motivation follows the pigeonhole bootstrap of Owen (2007).

Table 9 separates the item counts, human point estimates, and intervals. Each incongruent quintile contains only 58 or 59 items. Figure 9 shows the narrow range of human point estimates (4.1 to 5.8%) alongside their wide, substantially overlapping intervals. The data support a low human failure rate throughout the sampled range; a modest increase at high conflict remains possible.

Table 9: Human conditional error across the saliency-gap range. The human study uses 350 items with three responses each; quintile counts refer to items. SFR uses only incongruent questions. The model column gives the corresponding conditional benchmark summary, rather than unconditional answer error.

Stratum	Items	Human (%)	Human 95% CI	Model macro (%)
Congruent	58	3.7	[0.5, 12.5]	9.1
Q1	58	4.1	[0.8, 13.1]	39.6
Q2	59	4.7	[1.0, 13.8]	46.2
Q3	58	5.2	[1.2, 14.5]	53.1
Q4	59	4.9	[1.1, 14.2]	61.2
Q5	58	5.8	[1.5, 15.3]	68.4
Incongruent	292	4.9	[3.0, 7.5]	53.7

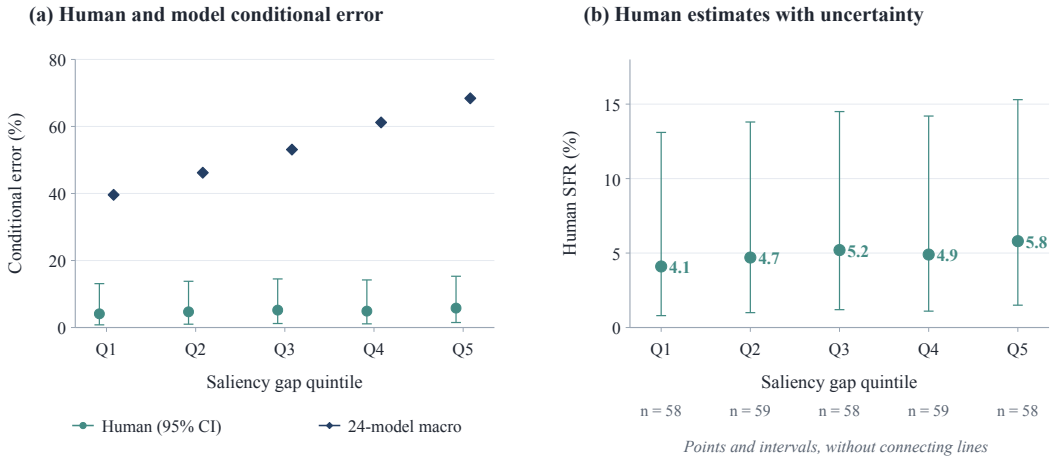


Figure 9: **Human suppression failures across saliency-gap quintiles.** (a) Human SFR with reported 95% intervals alongside the 24-model macro estimates. (b) The same human observations on an expanded vertical scale, with item counts shown below. The 292 incongruent items receive three responses each. Points are unconnected to preserve the uncertainty of the human estimates; model intervals are not reported.

The model column reports conditional error after successful salience identification and localization. It therefore differs from the unconditional answer error rates in Table 7. Conditioning changes which instances contribute for each model, and a model with poor SG or TG performance can have a smaller eligible set. Both summaries are useful, but their numerical levels answer different questions. The human comparison uses the conditional diagnostic definition, while the principal saliency-gap regression uses unconditional answering errors.

D Evaluation Protocol and Model Coverage

D.1 Dimension inputs and query accounting

Each of SG, TG, and AA uses an independent query with an image, a dimension-specific question, four labeled alternatives, and the instruction *"Answer with only one letter: A, B, C, or D."* SG asks which object is most visually salient in the image. TG supplies the task question and asks where the evidence needed to answer it is located. AA presents the task question itself. The alternatives are filled from the annotations described in Appendix A; they are not a shared answer vocabulary across all questions.

The common input order is image, question, candidates A through D, and the answer-format instruction. The three baseline queries do not share conversation history or responses. Image placement denotes the model's visual input at that point in its multimodal prompt, not a literal textual placeholder sent in place of the image. TG alternatives specify location without revealing answer content. In the baseline evaluation, a correct TG response is used to define eligibility; explicitly feeding its location back into AA is a separate intervention, not part of the baseline scoring rule.

SG is queried once per image, and its correctness is shared by that image's associated questions. TG and AA are each queried once per question, including the congruent controls. The total is consequently $2{,}163 + 2(9{,}596) = 21{,}355$ queries per model. Counting a separate SG call for every question would overstate the evaluation budget and, more importantly, conceal the shared SG response when estimating uncertainty.

Decoding uses temperature 0, top-p 1, one response, and at most 16 new tokens. These settings remove intentional sampling over answers; they do not imply that a hosted implementation can never vary across requests or backend revisions. Results describe the evaluated model configurations under the fixed protocol. A thinking variant is retained as its own evaluated system rather than pooled with the non-thinking model of the same nominal size.

The generation-budget sensitivity evaluation repeats Enumerate and Thinking runs with 64 new tokens while retaining temperature 0 and top-p 1. For Enumerate, truncation decreases from 2.8% to 0.3%, and parsing failure decreases from 1.6% to 0.8%. For Thinking models, the corresponding changes are 8.7% to 1.1% and 3.3% to 1.2%. Changes in the primary evaluation metrics remain below 0.7 percentage points.

The text-only GPT-5.6 Sol baseline (OpenAI, 2026) answers $2{,}507$ of $9{,}196$ incongruent four-choice questions correctly, yielding 27.3% accuracy against 25% chance. Language priors therefore provide limited but non-zero predictive information.

D.2 Letter parsing and failure handling

After normalization, the parser extracts a standalone candidate letter from A through D. Accepted forms include a bare letter, a letter followed by punctuation, a parenthesized letter, and a letter following `Answer:`. If multiple candidate letters appear, an explicit answer following `Answer:` takes precedence; otherwise the first standalone candidate letter is used. The parser does not resolve an answer by judging the semantic plausibility of a free-form explanation, since that would introduce an additional model or adjudication step into scoring.

An output that cannot be resolved by this rule is scored as incorrect. The same rule applies in all three dimensions, but its effect depends on the dimension. An SG or TG parsing failure removes the response from the eligible set. An AA parsing failure contributes to suppression failure if both SG and TG responses were correct. Thus, parse failures can affect both the numerator and denominator of SFR, whereas unconditional AA is affected directly through its error count.

Table 10 reports parse-failure rates by model and dimension; the largest AA parsing-failure rate is 4.9% for mplug-owl-7B. Rates are lower for the stronger evaluated systems, reaching 0.2% for GPT-5 and Gemini-3.1-Pro. This asymmetry qualifies comparisons at the weak end of the model range.

D.3 Model families and mechanistic access

The evaluation covers 24 systems: 19 open-weight models and five hosted models. The open-weight set spans InternLM-XComposer2, VILA, Llama 3.2 Vision, mplug-owl, MiniCPM-V, Gemma 3, Phi-4-Multimodal, DeepSeek-VL2, LLaVA, Qwen3-VL, InternVL3.5, Molmo2, and MolmoPoint. The hosted set comprises GPT-4o, GPT-5, Gemini-2.0-flash, Gemini-3.1-Pro, and Claude-Opus-4.8. GPT-5.6 Sol is evaluated separately as a text-only baseline and is not included in these 24 models. Exact evaluated names are retained in the full diagnostic and parse-rate tables.

Mechanistic analyses use seven open-weight systems for which the required hidden states and interventions are available: mplug-owl-7B, Qwen3-VL-2B, LLaVA-v1.5-7B, Qwen3-VL-8B, its Thinking variant, InternVL3.5-8B, and MolmoPoint-8B. This subset contains decoders of 28, 32, and 36 layers. Appendix J gives their complete layer audit. These are decoder lengths, not counts of vision-encoder blocks, and a common nominal relative depth must be mapped separately for each length.

Behavioral, probing, and intervention summaries use different model populations and eligible sets.

D.4 Recorded checkpoint and endpoint identifiers

Tables 11 and 12 record the evaluated open-weight checkpoints and hosted endpoints, respectively. Hosted names can resolve to provider-managed implementations, and repository paths can change

Table 10: Unparseable responses (%) under the fixed letter parser. Each dimension has its own denominator: SG is evaluated once per image, whereas TG and AA are evaluated per question. Overall uses query-weighted aggregation across all three dimensions of the diagnostic evaluation.

Model	SG	TG	AA	Overall
InternLM-XComposer2-7B	1.3	1.9	2.7	2.2
VILA-8B	0.9	1.4	2.1	1.7
Llama3.2-Vision-11B	0.4	0.7	1.1	0.8
mplug-owl-7B	2.7	3.8	4.9	4.2
MiniCPM-V-4.6	0.3	0.5	0.7	0.6
Gemma3-4B	0.8	1.2	1.7	1.4
Phi-4-Multimodal	0.4	0.6	0.9	0.7
DeepSeek-VL2-Tiny	2.1	3.2	4.4	3.6
LLaVA-v1.5-7B	1.9	2.9	4.1	3.3
LLaVA-OneVision-2-8B	0.5	0.8	1.2	0.9
Qwen3-VL-2B	0.8	1.3	1.9	1.5
Qwen3-VL-4B	0.6	1.0	1.4	1.1
Qwen3-VL-8B	0.3	0.5	0.8	0.6
Qwen3-VL-8B-Thinking	0.2	0.4	0.6	0.5
InternVL3.5-2B	0.8	1.1	1.6	1.3
InternVL3.5-4B	0.5	0.8	1.2	0.9
InternVL3.5-8B	0.3	0.5	0.8	0.6
Molmo2-8B	0.2	0.4	0.6	0.5
MolmoPoint-8B	0.1	0.3	0.4	0.3
GPT-4o	0.1	0.2	0.2	0.2
GPT-5	0.1	0.1	0.2	0.1
Gemini-2.0-flash	0.1	0.2	0.3	0.2
Gemini-3.1-Pro	0.1	0.1	0.2	0.1
Claude-Opus-4.8	0.1	0.2	0.3	0.2

independently of a paper's evaluation. The listed dates therefore delimit the recorded runs without implying that the same alias is an immutable snapshot. The local LLaVA-OneVision-2-8B identifier denotes the locally named evaluated configuration.

For the seven systems used in internal experiments, Table 28 additionally records the decoder length and absolute layer indices used by our analysis. These audited values define the actual probe and patching operations. Parameter counts embedded in model names are not substituted for decoder lengths, and a hosted model's internal depth is not inferred from its API label.

E Full Diagnostic Results

E.1 Task-level SG, TG, and AA

Tables 13 and 14 provide the complete task decomposition underlying the compressed main table. The split keeps the model names and dimension columns readable while retaining every task-level value. Each task reports SG, TG, and AA. The second panel also retains task-macro AA, eligibility, and SFR. These last two quantities use only incongruent questions, as in the main diagnostic definition.

Across the table, high accuracy in locating evidence does not translate into comparable accuracy in answering from that evidence. This observation motivates the conditional analysis, but marginal SG, TG, and AA alone do not identify the failed instances. Eligibility is the joint event that SG and TG responses are correct for a given model and question, not the product of the two marginal accuracies. SFR is then measured within that joint set. Multiplying marginal accuracies would assume independence that the protocol neither requires nor establishes.

The five task types also have unequal question counts. The main table's Avg column gives equal weight to task types, while an instance-weighted AA rate weights each question equally. These summaries can differ even when computed from the same responses. The 39.9% AA used in the probe

Table 11: Open-weight evaluation records. Identifiers and dates record the evaluated checkpoints or endpoints and evaluation days.

Model	Recorded identifier	August 2026
InternLM-XComposer2-7B	internlm/internlm-xcomposer2-vl-7b	18
VILA-8B	Efficient-Large-Model/VILA1.5-8B	18
Llama3.2-Vision-11B	meta-llama/Llama-3.2-11B-Vision-Instruct	19
mplug-owl-7B	MAGAer13/mplug-owl-llama-7b	19
MiniCPM-V-4.6	openbmb/MiniCPM-V-4_6	20
Gemma3-4B	google/gemma-3-4b-it	20
Phi-4-Multimodal	microsoft/Phi-4-multimodal-instruct	20
DeepSeek-VL2-Tiny	deepseek-ai/deepseek-vl2-tiny	21
LLaVA-v1.5-7B	liuhaotian/llava-v1.5-7b	21
LLaVA-OneVision-2-8B	local: LLaVA-OneVision-2-8B	21
Qwen3-VL-2B	Qwen/Qwen3-VL-2B-Instruct	22
Qwen3-VL-4B	Qwen/Qwen3-VL-4B-Instruct	22
Qwen3-VL-8B	Qwen/Qwen3-VL-8B-Instruct	22
Qwen3-VL-8B-Thinking	Qwen/Qwen3-VL-8B-Thinking	22
InternVL3.5-2B	OpenGVLab/InternVL3_5-2B	23
InternVL3.5-4B	OpenGVLab/InternVL3_5-4B	23
InternVL3.5-8B	OpenGVLab/InternVL3_5-8B	23
Molmo2-8B	allenai/Molmo2-8B	24
MolmoPoint-8B	allenai/MolmoPoint-8B	24

Table 12: Hosted evaluation records. GPT-5.6 Sol is the text-only baseline, evaluated on 9,196 incongruent questions, and is excluded from the 24-model benchmark set. Identifiers and dates record the evaluated checkpoints or endpoints and evaluation days.

Model	Recorded identifier	August 2026
GPT-4o	openai/gpt-4o	25
GPT-5	openai/gpt-5	25
Gemini-2.0-flash	google/gemini-2.0-flash	26
Gemini-3.1-Pro	google/gemini-3.1-pro	26
Claude-Opus-4.8	anthropic/claude-opus-4.8	26
GPT-5.6 Sol (text-only)	openai/gpt-5.6-sol	27

cross-tabulation has a further restriction: it includes only the 9,196 incongruent questions for each of the seven probe models. It should not be replaced by an average that also includes congruent controls.

E.2 Eligibility and conditional failure

For model m, let $a_{mi}^{(d)}$ indicate correctness in dimension $d \in \{\mathrm{SG}, \mathrm{TG}, \mathrm{AA}\}$ for incongruent instance i. Define $e_{mi} = a_{mi}^{(\mathrm{SG})} a_{mi}^{(\mathrm{TG})}$ and $f_{mi} = e_{mi}(1 - a_{mi}^{(\mathrm{AA})})$. Eligibility and suppression failure are then

$$\mathrm{Elig}_m = \frac{\sum_i e_{mi}}{9{,}196}, \qquad \mathrm{SFR}_m = \frac{\sum_i f_{mi}}{\sum_i e_{mi}}. \tag{3}$$

These ratios separate grounding correctness from answer correctness conditional on successful grounding. A lower SFR is therefore not, by itself, a complete ranking of model capability. It must be read alongside eligibility and final accuracy, since the eligible questions are model-dependent.

The intersection of eligible questions across all 24 evaluated models contains 1,327 questions. Relative to model-specific eligible sets, the mean absolute SFR change is 1.4 percentage points, with a maximum of 2.6 points. The two estimates have Spearman correlation $\rho = 0.94$, preserving the qualitative diagnostic conclusion on a shared question set.

Table 13: Full task-level diagnostic accuracy (%), HOI, Safety, and Retrieval. SG, TG, and AA denote Saliency Grounding, Target Grounding, and Answer Accuracy. SG, TG, and AA include congruent controls; conditional diagnostics use only incongruent questions.

Model	HOI			Safety			Retrieval		
	SG	TG	AA	SG	TG	AA	SG	TG	AA
InternLM-XComposer2-7B	75.2	78.4	43.8	76.1	79.2	47.6	75.5	78.8	45.5
VILA-8B	76.8	79.8	45.1	77.6	80.7	48.9	77.0	80.2	46.7
Llama3.2-Vision-11B	80.0	83.0	46.4	80.8	84.0	50.1	80.2	83.5	47.8
mplug-owl-7B	68.2	71.0	34.2	69.0	72.2	38.0	68.5	71.5	36.1
MiniCPM-V-4.6	82.0	84.8	49.0	82.8	85.7	52.8	82.2	85.2	50.4
Gemma3-4B	74.8	77.7	39.8	75.6	78.8	43.7	75.0	78.2	41.5
Phi-4-Multimodal	81.4	84.3	48.3	82.2	85.4	51.8	81.7	84.8	49.5
DeepSeek-VL2-Tiny	68.8	71.8	35.4	69.6	73.0	39.1	69.0	72.3	37.2
LLaVA-v1.5-7B	65.1	68.7	36.8	65.8	70.2	41.9	64.8	69.1	39.5
LLaVA-OneVision-2-8B	79.2	82.2	45.7	80.0	83.2	49.6	79.4	82.7	47.3
Qwen3-VL-2B	70.2	73.1	34.8	71.0	74.4	38.5	70.5	73.7	36.1
Qwen3-VL-4B	73.8	76.8	37.2	74.6	78.2	41.3	74.0	77.6	39.0
Qwen3-VL-8B	77.6	81.3	39.6	78.4	82.7	44.2	77.8	82.0	41.5
Qwen3-VL-8B-Thinking	78.4	82.0	41.5	79.2	83.3	46.0	78.6	82.6	43.4
InternVL3.5-2B	72.8	75.8	40.4	73.6	77.2	44.7	73.0	76.5	42.5
InternVL3.5-4B	76.2	79.6	43.2	77.0	81.0	47.5	76.5	80.3	45.3
InternVL3.5-8B	79.3	82.8	45.8	80.1	84.0	50.1	79.6	83.5	47.9
Molmo2-8B	80.2	83.4	48.2	81.0	84.7	52.5	80.5	84.0	50.0
MolmoPoint-8B	82.8	86.0	51.0	83.6	87.2	55.4	83.0	86.6	53.0
GPT-4o	83.0	86.2	52.8	83.8	87.3	57.0	83.2	86.6	54.5
GPT-5	84.6	87.7	56.8	85.0	88.5	61.0	84.7	88.0	58.2
Gemini-2.0-flash	82.3	85.3	49.6	83.0	86.6	54.0	82.5	85.9	51.4
Gemini-3.1-Pro	84.8	88.0	57.4	85.0	88.8	62.0	84.9	88.2	59.0
Claude-Opus-4.8	84.0	87.0	54.6	84.6	88.1	59.1	84.2	87.5	56.3

Table 14: Full task-level diagnostic accuracy (%), Attribute and Reasoning, with aggregate diagnostics. SG, TG, and AA denote Saliency Grounding, Target Grounding, and Answer Accuracy. SG, TG, and AA include congruent controls; conditional eligibility and SFR exclude them.

Model	Attribute			Reasoning			Avg	Elig.	SFR
	SG	TG	AA	SG	TG	AA			
InternLM-XComposer2-7B	75.0	77.7	43.4	74.8	77.5	42.7	44.6	59.4	56.8
VILA-8B	76.5	79.2	44.3	76.3	79.0	43.8	45.8	61.2	55.3
Llama3.2-Vision-11B	79.7	82.7	45.6	79.6	82.5	45.1	47.0	66.1	53.5
mplug-owl-7B	67.9	70.6	34.5	67.8	70.4	33.8	35.3	48.3	64.0
MiniCPM-V-4.6	81.7	84.3	48.2	81.6	84.1	47.7	49.6	69.2	51.2
Gemma3-4B	74.4	77.3	40.2	74.2	77.1	39.6	41.0	58.1	58.6
Phi-4-Multimodal	81.1	83.9	47.5	80.9	83.7	46.9	48.8	68.4	52.0
DeepSeek-VL2-Tiny	68.5	71.3	35.8	68.3	71.1	35.0	36.5	49.1	62.7
LLaVA-v1.5-7B	65.3	68.5	37.5	64.5	68.0	36.8	38.5	45.6	60.9
LLaVA-OneVision-2-8B	78.8	81.8	45.1	78.6	81.6	44.5	46.4	64.8	54.4
Qwen3-VL-2B	69.9	72.7	35.2	69.7	72.5	34.5	35.8	51.2	62.5
Qwen3-VL-4B	73.5	76.5	37.8	73.3	76.3	37.0	38.5	56.9	59.8
Qwen3-VL-8B	77.2	80.8	40.3	77.0	80.6	39.5	41.0	62.7	56.8
Qwen3-VL-8B-Thinking	78.0	81.5	42.1	77.8	81.3	41.4	42.9	64.1	54.9
InternVL3.5-2B	72.5	75.5	41.3	72.3	75.3	40.6	41.9	55.7	57.2
InternVL3.5-4B	75.9	79.2	43.8	75.7	79.0	43.0	44.6	61.0	54.8
InternVL3.5-8B	79.0	82.4	46.5	78.8	82.2	45.7	47.2	65.6	52.0
Molmo2-8B	79.8	83.0	48.4	79.6	82.8	47.6	49.3	67.2	51.4
MolmoPoint-8B	82.4	85.5	51.2	82.2	85.3	50.4	52.2	71.1	48.2
GPT-4o	82.7	85.8	52.9	82.5	85.6	52.1	53.9	72.0	47.0
GPT-5	84.3	87.3	56.5	84.1	87.1	55.8	57.7	74.3	37.7
Gemini-2.0-flash	82.0	85.0	49.8	81.8	84.8	48.9	50.7	70.2	49.8
Gemini-3.1-Pro	84.5	87.6	57.6	84.3	87.4	56.7	58.5	75.0	41.6
Claude-Opus-4.8	83.8	86.8	54.8	83.6	86.6	54.0	55.8	73.1	44.6

E.3 CORRECT-OPTION POSITION

AA is 46.1%, 45.7%, 46.3%, and 45.8% when the correct answer occupies A, B, C, and D, respectively. The maximum spread is 0.6 percentage points. The image-clustered, model-controlled

comparison yields $\chi^2(3) = 1.42$ and $p = .70$. These results provide no evidence of a large aggregate accuracy advantage for one answer position under the evaluated presentation scheme.

The candidate-reordering control in Appendix I separately tests the content probe on reordered versions of the same alternatives.

E.4 Congruent questions on matched images

The paired-image comparison uses 400 congruent questions and the corresponding incongruent questions on the same photographs. Congruent answer error is 10.7%, compared with 54.8% for the matched incongruent questions. The paired difference is 44.1 percentage points, with a 95% paired-bootstrap interval of $[39.6, 48.5]$ and $p < .001$. Pairing keeps the photograph fixed, so the large difference cannot be attributed simply to presenting an entirely different collection of images.

The question is nevertheless changed to make the most salient object relevant. The comparison therefore does not isolate a saliency manipulation while holding semantics constant. Its role is to show that the same scenes can support much better performance when relevance aligns with prominence. The attenuation ladder adds a stricter within-item test by holding the question and answer fixed while changing only the competing region.

F Alternative Explanations and Statistical Sensitivity

F.1 Geometric specifications

The primary regression models unconditional answer error on all $24 \times 9{,}196 = 220{,}704$ incongruent model-question observations. It includes fixed effects for corpus, task type, and model. In schematic form, for error indicator Y_{mi} and standardized gap z_i,

$$\operatorname{logit} \Pr(Y_{mi} = 1) = \alpha + \beta z_i + \gamma_{c(i)} + \tau_{t(i)} + \eta_m. \tag{4}$$

The reported odds ratio is $\exp(\beta)$. Image-clustered uncertainty allows questions and model responses sharing an image to be dependent. The baseline in Table 15 is uncontrolled for geometry, but it already contains these corpus, task, and model effects.

Table 15: Geometric sensitivity of the saliency-gap association. All specifications include corpus, task, and model fixed effects and use image-clustered uncertainty. Each row contains 220,704 model-question observations; odds ratios and intervals refer to the standardized gap in that specification.

Specification	OR	95% CI	p	Pseudo-R^2
Uncontrolled	1.73	[1.64, 1.83]	$< .001$	0.172
Plus area-ratio covariate	1.69	[1.60, 1.79]	$< .001$	0.176
Area-normalized Δ	1.65	[1.56, 1.75]	$< .001$	0.169
Plus target-area covariate	1.67	[1.58, 1.77]	$< .001$	0.177
Plus distractor-area covariate	1.63	[1.54, 1.73]	$< .001$	0.178
Plus eccentricity covariate	1.54	[1.45, 1.64]	$< .001$	0.184

Adding the area-ratio covariate changes the gap odds ratio from 1.73 to 1.69. Separate target-area and distractor-area controls give 1.67 and 1.63, respectively. Replacing cumulative mass by the area-normalized gap yields 1.65. All reported intervals exclude one. Thus, the association is not eliminated by these geometric adjustments, although the reduced coefficients show that geometry contributes to its magnitude.

The cumulative-mass model describes competition as operationalized in the benchmark; the area-normalized model changes that predictor. Covariate-adjusted models instead ask whether the mass-based gap remains informative conditional on additional geometry. Their agreement in direction is useful robustness evidence, but it does not identify a unique causal decomposition of saliency, object size, and image difficulty.

A joint specification over $220{,}704$ model-question pairs controls simultaneously for target area, distractor area, distance between target and distractor, object count, corpus, task, and model fixed

effects. The saliency-gap association remains positive ($\beta = 0.587$, SE $= 0.041$, OR $= 1.80$). Restricting the analysis to eligible questions yields $\beta = 0.521$, SE $= 0.047$, and OR $= 1.68$; excluding parsing failures yields $\beta = 0.569$ and SE $= 0.042$. The positive association persists under joint adjustment and both sample restrictions.

F.2 Dependence across images and models

Table 16 holds the fitted regression fixed and varies its uncertainty calculation. Clustering by image gives interval $[1.64, 1.83]$, whereas clustering by model gives $[1.43, 2.09]$. Two-way clustering gives $[1.42, 2.10]$, and an image-clustered bootstrap gives $[1.63, 1.84]$. The wider model-based intervals show why the number of model-question rows should not be mistaken for the number of independent units supporting inference.

Table 16: Sensitivity to the dependence structure used for inference. The fitted mean model is held fixed, so the odds ratio is unchanged across rows. The bootstrap resamples image clusters 10,000 times; the two-way estimate allows dependence within both images and models.

Uncertainty estimate	OR	95% CI	Clusters
Clustered by image (main)	1.73	[1.64, 1.83]	2163
Clustered by model	1.73	[1.43, 2.09]	24
Two-way, image $\times$ model	1.73	[1.42, 2.10]	2163 / 24
Image-clustered bootstrap	1.73	[1.63, 1.84]	2163 ($B = 10000$)

There are 2,163 image clusters but only 24 model clusters. Repeated predictions from the same model can share systematic behavior that is not captured by treating every row as independent. Likewise, multiple questions and models evaluated on one image can share a difficult scene feature. Cluster-robust methods address such dependence in uncertainty estimates; they do not remove omitted-variable bias in the fitted mean relationship (Cameron & Miller, 2015). The persistence of intervals above one across the reported choices supports the association under several dependence assumptions.

F.3 Why eccentricity warrants separate attention

Eccentricity produces the largest attenuation among the geometric controls: the odds ratio falls from 1.73 to 1.54, with interval $[1.45, 1.64]$. On the log-odds scale, the reduction is $(\log 1.73 - \log 1.54)/\log 1.73 \approx 21\%$. The adjusted association remains positive, but its magnitude is smaller once this aspect of spatial arrangement is included.

The roughly one-fifth attenuation is a descriptive comparison of specifications and does not distinguish mediation through gaze from confounding by scene layout.

The primary specification retains the gaze-defined gap because the benchmark concerns the total visual competition present in the scene, including contributions associated with spatial layout. The eccentricity-adjusted model asks how much association remains after conditioning on that layout variable. Reporting both makes the scope explicit. Neither observational specification replaces the within-item experiment, where the target and question are fixed and the distractor is directly attenuated.

G Within-Item Distractor Attenuation

G.1 The intervention and its fixed quantities

The attenuation experiment uses 1,297 questions on 318 images per model and changes the visual prominence of the distractor within each item. The image identity, downstream question, correct answer, candidate set, and task-relevant target are held fixed. L0 is the original image, while L1, L2, and L3 apply progressively stronger local attenuation. Unlike Removal, the operator retains the distractor object: the intended contrast concerns its prominence in the same scene, rather than the presence or absence of a competing object.

The operator acts on local chroma, luminance contrast, and spatial detail, with level-specific settings in Table 17. In CIELAB coordinates, chromatic components are contracted toward neutral, while

luminance contrast is reduced around a local mean. Additional smoothing attenuates fine structure. The edited region is blended back through a soft boundary. These operations preserve the local mean-luminance reference while reducing several sources of prominence together; the experiment therefore tests a composite attenuation, not the separate causal effects of color, contrast, and sharpness.

Table 17: Parameters of the attenuation ladder. Chroma and local-contrast scales decrease with level, while the smoothing scale increases. The final column is the implementation's smoothing-mixture coefficient. Pixel scales are specified at a 1024-pixel image long side and scale with image resolution.

Level	Chroma scale	Contrast scale	Blur σ (px)	Mixture
L0	1.00	1.00	0.0	1.00
L1	0.82	0.88	0.8	0.92
L2	0.64	0.72	1.6	0.80
L3	0.48	0.58	2.4	0.68

The distractor box is expanded by 6% of its width and height on each side, with a soft transition specified at 12 pixels for a 1024-pixel long side. The target is held unchanged. The experimental interpretation rests on retaining the object and validating a graded change in attention, rather than assuming that any blur or desaturation operation automatically isolates saliency.

The neutral condition applies the L3 operator to a task-irrelevant region. Its area is matched within 10% of the distractor region and its eccentricity within 0.05 image diagonals. It therefore controls for a comparable local edit without directly attenuating the competing evidence. This control tests whether a generic editing operation accounts for the behavioral change.

G.2 Manipulation check

The manipulation check uses 24 raters and the same 318 base images. Conditions are assigned in a balanced incomplete-block design so that a participant sees a base image at most once, with conditions rotated across participants. The image is displayed for $1{,}200$ milliseconds. A forced choice then asks which of the two regions first drew attention. Fresh gaze measurements are processed with the benchmark's KDE bandwidth of 0.025 image diagonals to obtain the gap for the edited conditions.

Table 18: Manipulation check on 318 images with 24 raters. Preference is the percentage selecting the distractor as the region that first drew attention. Gap statistics use the same cumulative-fixation-mass definition as the benchmark. Neutral applies L3-strength editing to an irrelevant region.

Condition	Preference (%)	Mean Δ	s.d.	Change from L0
L0	73.6	0.371	0.143	0.000
L1	64.9	0.319	0.138	-0.052
L2	56.7	0.258	0.130	-0.113
L3	48.1	0.196	0.122	-0.175
Neutral	72.8	0.364	0.141	-0.007

Table 18 reports subjective preference and gaze-derived gap as complementary manipulation checks. Distractor preference decreases from 73.6% at L0 to 48.1% at L3, while mean Δ decreases from 0.371 to 0.196. The intermediate levels lie between these endpoints on both measures. The neutral edit changes preference to 72.8% and mean gap to 0.364, close to the original condition. Thus, the intended visual competition varies with the ladder, while the same-strength edit elsewhere produces a much smaller change in these summaries.

Preference and cumulative gaze mass should not be treated as interchangeable measurements. A forced choice concerns a participant's judgment about initial attention, whereas the gap integrates the measured fixation distribution over regions. Agreement in their ordering supports the manipulation, but their numerical scales need not match.

G.3 BEHAVIORAL RESPONSE WITHIN ITEMS

Figure 10 summarizes the manipulation check and behavioral ladder. Table 19 reports answer error under all five conditions. Model-macro error decreases from 61.4% at L0 to 57.6%, 52.8%, and 47.9% at L1, L2, and L3. The L0-to-L3 reduction is 13.5 percentage points. All seven models show the same ordering over the attenuation levels, with fitted slopes ranging from −4.97 to −3.99 percentage points per level.

Table 19: Within-item answer error (%) on 1,297 questions from 318 images per model. L0 to L3 retain the same question, target, answer, and candidates; Neutral edits an irrelevant region. Slopes are percentage points per attenuation level and are negative for all seven models ($p < .001$ in each reported fit).

Model	L0	Neutral	L1	L2	L3	Slope
mplug-owl-7B	68.6	67.9	63.8	59.1	53.8	-4.91
Qwen3-VL-2B	67.0	66.4	62.3	56.9	52.2	-4.97
LLaVA-v1.5-7B	64.2	63.5	60.4	55.0	50.6	-4.59
Qwen3-VL-8B	61.3	61.6	57.5	52.5	47.5	-4.65
Qwen3-VL-8B-Thinking	59.1	58.5	55.7	51.3	46.5	-4.21
InternVL3.5-8B	56.6	56.9	53.1	48.4	43.4	-4.43
MolmoPoint-8B	53.1	53.8	50.3	46.2	41.2	-3.99
Macro mean	61.4	61.2	57.6	52.8	47.9	-4.54

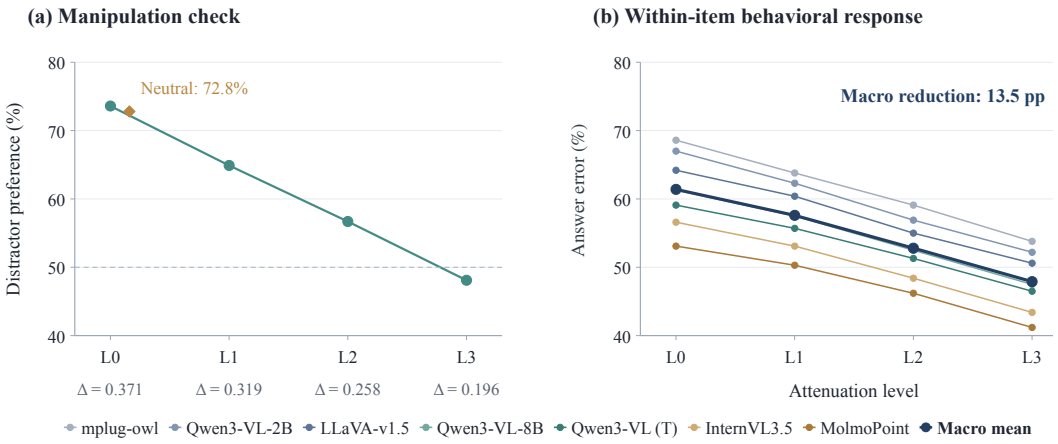


Figure 10: **Manipulation validity and behavioral response.** (a) Distractor preference from the 24-rater check, with mean saliency gaps below each attenuation level; the diamond denotes the irrelevant-region control. (b) Answer error on 1,297 questions from 318 images per model, holding questions, targets, answers, and options fixed. Thick lines average seven models equally. The legend abbreviation Qwen3-VL (T) denotes Qwen3-VL-8B-Thinking throughout the supplementary figures.

Neutral error is 61.2%, only 0.2 points below L0 in the macro summary. The large reduction produced by attenuating the distractor therefore exceeds the change produced by an equally strong edit to an irrelevant region. The comparison supports a role for the competing object, subject to the operator's joint changes to appearance.

The 13.5-point contrast measures the response to a defined edit on fixed instances within this attenuation range and selected item set.

G.4 PATCHING UNDER DIFFERENT ATTENUATION LEVELS

Table 20 extends Table 4(a) with mean gaps and half-decay depths. Distractor-region patching is repeated within the manipulation subset. Table 20 reports recovery at the shallowest and deepest sampled depths, together with the level-specific half-decay summary. As mean gap decreases from 0.371 to 0.196, the reported half-decay depth moves from 0.64 to 0.73. This ordering is consistent with a later loss of patching efficacy under weaker competition.

Table 20: Patching summaries within the manipulation subset. Recovery is evaluated at nominal depths 0.15 and 0.95; N counts eligible model-item pairs at each attenuation level. Eligibility changes with attenuation, so these summaries do not constitute a fixed-source paired comparison across all levels.

Level	Mean Δ	N	$R(0.15)$ (%)	$R(0.95)$ (%)	Half depth
L0	0.371	1342	73.4	9.8	0.64
L1	0.319	1251	71.6	10.9	0.67
L2	0.258	1137	69.3	12.7	0.70
L3	0.196	1024	66.8	14.1	0.73

The eligible source population changes across levels: there are $1{,}342$ eligible model-item pairs at L0 and $1{,}024$ at L3, with intermediate counts at L1 and L2. Attenuation can remove an original error and therefore alter which cases qualify for patching. The depth summaries consequently mix within-item changes in visual input with changes in source-set composition.

The manipulation table uses level-specific eligible cases on selected images, whereas the main depth sweep uses fixed Removal-recovered failures; the two half-depth summaries therefore concern different populations.

H Intervention Implementation and Extended Results

H.1 Fixed failure sets and recovery

For each model, interventions are evaluated on its original strict suppression failures: SG and TG responses are correct, and the AA response is wrong. Let F_m denote this set, and let $a_{mi}^{(v)}$ denote answer correctness under intervention v. Recovery is

$$R_m(v) = 100\,\frac{\sum_{i \in F_m} a_{mi}^{(v)}}{|F_m|}. \tag{5}$$

The denominator is fixed before applying any intervention. Original accuracy is consequently zero on this selected population by construction.

Figure 11 shows recovery rates across the ten-model comparison, which contains $29{,}582$ model-instance failures. Its diamonds report the arithmetic mean of the ten model-specific percentages. This model-macro quantity differs from pooling all recovered instances and dividing by the total failure count, because the models contribute different denominators. Direct contrasts between interventions use the same original failure set within each model. Group averages over intervention types are descriptive summaries of the named conditions.

H.2 Removal and matched control inpainting

Removal uses LaMa (Suvorov et al., 2022) to inpaint the salient distractor region. The rectangular mask is expanded by 12% of the original box width and height on each side. The processing long side is 1024 pixels, with a 7-pixel feather at that scale. Target overlap is forbidden so that the operation does not directly replace the evidence needed for the answer.

Control inpainting uses the same processing settings on a neutral region matched to the distractor in area and eccentricity. The matching tolerances are 10% for area and 0.05 image diagonals for eccentricity. The control tests whether inpainting a comparable patch can itself improve answering. Because inpainting synthesizes pixels conditioned on the surrounding image, matching the mask geometry cannot guarantee identical visual artifacts or semantic consequences at two different locations.

Removal changes more than the saliency score: it removes a competing object and fills the region with generated content. Its large recovery rate therefore supports the importance of that competitor under this edit, while the attenuation ladder provides a complementary test that retains the object.

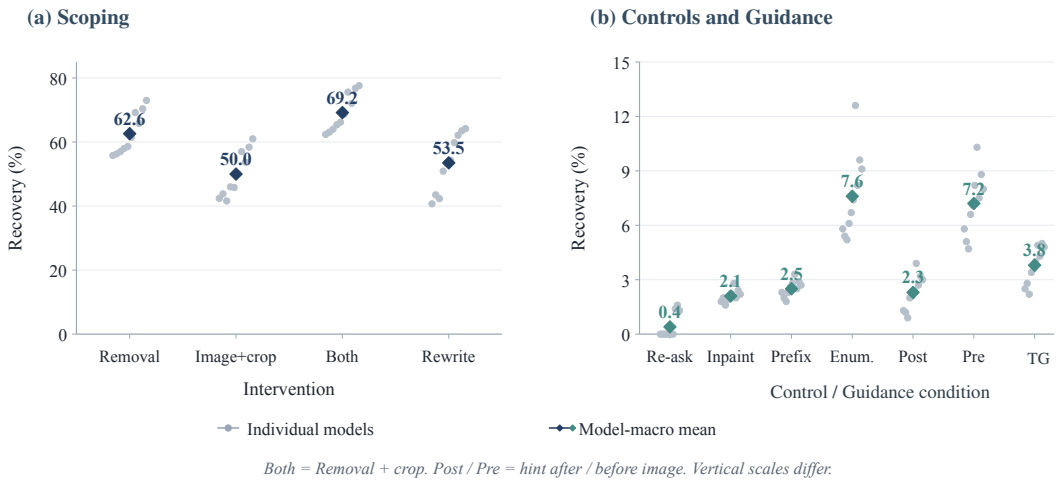


Figure 11: **Recovery under Controls, Scoping, and Guidance.** Each condition uses the same original failure set within a model (29,582 model-instance pairs across ten models). Small points show model rates; diamonds show equal-model means. Horizontal offsets only separate overlapping points. Both combines Removal and Crop; Post and Pre place the hint after and before the image. The panels use different vertical scales.

H.3 Target crops and the joint comparison

Image+Crop presents the original image followed by an enlarged crop of the annotated target. For compositional questions, the crop encloses the union of the task-relevant boxes. The crop rule adds a 15% margin on each side and uses a square extent with side 1.3 times the longer box dimension, subject to a minimum of 160 pixels. The crop is centered on the target region and clipped at image boundaries. Clipping can make the resulting extent non-square near an edge.

The crop is supplied as a second visual input rather than overlaid on the original image. Removal+Crop combines distractor removal in the scene with the same target-focused presentation. This provides two direct comparisons: how much cropping adds after the distractor is removed, and how much removal adds when a target crop is already available. Table 21 reports both contrasts on the fixed original failure subsets.

Table 21: Recovery (%) for the joint removal/crop comparison on ten fixed model-specific failure subsets. The two contrast columns add cropping after removal, or removal when a crop is already present. They are conditional differences within this selected sample.

Model	Crop	Removal	Both	Both − Rem.	Both − Crop
mplug-owl-7B	42.4	55.8	62.4	6.6	20.0
Qwen3-VL-2B	43.8	56.3	63.1	6.8	19.3
LLaVA-v1.5-7B	41.6	57.0	64.0	7.0	22.4
Qwen3-VL-8B	46.0	58.0	65.4	7.4	19.4
Qwen3-VL-8B-Thinking	45.8	58.6	66.2	7.6	20.4
InternVL3.5-8B	50.1	61.5	68.7	7.2	18.6
MolmoPoint-8B	57.0	69.2	75.6	6.4	18.6
Claude-Opus-4.8	53.7	65.8	72.1	6.3	18.4
Gemini-3.1-Pro	58.4	70.4	76.8	6.4	18.4
GPT-5	61.0	73.0	77.6	4.6	16.6
Macro mean	50.0	62.6	69.2	6.6	19.2

Macro recovery is 50.0% for Image+Crop, 62.6% for Removal, and 69.2% for their combination. Cropping adds 6.6 points after removal, ranging from 4.6 to 7.6 across models. Removing the distractor adds 19.2 points when a crop is already present. Thus, exposing the target at a larger scale does not exhaust the benefit of eliminating the competitor, while removing that competitor does not exhaust the benefit of a target-focused view.

These selected-failure contrasts are consistent with overlapping intervention benefits, although bounded recovery and selection for baseline failure limit their mechanistic interpretation.

H.4 Language templates and input order

All language conditions retain the original image and the original four AA answer candidates. Below, Q denotes the original question, O the ordered four-candidate block, L the annotated TG location, and $\widehat{L}$ the model's TG response. The final instruction is to return one option letter. These symbols specify fields inserted into the prompt.

Enumeration. The image is followed by: *"Before answering, list the distinct visible objects that may be relevant to the question. Then answer the multiple-choice question with one letter only."* The original question and four options follow this instruction. Enumeration changes the requested processing of the scene without restricting the question to the target region. Its performance describes the evaluated prompt under the common decoding budget.

Location hint after the image. The image is followed by *"The evidence needed to answer is located in: L."* The question, options, and single-letter instruction follow. This condition states where the evidence lies while preserving the original question's scope. The location must identify the region without stating the answer, consistent with the TG construction rule.

Location hint before the image. The same sentence, *"The evidence needed to answer is located in: L."*, precedes the image, which is then followed by the original question, options, and single-letter instruction. The relevant change is the hint's position relative to the visual input. This contrast tests input order within the evaluated prompt sequence.

Irrelevant prefix. A length-matched, task-irrelevant prefix is placed before the image, followed by the original question, options, and single-letter instruction. It controls for adding text at that position without providing the task's location information. The comparison with a relevant prefix is consequently more specific than comparing the pre-image hint only with a prompt that has no prefix at all.

TG feedback. The image is followed by *"In the previous localization step, you answered: $\widehat{L}$. Use that localization when answering:"*, and then the question and options. The condition ends with the single-letter instruction. It explicitly carries the model's own localization into the answering prompt. Since the intervention population already requires a correct TG response, this condition tests whether reusing that successful localization is sufficient to repair the final answer.

Region-anchored rewriting. The image is followed by *"Restrict the question to the region: L."*, the region-scoped question, the unchanged options, and the single-letter instruction. The rephrased question directs its reference or operation to the target region. It therefore changes the scope over which evidence must be read, in addition to naming a location. This is the distinguishing feature relative to appending a location hint to the original wording.

H.5 Prefix controls and the effect of position

Table 22 gives the full ten-model prefix comparison. A relevant hint after the image recovers 2.3% on average; a length-matched irrelevant prefix recovers 2.5%; moving the relevant hint before the image recovers 7.2%. The relevant-prefix advantage over its matched prefix control is therefore 4.7 points, rather than the 4.9-point difference between pre-image and post-image placement.

The prefix control is not numerically identical to the post-image hint in every model. For MolmoPoint-8B, Claude-Opus-4.8, Gemini-3.1-Pro, and GPT-5, irrelevant-prefix recovery is slightly lower than post-image hint recovery. The stronger supported comparison is the consistently larger recovery of the relevant pre-image hint relative to the irrelevant prefix.

Enumeration and the pre-image hint are also close for the weakest models. For mplug-owl-7B, both recover 5.8%; Qwen3-VL-2B and LLaVA-v1.5-7B show similarly small differences. The

Table 22: Recovery (%) for location-hint and prefix controls on the same original failure subsets. Post and Pre place the relevant hint after and before the image, respectively; Irrel. is a length-matched irrelevant prefix. The final column isolates the relevant-prefix advantage over that control.

Model	N_{SF}	Post	Irrel.	Pre	Pre − Irrel.
mplug-owl-7B	2843	1.3	2.3	5.8	3.5
Qwen3-VL-2B	2943	1.2	2.0	5.1	3.1
LLaVA-v1.5-7B	2554	0.9	1.8	4.7	2.9
Qwen3-VL-8B	3275	2.0	2.3	6.6	4.3
Qwen3-VL-8B-Thinking	3236	2.1	2.6	7.1	4.5
InternVL3.5-8B	3137	2.4	2.8	8.2	5.4
MolmoPoint-8B	3151	3.9	3.3	10.3	7.0
Claude-Opus-4.8	2998	2.7	2.5	7.5	5.0
Gemini-3.1-Pro	2869	3.2	2.9	8.8	5.9
GPT-5	2576	3.0	2.7	8.0	5.3
Macro mean	2958	2.3	2.5	7.2	4.7

ten-model enumeration mean of 7.6% therefore does not imply that enumeration outperforms pre-image hints in every row. Both belong to the group of auxiliary language conditions with limited recovery relative to removal, cropping, and scoped rewriting.

H.6 Region-anchored rewriting across all models

The purpose of a scoped rewrite is to narrow the region over which the original task is evaluated. For an attribute question, the restriction identifies the object whose attribute is requested; for retrieval, it limits where the named entity is sought; for compositional reasoning, it must preserve all regions needed for the relation or count. The correct answer and candidate meanings remain fixed. A rewrite that reveals the attribute itself or removes a required part of a compositional target would change the task rather than instantiate this condition.

Table 23 extends region-anchored rewriting to all 24 models, with mean recovery of 51.0% on original strict suppression failures. On the ten models shared with the full intervention comparison, its mean is 53.5%. These are different model averages, not alternative roundings of one result. The recovery range in the full table extends from 40.7% for mplug-owl-7B to 64.2% for GPT-5, showing that substantial gains are not confined to one open-weight family.

The result prevents a simple division between visual and language interventions. Scoped rewriting is entirely textual yet reaches the same broad recovery scale as target-focused visual presentation. A more useful organizing distinction is whether the intervention restricts the evidence search or reading scope, while retaining the caveat that the conditions change different aspects of the input. This grouping is an interpretation supported by the observed contrasts.

I Probing Implementation and Additional Controls

I.1 Extraction from the AA pass

Hidden states are taken from the actual AA generation pass. For each target annotation, visual tokens are associated with the image-patch footprints they represent. We pool the tokens whose footprints overlap the target box; for compositional questions, the overlap set is defined by the union of all task-relevant boxes. The pooled representation is the arithmetic mean of the selected hidden states at the chosen decoder layer.

If fewer than four visual tokens overlap the target, the readout uses the four tokens with the largest overlap area. This fallback limits the instability of a mean computed from a very small token set. It does not increase the original image resolution or guarantee that the selected tokens isolate the target perfectly. The same pooling rule is applied across models and layers, while the underlying visual tokenization remains model-dependent.

Table 23: Region-anchored rewriting on all 24 models. Recovery is the percentage of original strict suppression failures answered correctly after rewriting. SFR retains the original incongruent denominator; N_{SF} counts eligible failures. The macro summary weights models equally.

Model	SFR (%)	N_{SF}	Recovery (%)
InternLM-XComposer2-7B	56.8	3103	43.8
VILA-8B	55.3	3112	44.6
Llama3.2-Vision-11B	53.5	3252	47.9
mplug-owl-7B	64.0	2843	40.7
MiniCPM-V-4.6	51.2	3258	49.5
Gemma3-4B	58.6	3131	45.1
Phi-4-Multimodal	52.0	3271	50.2
DeepSeek-VL2-Tiny	62.7	2831	41.8
LLaVA-v1.5-7B	60.9	2554	42.3
LLaVA-OneVision-2-8B	54.4	3242	48.7
Qwen3-VL-2B	62.5	2943	43.5
Qwen3-VL-4B	59.8	3129	46.8
Qwen3-VL-8B	56.8	3275	50.9
Qwen3-VL-8B-Thinking	54.9	3236	53.4
InternVL3.5-2B	57.2	2930	46.2
InternVL3.5-4B	54.8	3074	49.7
InternVL3.5-8B	52.0	3137	54.3
Molmo2-8B	51.4	3176	55.1
MolmoPoint-8B	48.2	3151	59.8
GPT-4o	47.0	3112	61.7
GPT-5	37.7	2576	64.2
Gemini-2.0-flash	49.8	3215	58.6
Gemini-3.1-Pro	41.6	2869	63.5
Claude-Opus-4.8	44.6	2998	62.1
Macro mean	53.7	3059	51.0

The readout layer is the nearest layer to relative depth 0.75, using each model's decoder length L. It is layer 24 for the 32-layer models, layer 21 for Qwen3-VL-2B with 28 layers, and layer 27 for the 36-layer models. A common relative depth places the readout late in each decoder without treating layer 24 as equivalent across all architectures. The mapping is audited together with the patching grid in Table 28.

Candidate representations are extracted from the same AA forward pass and combined with the pooled target state to construct visual-candidate interaction features. Under causal attention, visual-token positions preceding the question and options cannot attend to those subsequent tokens, whereas candidate representations can incorporate the preceding image and question context. Probe accuracy therefore characterizes information accessible in the joint feature set and cannot be attributed entirely to the pooled target states.

I.2 Probe training and selectivity

The content probe applies a shared linear scorer to the visual-candidate interaction features of each AA candidate and predicts the highest-scoring candidate. Supervision identifies the correct candidate for each question, with scoring parameters shared across the four candidate positions. The scorer uses L2 regularization, $C = 1.0$, and is fitted by LBFGS for at most $1{,}000$ iterations. Evaluation uses image-grouped five-fold cross-validation with split seeds $\{13, 17, 23, 29, 31\}$. Questions and regions sharing an image remain in the same fold. This grouping prevents a probe from being trained on one question about a photograph and tested on another question about the same photograph.

Hyperparameters are fixed rather than searched separately to maximize the reported test result for each model. A permuted-label control uses the same probe capacity, allowing selectivity to be expressed as the difference in accuracy between the content task and the control. The model-macro content

accuracy is 70.5%, and selectivity is 45.1 percentage points. These are properties of the evaluated readout and label task; they do not assert preservation of every visual detail in the hidden state.

Control-task comparisons are important because a sufficiently expressive readout can exploit regularities that are not the intended linguistic or visual property (Hewitt & Liang, 2019). More generally, probe accuracy alone does not establish how a representation is used by the original network (Belinkov, 2022). The benchmark therefore combines candidate reordering, corpus holdout, and pooling diagnostics with a separate causal patching analysis. Each addresses a different limitation rather than serving as an interchangeable certificate of representational validity.

I.3 Layerwise evaluation and selection of reporting depth

Figure 12 separates the layerwise test profile from the validation record used to select the reporting depth. A fresh probe is fitted at every decoder layer using the same regularization, optimization, target pooling, and image-grouped folds described above. Across the seven models, the sweep covers 236 layers and 5,900 model-layer-seed-fold fits. The test curves aggregate correct predictions and test counts within each model and layer. Each seed evaluates the same 9,196 distinct incongruent questions across its five folds; combining seeds therefore gives repeated test exposures rather than additional independent questions.

The candidate grid is fixed at relative depths $\{0.15, 0.30, 0.45, 0.60, 0.75, 0.85, 0.95\}$. Validation macro accuracy is highest at 0.75 (70.53%), which is selected before reporting the corresponding test results. At that depth, the seven test accuracies are 67.6, 66.4, 69.5, 69.7, 71.1, 72.4, and 76.8%, giving the reported 70.5% macro mean. Selection uses the validation grid rather than each model's highest test point. The test curves show a broad region of high decodability, followed by a modest late decline, rather than a sharp optimum shared by every architecture.

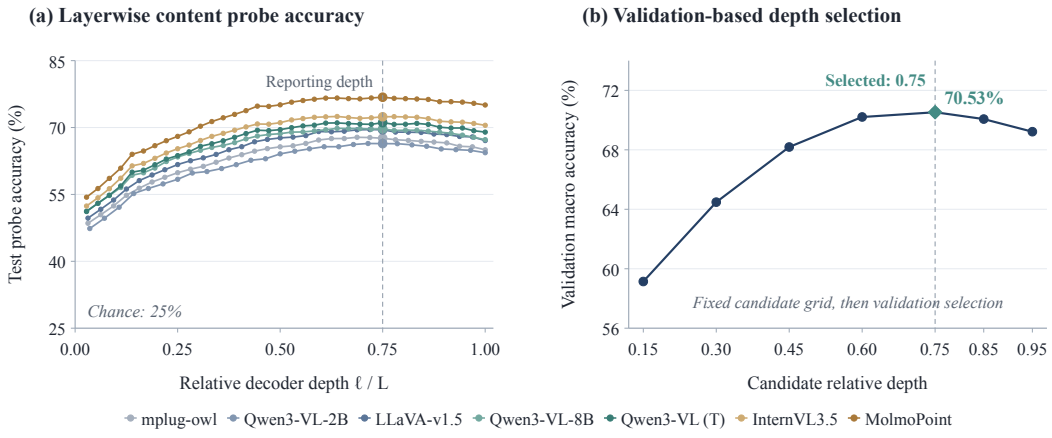


Figure 12: **Layerwise decoding and validation-based depth selection.** (a) Content-probe test accuracy at every decoder layer, aggregated over five seeds and five image-grouped folds. The horizontal coordinate is the actual one-based layer divided by decoder length. (b) Validation macro accuracy on the fixed candidate grid selects relative depth 0.75. Validation scores and test scores refer to different evaluations; the selected test macro accuracy is 70.5%.

I.4 Candidate reordering

Table 24 reports the seven-model candidate-reordering control. Macro accuracy changes from 70.5% to 70.0%, with per-model reductions of 0.4 to 0.6 points. Reordered selectivity remains 44.6 points, and the permuted-label control is close to the four-way chance rate. The observed stability supports limited sensitivity to the tested presentation change.

The conclusion is specific to this control: changing candidate order leaves the reported aggregate readout performance largely intact. It is not equivalent to demonstrating that every prediction is unchanged. The prediction identifies a question-specific candidate, and its features combine target

Table 24: Content-probe accuracy (%) under candidate reordering. Difference is reordered minus original accuracy; selectivity subtracts the permuted-label control from reordered accuracy. These results assess sensitivity to candidate presentation.

Model	Original	Reordered	Difference	Control	Selectivity
mplug-owl-7B	67.6	67.2	-0.4	25.4	41.8
Qwen3-VL-2B	66.4	66.0	-0.4	25.2	40.8
LLaVA-v1.5-7B	69.5	68.9	-0.6	25.6	43.3
Qwen3-VL-8B	69.7	69.1	-0.6	25.3	43.8
Qwen3-VL-8B-Thinking	71.1	70.5	-0.6	25.5	45.0
InternVL3.5-8B	72.4	72.0	-0.4	25.7	46.3
MolmoPoint-8B	76.8	76.3	-0.5	25.4	50.9
Macro mean	70.5	70.0	-0.5	25.4	44.6

and candidate representations. In particular, order stability cannot by itself show that the native decoder consults the recovered evidence when generating its answer.

I.5 Generalization across source corpora

Image-grouped cross-validation prevents direct image overlap but can still train and test on questions drawn from the same source corpus. The held-out-corpus analysis is stricter: one source is excluded from probe training and used for evaluation, repeated across all four corpora. Table 25 gives each held-out result rather than reporting only the average transfer score.

Table 25: Content-probe transfer to a source corpus excluded from training. Columns identify the held-out corpus; Mean gives the average across four held-out evaluations. In-corpus is the original image-grouped result, and Drop is its difference from the held-out mean, in percentage points.

Model	MIT	OSIE	PASCAL	CAT	Mean	In-corpus	Drop
mplug-owl-7B	62.1	63.4	60.7	62.8	62.3	67.6	5.3
Qwen3-VL-2B	60.8	62.2	59.6	61.7	61.1	66.4	5.3
LLaVA-v1.5-7B	64.7	65.8	62.9	65.2	64.7	69.5	4.8
Qwen3-VL-8B	64.9	66.1	63.5	65.7	65.1	69.7	4.6
Qwen3-VL-8B-Thinking	66.5	67.4	64.9	66.9	66.4	71.1	4.7
InternVL3.5-8B	67.6	68.7	66.3	68.1	67.7	72.4	4.7
MolmoPoint-8B	71.2	72.3	69.0	70.5	70.8	76.8	6.0
Macro mean	65.4	66.5	63.8	65.8	65.4	70.5	5.1

Figure 13 contrasts candidate reordering with corpus holdout and displays the model-corpus transfer matrix. Mean held-out accuracy is 65.4%, compared with 70.5% under the original image-grouped evaluation, a drop of 5.1 points. Transfer remains above chance for every model-corpus cell, while the decline shows that within-corpus performance benefits from source-specific regularities. PASCAL-S has the lowest column macro at 63.8% and OSIE the highest at 66.5%. These differences describe transfer under the benchmark's source partition, not a controlled ranking of corpus difficulty.

Source holdout changes the mix of images, questions, tasks, and target geometry, so the transfer drop cannot be attributed to a single factor.

I.6 The four-token fallback

Table 26 reports how often the fallback is triggered and the probe accuracies within each subset. Trigger rates vary from 18.7% for mplug-owl-7B to 6.4% for MolmoPoint-8B. Accuracy is lower on triggered instances in every model, with differences ranging from -5.4 to -2.8 percentage points. This pattern is consistent with reduced visual coverage being associated with harder content decoding.

Because fallback status is associated with target coverage and task difficulty, the subset accuracy gap does not isolate the cost of the fallback itself.

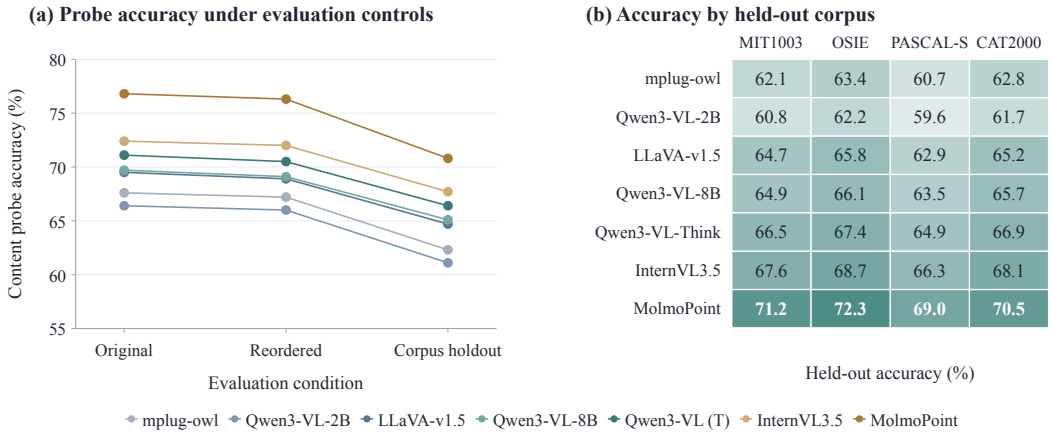


Figure 13: **Content-probe robustness across evaluation conditions.** (a) Accuracy with original candidates, reordered candidates, and corpus holdout for seven models. (b) Held-out test accuracy for each source corpus, with training restricted to the other three corpora. Each heatmap cell is a separate held-out evaluation, not an uncertainty bound. Four-way chance accuracy is 25%; corpus shifts reduce accuracy more than the tested candidate reordering.

Table 26: Content-probe results for the four-token fallback. Trigger rates and accuracies are percentages; Difference compares triggered with non-triggered items. The split is observational: small or poorly covered targets can be more difficult, so it does not estimate the effect of invoking the fallback.

Model	Trigger	Acc., yes	Acc., no	Difference	p
mplug-owl-7B	18.7	63.2	68.6	-5.4	.004
Qwen3-VL-2B	15.9	62.4	67.2	-4.8	.006
LLaVA-v1.5-7B	12.8	65.9	70.0	-4.1	.012
Qwen3-VL-8B	10.4	66.6	70.1	-3.5	.021
Qwen3-VL-8B-Thinking	9.7	68.2	71.4	-3.2	.028
InternVL3.5-8B	8.1	69.6	72.6	-3.0	.036
MolmoPoint-8B	6.4	74.2	77.0	-2.8	.081
Macro mean	11.7	67.2	71.0	-3.8	.003

I.7 Decodability and native AA correctness

The probe-model comparison contains $64{,}372 = 7 \times 9{,}196$ incongruent model-question pairs. Of these, the model answers $25{,}684$ correctly and $38{,}688$ incorrectly. The probe is correct on $18{,}110$ of the model-correct pairs and $27{,}272$ of the model-error pairs; it is wrong on the remaining $7{,}574$ and $11{,}416$, respectively. These counts ensure that congruent questions do not enter the mechanistic accuracy denominator.

Probe accuracy is 70.5% on both model-correct and model-error subsets to one decimal place. Conversely, native AA is 39.9% in both probe-outcome groups. After adjustment for saliency gap, task, corpus, and model, the association between probe success and native success is OR 1.01, with interval $[0.96, 1.06]$ and $p = .71$. The corresponding crude odds ratio is 1.00. The result provides no evidence of a substantial positive instance-level association under this specification.

Among the probe's three false candidates, the error shares are 34.0%, 32.8%, and 33.2% for the distractor-linked alternative and the other two false alternatives. This near-balanced distribution is compatible with the readout not strongly favoring the distractor-linked answer. It remains a distribution conditional on probe error, rather than a direct measure of the information lost at a particular depth. The patching experiment asks the separate question of whether changing selected internal states can alter the model's generated answer.

Native answer errors exhibit a larger distractor-linked share, with an equal-model mean of 48.6% across the same seven models (Table 27). Each model-specific share conditions on that model's incorrect AA responses. Uniform allocation across the three incorrect options gives

a distractor-linked share of 33.3%; the probe's 34.0% is close to this reference, whereas all seven model-specific shares (44.7–52.7%) exceed it.

Table 27: Distractor-linked share (%) of native answer errors on the incongruent evaluation set. Each row uses the model's own incorrect responses as its denominator. The macro mean weights the seven model-specific shares equally.

Model	Distractor-linked share (%)
mplug-owl-7B	52.7
Qwen3-VL-2B	50.9
LLaVA-v1.5-7B	49.4
Qwen3-VL-8B	48.8
Qwen3-VL-8B-Thinking	47.6
InternVL3.5-8B	46.1
MolmoPoint-8B	44.7
Macro mean	48.6

I.8 Where-probe implementation

Assessing region-level decodability requires a readout aligned with TG. The where probe extracts hidden states from the TG forward pass at layer $\ell = \mathrm{round}(0.75L)$, where L is decoder depth. For candidate region j, r_j is the mean of its visual-token states; q is the mean of the TG question-token states. A shared linear scorer with L2 regularization operates on the elementwise interaction $q \odot r_j$. Supervision identifies the task-relevant region among the four candidates, and the highest-scoring region determines the prediction. The probe attains model-macro accuracy of 79.8%, measuring localization information accessible to this question-conditioned readout.

J Activation Patching and Depth Sensitivity

J.1 A depth audit across decoder lengths

Patching uses the nominal relative grid $d \in \{0.15, 0.30, 0.45, 0.60, 0.75, 0.85, 0.95\}$. For decoder length L, the actual layer is $\ell(d) = \mathrm{round}(dL)$, using one-based layer numbering. Table 28 lists every mapping, together with the probe layer. The 36-layer configuration is explicitly included: its patched layers are $5, 11, 16, 22, 27, 31, 34$, rather than the 32-layer sequence reused across models.

Table 28: Decoder-layer audit for probing and patching. Absolute layer numbers are one-based, with $\ell = \mathrm{round}(dL)$. The probe reads at $d = 0.75$ in every model; the remaining columns enumerate the actual patched layers for each nominal depth and model configuration.

Model	L	ℓ_{probe}	.15	.30	.45	.60	.75	.85	.95
mplug-owl-7B	32	24	5	10	14	19	24	27	30
Qwen3-VL-2B	28	21	4	8	13	17	21	24	27
LLaVA-v1.5-7B	32	24	5	10	14	19	24	27	30
Qwen3-VL-8B	36	27	5	11	16	22	27	31	34
Qwen3-VL-8B-Thinking	36	27	5	11	16	22	27	31	34
InternVL3.5-8B	36	27	5	11	16	22	27	31	34
MolmoPoint-8B	36	27	5	11	16	22	27	31	34

The nominal grid facilitates comparison but is not an exact common fractional layer location away from $d = 0.75$. Rounding produces small model-specific differences between d and ℓ/L. For example, nominal 0.30 maps to $8/28$, $10/32$, or $11/36$. The probe depth is exact for all three decoder lengths: $21/28 = 24/32 = 27/36 = 0.75$. This is why absolute layer labels from a single architecture should not be placed on a pooled depth axis.

Our depth summaries use the common nominal coordinates of the sampled curves. They report where recovery declines on that grid, with interpolation between observed points. A later rel-

ative crossing can be meaningfully compared descriptively while leaving the specific internal computation responsible for that crossing unresolved.

J.2 Source construction and coverage

For each model, the source activation comes from its own Removal run on an instance that it fails under the original image and answers correctly after removal. Both the original suppression-failure criterion and successful removal are required. This yields $12{,}604$ model-instance pairs from the $21{,}139$ original strict failures across the seven open-weight models, giving coverage $12{,}604/21{,}139 = 59.6\%$ after rounding.

The denominator is a count of model-instance pairs, not distinct questions or images. The same question can contribute to one model's source set but not another's, because original failure and removal success are model-specific. Within a model, the source set is held fixed across the depth sweep. The recovery curve then measures how often patching changes the answer on that fixed set, rather than changing the evaluated cases at each layer.

All four patch sites in Figure 5(a) and the distractor-region curves in Figure 5(b) are evaluated on these same $12{,}604$ Removal-recovered model-question pairs. Recovery is computed separately for each model and then averaged with equal model weights. The 95% confidence intervals use $1{,}000$ image-clustered bootstrap replicates, with paired resamples shared across panels. Both panels use the model-macro distractor recovery values reported in Table 29.

Failures that Removal does not repair are outside this source construction. They do not provide the required successful counterpart under the chosen intervention, and their exclusion limits the mechanistic scope. The curves therefore characterize removal-recoverable suppression failures. They cannot establish that the remaining 40.4% fail for the same reason, nor that an alternative successful intervention would yield the same source activations or depth profile.

Patching replaces selected hidden states in the original run with the corresponding states from the successful source run, then continues computation to obtain an answer. The measured effect depends on the source, patched site, layer, and behavioral outcome. This dependence is central to activation-patching methodology: a positive effect is evidence about the specified intervention, while the absence of an effect at one site is not a proof that no information remains anywhere in the network (Zhang & Nanda, 2024).

J.3 Complete distractor-region curves

Table 29 reports all seven sampled depths for every model; Figure 14(a) plots these trajectories with their model-macro mean. Model-macro recovery declines from 72.6% at nominal depth 0.15 to 8.5% at 0.95. All seven models follow a decreasing trajectory over the sampled grid. The rate of decline differs, with mplug-owl-7B losing more efficacy earlier and MolmoPoint-8B retaining more recovery into the later samples.

Table 29: Distractor-region recovery (%) at every sampled depth. Each row uses the model-specific source set of Removal-recovered suppression failures throughout the depth sweep. The macro row averages model recovery rates, giving each model equal weight independently of source-set size.

Model	.15	.30	.45	.60	.75	.85	.95
mplug-owl-7B	67.9	65.9	58.4	39.1	17.2	9.2	5.5
Qwen3-VL-2B	69.1	67.8	62.2	44.9	20.5	10.6	6.1
LLaVA-v1.5-7B	70.9	69.8	65.1	49.2	23.8	12.2	6.7
Qwen3-VL-8B	72.9	72.3	68.9	55.7	29.0	14.8	7.8
Qwen3-VL-8B-Thinking	73.9	73.4	70.9	60.4	35.0	18.5	9.4
InternVL3.5-8B	75.7	75.4	73.6	64.9	39.8	21.1	10.4
MolmoPoint-8B	77.9	77.7	76.4	69.7	47.4	27.0	13.4
Macro mean	72.6	71.8	67.9	54.8	30.4	16.2	8.5

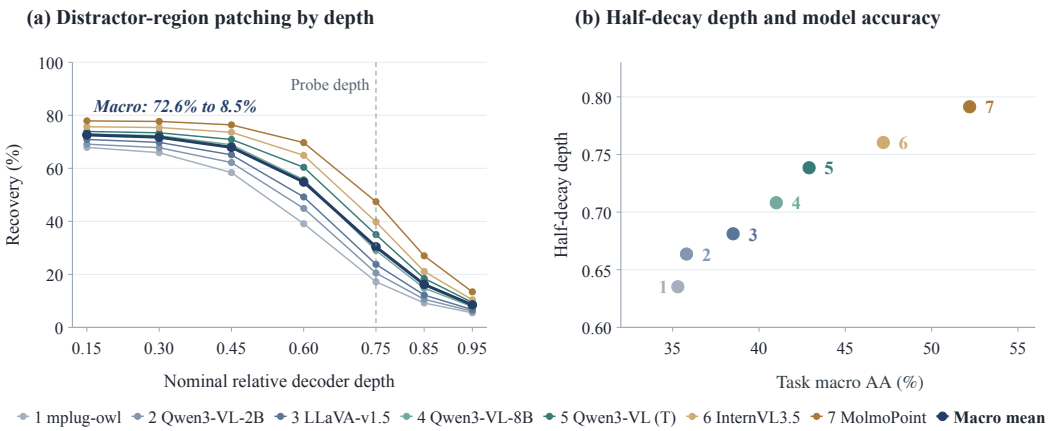


Figure 14: **Depth-dependent visual patching and model accuracy.** (a) Recovery on Removal-recoverable failures, using each model's fixed source set and nearest-layer mapping. The thick curve is the model-macro mean; the dashed line marks probe depth. (b) Interpolated half-decay depth against task-macro AA. Numbers identify the models in the legend. The monotone ordering is descriptive across seven systems with different source sets.

At the probe depth 0.75, macro patching recovery is 30.4%, compared with content-probe accuracy 70.5%. These percentages have different denominators and meanings: probe accuracy uses the full incongruent set for each of seven models, while patching uses the selected successful-source pairs. Their juxtaposition illustrates coexistence of decodable content and reduced intervention efficacy.

The decline at the probe layer also varies across models. InternVL3.5-8B retains 39.8% recovery and MolmoPoint-8B 47.4%, both above half of their shallow values. The other five models have already crossed that threshold by 0.75. Consequently, the data do not support a universal statement that visual patching has lost more than half its efficacy at the probe layer in every model. The half-decay summaries preserve precisely this variation.

J.4 Interpolation and alternative decay thresholds

Let $R_m(d_j)$ be recovery at sampled nominal depth d_j. For a retained fraction q of the shallow value, define $d_m(q)$ by the first sampled interval satisfying $R_m(d_{j-1}) \geq qR_m(d_1) \geq R_m(d_j)$ and linearly interpolate:

$$d_m(q) = d_{j-1} + \frac{R_m(d_{j-1}) - qR_m(d_1)}{R_m(d_{j-1}) - R_m(d_j)}(d_j - d_{j-1}). \tag{6}$$

Half-decay corresponds to $q = 1/2$. This threshold is relative to the shallow recovery, with no subtraction of a neutral-control floor. It is a convention for summarizing the curve.

Table 30: Sensitivity to the retained fraction of shallow recovery. Depths are linearly interpolated from the rounded rates in Table 29; they are descriptive summaries, not additional measurements. AA is the task-macro accuracy from the main model table and is used only for the rank comparison.

Model	AA (%)	Retain $2/3$	Retain $1/2$	Retain $1/3$
mplug-owl-7B	35.3	0.552	0.635	0.713
Qwen3-VL-2B	35.8	0.590	0.664	0.734
LLaVA-v1.5-7B	38.5	0.611	0.681	0.751
Qwen3-VL-8B	41.0	0.640	0.708	0.783
Qwen3-VL-8B-Thinking	42.9	0.666	0.738	0.813
InternVL3.5-8B	47.2	0.686	0.760	0.828
MolmoPoint-8B	52.2	0.720	0.791	0.858

The half-decay depths range from approximately 0.64 to 0.79, with a mean of approximately 0.71 across models. Table 30 also reports retaining two-thirds and one-third of shallow recovery, corresponding to a one-third and two-thirds decline, respectively. These are calculations from the existing rounded recovery table, rather than new experimental measurements or confidence bounds.

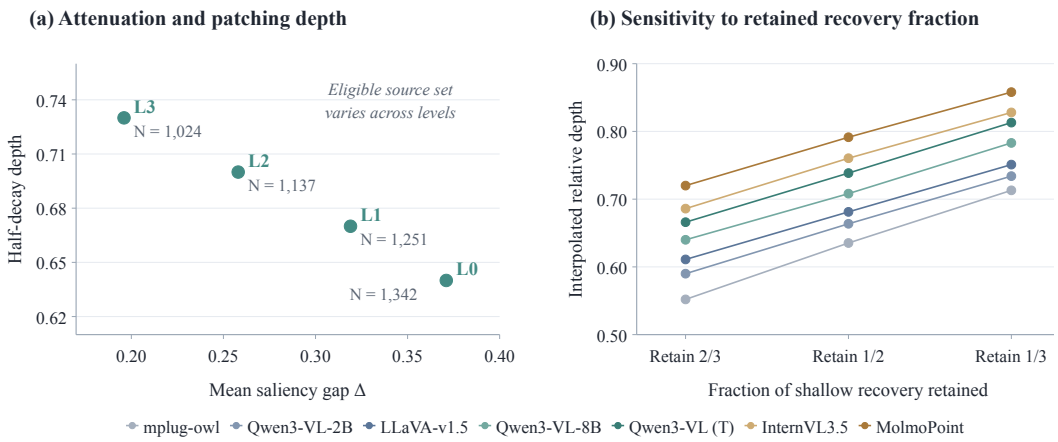


Figure 15: **Conflict strength and decay-threshold sensitivity.** (a) Reported half-decay depths across attenuation levels, annotated with eligible source counts. These populations change across levels, so the points do not estimate a fixed-source paired effect. (b) Depths retaining two-thirds, one-half, or one-third of shallow recovery, interpolated from the rounded seven-depth curves. All three criteria preserve the model ordering; the curves are descriptive summaries, not confidence intervals.

Figure 15(b) shows that all three threshold choices preserve the same model ordering. Thus, the descriptive conclusion that some models retain patching efficacy later in the decoder is not dependent on selecting exactly a half-decay rule. The threshold location itself does change, as expected: a stricter required decline is reached later. Averaging per-model crossing depths is also conceptually different from finding the crossing of an averaged recovery curve, even when both round to a similar macro value.

J.5 Depth and model accuracy

Across the seven systems, half-decay depth increases monotonically with task-macro AA, as shown in Figure 14(b). The sequence runs from mplug-owl-7B through Qwen3-VL-2B, LLaVA-v1.5-7B, Qwen3-VL-8B, Qwen3-VL-8B-Thinking, InternVL3.5-8B, and MolmoPoint-8B. Table 30 places accuracy beside the depth summaries so that the ordering can be inspected directly.

The ordering is descriptive across seven models.

Within the attenuation experiment, weaker conflict is likewise associated with a later reported crossing (Figure 15(a)), but the changing eligible population introduces a separate limitation. The agreement in direction between these observations supports studying how evidence remains causally effective across depth. It does not justify inferring a full-benchmark mean gap by extrapolating a line fitted to four selected attenuation levels.

J.6 What the other patch sites establish

The four-site comparison includes a neutral patch, a target-region patch, a distractor-region patch, and an answer-token residual patch. These sites address different interventions. In particular, the answer-token residual reaches 94.5% recovery at deep decoder layers, while the neutral control remains near its low floor. The high residual-patch value shows that a deep-layer change to the answer pathway can still alter the output under that intervention.

This finding rules out describing the visual-patching decline as a global point after which the answer is impossible to change. A residual stream near readout can remain writable even when replacing selected visual states has little downstream effect. The appropriate conclusion concerns the diminishing efficacy of a particular route from the visual representation to the answer. It does not assert irreversible commitment throughout the entire network.

The four-site overview supplies location controls, whereas Table 29 supports the numerical depth summaries. Together, the experiments distinguish recoverable information, the causal efficacy of changing a visual state, and the continued ability to change the answer through a later residual intervention.