CHANGE THE IMAGE, CHANGE THE ANSWER? PREDICTING VLM RESPONSES TO REGIONAL EDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

A vision–language model that answers an image question correctly may follow a regional edit intended to change the answer-relevant facts, or retain its original answer. We ask whether original-image measurements can rank these responses in multiple-choice VQA without observing the edited image. We propose GLANCE, a three-view score fixed before evaluation contrasting original-answer support on the full image and an enlarged regional crop with support remaining after masking. Edited images supply labels and determine admission to the primary cohorts but never enter the score. On six model-specific Visual7W cohorts selected by post-edit checks, GLANCE ranks target-following above preservation with AUROC 0.707–0.832 and exceeds the strongest tested attention and geometry readouts in five. On two models, GLANCE also ranks changes among originally correct answers without applying crop or placebo admission to the new edits. In matched diagnostic probes, adding masked support to visible supports raises AUROC by 0.124–0.243 across six Visual7W and two single-layer MEGA cohorts, with all refitting-sensitivity intervals above zero. Positive increments also persist in two small subsets where a majority of human raters selects the target answer. Component analyses show that masked-view access matters more than the exact three-view weighting; the tested binary answer-agreement rules generally retain less ranking information than continuous support. The score ranks response tendencies under the tested edits; among changed answers, it does not consistently separate target from third-option responses.

1 INTRODUCTION

Vision–language models (VLMs) answer visual questions with high accuracy (Bai et al., 2025; Li et al., 2025; Gemini Team, 2025). A correct answer does not by itself establish reliance on the relevant visual evidence. A valid regional edit that changes the correct answer tests whether the model updates accordingly. Counterfactual evaluations test this by editing images and checking whether answers update, and they find answers that persist after the evidence has changed (Agarwal et al., 2020; He et al., 2025; Vo et al., 2026). Observing the edited answer and predicting it before the edit are different problems.

Consider a VLM that answers correctly on the original image. After a regional edit intended to change the answer-relevant facts, the model may follow the target answer or preserve its original one. Figure 1 illustrates the problem: two originally correct answers with similar visible-view support respond differently to edits; we test this relation at the population level. ERASER and FPVG assess model responses when evidence is retained or removed (DeYoung et al., 2020; Reich et al., 2023). We instead test whether such measurements predict the same VLM's response to a separately generated regional edit. The score observes neither the realized edit nor an editing instruction. **For a VLM that answers correctly on the original image, can original-image measurements rank subsequent answer changes versus preservation after a regional edit, without observing the edited image?**

GLANCE averages candidate-normalized support for the original answer on the full image and an enlarged regional crop, then subtracts masked-view support. High scores mark answers whose support falls under masking; we test whether these answers follow regional edits. The rule was selected from six candidates on disjoint development data, then frozen. Edited responses determine admission and evaluation labels.

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.

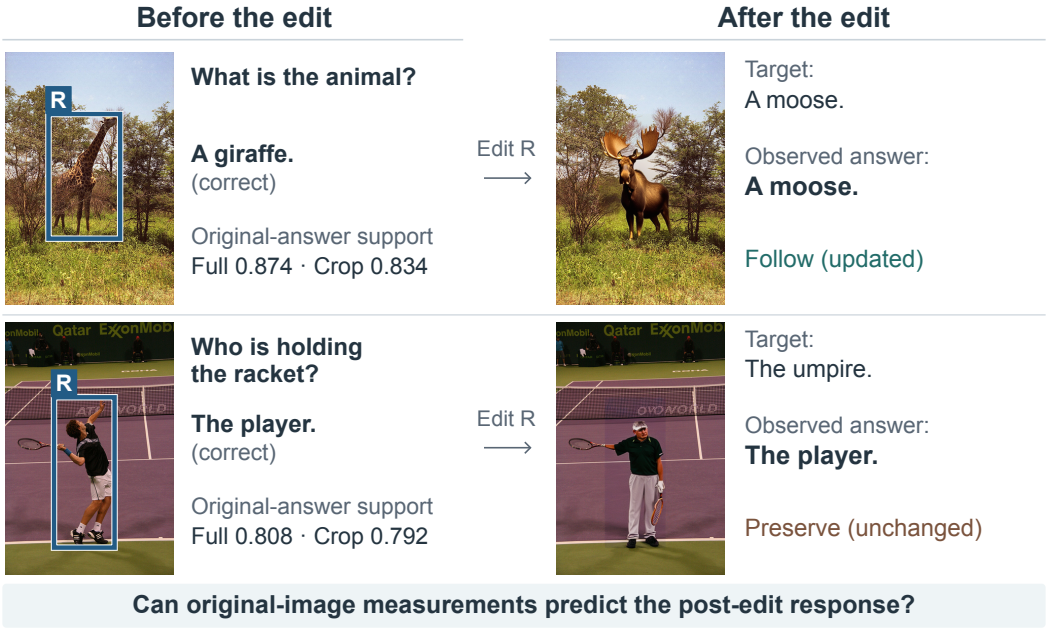


Figure 1: Similar visible support, different edit responses. Two Qwen2.5-VL-3B examples illustrate the prediction problem. Both satisfy the paper's admission checks. Full and crop support are computed from the original image; Figure 2 introduces the scoring path with masked support. These separate examples are illustrative, not paired interventions.

GLANCE exceeds the strongest tested attention and geometry readouts in five of six admission cohorts. Matched probes locate additional predictive information in masked support, including beyond geometry and blank-image references. Among changed answers, the score does not reliably separate target from third options (Section 4.6).

Our contributions are:

- **A task formulation for pre-edit response prediction.** We formulate the problem of ranking a VLM's responses to separately generated regional edits using original-image measurements. The protocol separates score computation from edited-image labels and distinguishes follow–preserve ranking from any-change ranking (Section 3.1; Figure 2).
- **GLANCE: a frozen three-view response predictor.** We introduce a three-view support contrast selected on disjoint development data and frozen during evaluation. GLANCE ranks follow versus preserve with AUROC 0.707–0.832 on six admission cohorts, with additional two-model checks under alternative selection and human-validity restrictions (Tables 1 and 2).
- **A controlled analysis of masked support's predictive value.** We characterize masked support through matched probes, spatial controls, and component analyses. Adding it to visible supports raises matched-probe AUROC in all eight evaluated cohorts. Component comparisons favor masked-view access over the exact weighting, while the tested binary agreement rules generally retain less ranking information than continuous support (Figures 3 and 4).

2 Related Work

Counterfactual evaluation of vision–language models. Counterfactual evaluations distinguish visual changes that should preserve an answer from those that should update it (Agarwal et al., 2020). Controlled visual and question perturbations probe the balance between visual evidence and priors (He et al., 2025; Golovanevsky et al., 2025; Vo et al., 2026; Singla et al., 2026). Edit-based tests assess whether textual reasoning or claimed visual re-examination tracks current evidence (Ding et al., 2025; Menzat et al., 2026; Shi et al., 2026). Generated counterfactual images require validation: benchmark constructions verify them with human raters (Agarwal et al., 2020; Ye-Bin et al., 2024; Awal et al., 2024; Le et al., 2023), and automatic judges agree with human ratings only partially on editing tasks

(Ku et al., 2024). CoSIm predicts answers to described scene changes (Kim et al., 2022), while counterfactual edits have also been used to help users anticipate VQA responses (Alipour et al., 2021). GLANCE instead ranks the same VLM's responses using original-image support measurements, without observing the realized edit or an editing instruction, and tests target identity separately.

Visual attribution and evidence-removal measures. Occlusion, perturbation, and gradient-based attribution locate the input regions a model uses (Zeiler & Fergus, 2014; Petsiuk et al., 2018; Selvaraju et al., 2017). The faithfulness of attention weights is contested (Jacovi & Goldberg, 2020; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), saliency methods can be insensitive to the model and the data (Adebayo et al., 2018), and removal-and-retrain tests rank many attribution methods near random (Hooker et al., 2019). Evidence-removal tests measure dependence directly: the comprehensiveness and sufficiency measures of ERASER compare support under full, rationale-removed, and rationale-only inputs (DeYoung et al., 2020), FPVG tests answer stability under retention and removal of relevant objects (Reich et al., 2023), and VIHD masks visual tokens inside the decoder (Chen et al., 2026). Contrastive decoding also contrasts two output distributions, but to steer generation (Li et al., 2023a; Leng et al., 2024). First-token and likelihood readouts can disagree with generated answers, and multiple-choice selection carries option-position bias (Wang et al., 2024a;b; Zheng et al., 2024). Counterfactual simulatability scores explanations by how well they help predict a model's outputs on counterfactual inputs (Chen et al., 2024; Karvonen et al., 2026). We study a different prediction setting: a pixel-space support contrast of a fixed original answer is evaluated against the same VLM's responses to separately generated regional edits, without changing decoding.

Reliability, confidence, and answer verification. Visual-reliability studies document perceptual errors, language shortcuts, and unsupported claims, and propose debiasing or visual verification (Tong et al., 2024; Guan et al., 2024; Agrawal et al., 2018; Li et al., 2023b; Cadene et al., 2019; Yin et al., 2024). A complementary line estimates when to trust an answer through calibration and selective prediction (Guo et al., 2017; Whitehead et al., 2022), semantic and sampling consistency (Kuhn et al., 2023; Manakul et al., 2023; Zhang et al., 2024), verification questions (Srinivasan et al., 2024; Jin et al., 2026), and probes on internal states (Kogilathota et al., 2026). Whereas these methods target correctness, we condition on an already correct answer and predict its response to a later edit.

3 METHOD

GLANCE uses only original-image views, the prompt, ordered candidates, the original answer, and the specified region. A separate audit path generates edits, applies admission checks, and records response labels (Figure 2); edited images, editing instructions, and labels never enter the score.

3.1 PREDICTION PROBLEM

An item contains image I, region R, prompt q with ordered option texts, candidate letters $\mathcal{A}$, and correct answer a. Model M returns $r_M(J, q) \in \mathcal{A} \cup \{\perp\}$ on image J, with $\perp$ denoting parsing failure. Target a' is the first option whose text differs from that of a. With four distinct option texts, the target is B when the original answer is A, and A otherwise; C and D are never targets. Edit I', meant to make a' correct, changes pixels inside R. Placebo I^{pl} uses the same procedure at a heuristically displaced location that can overlap R.

The response on the edited full image is follow ($r_M(I', q) = a'$), preserve ($r_M(I', q) = a$), or indeterminate, which covers third options and $\perp$. Follow and preserve denote response identities, not judgments of correctness. The originally correct population $C_M = \{x : r_M(I, q) = a\}$ depends only on the original response. Admission adds two edited-image checks,

$$G_M = \mathbf{1}[r_M(I, q) = a \ \wedge \ r_M(\mathrm{crop}(I', R), q) = a' \ \wedge \ r_M(I^{\mathrm{pl}}, q) = a], \tag{1}$$

requiring the target on the enlarged edited crop and the original answer under the displaced edit. The primary population $\mathcal{D}_M = \{x \in C_M : G_M(x) = 1, \ r_M(I', q) \in \{a, a'\}\}$ therefore conditions on post-edit responses; the crop check records model behavior, not semantic validity. We evaluate T1, follow versus preserve on $\mathcal{D}_M$; T2, any parsed change $r_M(I', q) \neq a$ on C_M without post-edit admission; and T3, target a' versus a third option among changed, parsed responses in a stated starting population. We report coverage per target, and scoring uses only (I, q, R, a).

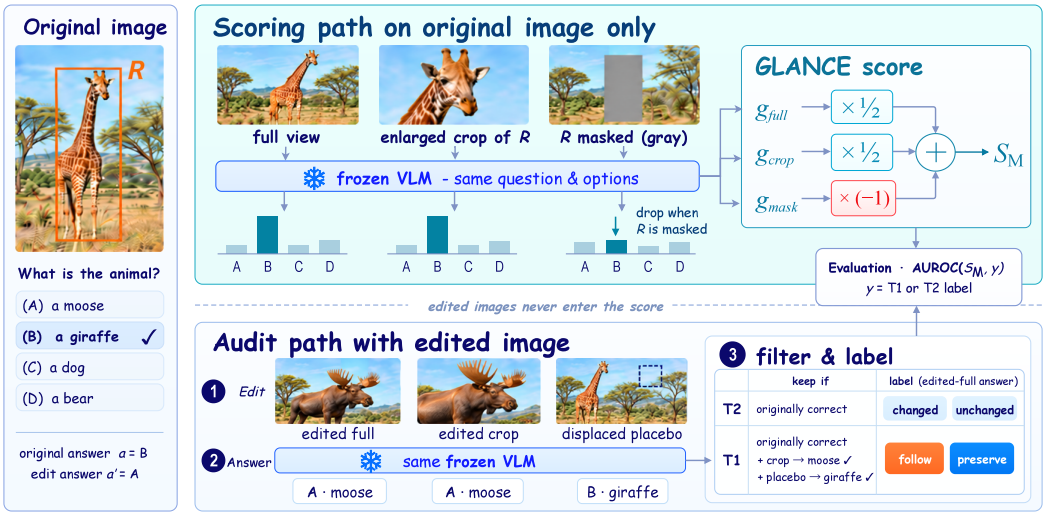


Figure 2: Scoring and audit paths meet only at evaluation. The frozen GLANCE score reads full, crop, and masked views of the original image, while edited images supply admission and labels but never enter the score. T2 evaluates parsed answer changes among originally correct items without crop/placebo admission, and T1 evaluates follow versus preserve on admitted items.

3.2 GLANCE Score

GLANCE reads three views of the unedited image: the full image $J_{\text{full}} = I$, an enlarged crop of R (J_{crop}), and a masked view J_{mask} in which the pixels inside R are replaced by uniform gray (RGB 128, 128, 128), with the surroundings, prompt, and option order unchanged. For candidate letter j and view v, $\ell_{v,j}$ is the mean token log-score of the teacher-forced completion of j, and the support for the original answer is its candidate-normalized probability,

$$p_v(j) = \frac{\exp(\ell_{v,j})}{\sum_{k \in \mathcal{A}} \exp(\ell_{v,k})}, \qquad g_v = p_v(a), \tag{2}$$

which is relative to the candidate set and is not a calibrated probability of correctness. The GLANCE score contrasts the two visible views with the masked view,

$$S_M = \tfrac{1}{2}\left(g_{\text{crop}} + g_{\text{full}}\right) - g_{\text{mask}}. \tag{3}$$

Larger scores rank an item as more likely to follow. The residual original-answer support g_{mask} combines surrounding visual context, question and candidate priors, and the masking intervention. Its predictive value does not identify evidence located outside the region.

With pixel-view ERASER contrasts (DeYoung et al., 2020), $\text{Comp} = g_{\text{full}} - g_{\text{mask}}$ and $\text{Suff} = g_{\text{full}} - g_{\text{crop}}$, GLANCE equals $\text{Comp} - \frac{1}{2}\text{Suff}$. On a V* development subset (Wu & Xie, 2024) disjoint from Visual7W, this rule achieved the highest minimum AUROC across Qwen2.5-VL-7B and 3B among six candidates and was then frozen. Selection used full-option-text scoring; subsequent evaluations retained the formula with option-letter scoring (Appendix B). In $s_{\alpha,\beta} = \alpha g_{\text{crop}} + (1-\alpha) g_{\text{full}} - \beta g_{\text{mask}}$, the frozen weights preserve the score under a common support shift and exchange of the visible supports; Section 4.3 compares other weights. The score is unchanged by the designated target when its inputs are fixed, motivating the separate T3 test.

3.3 Evaluation

We report AUROC with follow as the positive class for T1, any parsed answer change for T2, and the designated target response for T3. Main paired comparisons use 4,000 source-image-group bootstrap draws; the crop−mask contrasts in Table 1B resample items. Each draw evaluates both readouts. Family margins re-select the strongest candidate in panel $\mathcal{P}$: $\Delta = \text{AUROC}(S_M) - \max_{b \in \mathcal{P}} \text{AUROC}(b)$. Main-cohort absolute AUROC intervals use item bootstraps; probe-refitting intervals use 1,000 source-image-group draws. Each table specifies its resampling unit. Intervals are

unadjusted unless a family is named. Without prespecified equivalence bounds, intervals including zero indicate no separation, not equivalence.

Comparators are nine region-blind confidence signals, three pre-registered families of region-aware readouts (attention on the region, gradient attribution on the region, and region geometry), the same-region continuous contrasts $-g_{\mathrm{mask}}$, full$-$mask, and crop$-$mask, and binary answer-agreement rules on the greedy answers F, C, and M of the three original-image views: masked-greedy $\mathbf{1}[F \neq M]$, pixel-FPVG $\mathbf{1}[F = C \wedge F \neq M]$ (Reich et al., 2023), and, within the likelihood channel, argmax agreement b_{TF} over the distributions of Equation 2 (Appendices D and F). Main-text paired comparisons use the same items and labels; table notes identify other sampling schemes. Comparisons requiring parsed answers or internal readouts report coverage.

GLANCE is fixed during evaluation. To assess predictive information in each input, we separately fit matched logistic-regression probes using visible supports $(g_{\mathrm{crop}}, g_{\mathrm{full}})$, masked support g_{mask}, or all three. They share image-grouped five-fold splits over 20 repeats, training-fold standardization, and fixed regularization. Out-of-fold (OOF) AUROC is averaged over repeats. The probe-AUROC increments are $\Delta_{\mathrm{visible}} = A_{\mathrm{all3}} - A_{\mathrm{visible}}$ for adding masked support and $\Delta_{\mathrm{mask}} = A_{\mathrm{all3}} - A_{\mathrm{mask}}$ for adding visible supports.

4 Experiments

4.1 Data and Setup

We audit six VLMs on Visual7W (V7W; Zhu et al., 2016): what, where, and who questions with one human grounding box and four options (correct answer plus dataset decoys). Geometric screening keeps 2,234 of 3,000 sampled items with regions under half the image. SDXL (Podell et al., 2024) edits R with prompts the audited model writes from the question and target. Qwen2.5-VL-7B and 3B produced two edit families; LLaVA-OneVision-7B, Qwen3-VL-32B, Gemma-3-12B-it, and non-thinking Qwen3.6-27B reuse the 7B family (Bai et al., 2025; Li et al., 2025; Qwen Team, 2025; Gemma Team et al., 2025; Qwen Team, 2026). A supplementary bank from the same source and boxes (625 items for 7B, 634 for 3B; two edit seeds each) supplies originally correct pools without post-edit admission. MEGA, our second item pool, uses the GQA and COCO image-prefix subsets of LLaVA-665K (Hudson & Manning, 2019; Liu et al., 2024; Lin et al., 2014), with Qwen2.5-VL-3B-proposed regions; we analyze a post-hoc subset containing only one layer of answer options (7B/3B: 404/234 items).

Model-specific admission responses, labels, and scores define distinct cohorts on shared edits: 3,530 model–item records over 1,610 edit assets. Three raters independently audited class-balanced samples from the two Qwen2.5-VL cohorts (90 and 89 items; 70 and 69 valid). One author manually admitted 202 items, and a blind Qwen3.6-27B judge covers five models' admission cohorts. Scoring uses option letters. Appendix C details construction, funnels, coverage, asset audits, and compute.

4.2 Ranking Edit Responses from Original-Image Measurements

T1 asks whether the frozen score ranks follow above preserve among admitted edits; in the six admission cohorts, which keep only these two responses, 26–33% of items preserve the original answer. GLANCE ranks follow above preserve in all six cohorts, with AUROC 0.707–0.832 (Table 1A), and exceeds the attention and gradient families in all six cohorts and the geometry family in five, with paired intervals above zero (Appendix D.3). Its margin over the strongest of the 18 attention and geometry readouts is 0.105–0.189 in those five and includes zero on Gemma-3-12B-it (Table 1B). On the single-layer MEGA cohorts, GLANCE reaches 0.774 (7B) and 0.745 (3B). Appendix C.7 shows individual items, including a pair that the score ranks in the wrong order.

T2 evaluates answer-change ranking in four originally correct supplementary pools of the two Qwen2.5-VL models, without applying crop or placebo admission to the new edits (Table 25). GLANCE ranks any parsed answer change with AUROC 0.753–0.792 ($n = 611$–618), follow versus preserve with 0.748–0.784, and edited-crop target correctness with 0.612–0.643 (Table 28). Some items were selected using earlier edit outcomes. Post-hoc source strata retain any-change AUROC 0.740–0.791 on previously unused items and 0.732–0.789 on those whose source images are absent from the main sample but share its data source and editing pipeline (Table 2A; full strata in Table 29).

Table 1: Follow–preserve ranking on six model-specific V7W admission cohorts. A: AUROC (follow positive). B: paired differences, GLANCE minus comparator.

A. AUROC estimates

Readout	Qwen2.5 VL-7B $n = 612$	Qwen2.5 VL-3B $n = 688$	LLaVA OneVision 7B, $n = 548$	Qwen3-VL 32B $n = 602$	Gemma-3 12B-it $n = 512$	Qwen3.6 27B $n = 568$
Region-blind signals						
Max-softmax	0.509	0.545	0.472	0.518	0.519	0.542
−Option entropy	0.510	0.547	0.472	0.518	0.520	0.546
Answer mean logprob	0.511	0.536	0.445	0.521	0.440	0.491
Blur: −entropy	0.457	0.507	0.419	0.477	0.489	0.495
Blur: −disagreement	0.478	0.479	0.447	0.478	0.455	0.506
Self-verify: Yes weight	0.426	0.543	0.471	0.533	0.432	0.520
Swap: $1 - g_{\text{swap}}$	0.549	0.520	0.513	0.492	0.460	0.462
Original-answer support	0.509	0.545	0.472	0.518	0.517	0.542
Blank: $1 - g_{\text{blank}}$	0.541	0.525	0.510	0.513	0.456	0.480
Region-aware comparator families						
Best attention	0.586	0.622	0.542	0.562	0.625	0.601
Best gradient attr.	0.599	0.628	0.542	0.577	0.600	0.568
Best geometry	0.645	0.694	0.669	0.657	0.674	0.644
Same-region support readouts						
$-g_{\text{mask}}$	0.777	0.741	0.762	0.755	0.672	0.781
full−mask	0.788	0.769	0.781	0.778	0.676	0.813
crop−mask	**0.814**	**0.803**	0.796	0.782	**0.718**	0.823
GLANCE point estimate	0.812	0.798	**0.797**	**0.789**	0.707	**0.832**
95% CI lower limit	0.772	0.762	0.749	0.743	0.658	0.791
95% CI upper limit	0.848	0.834	0.840	0.833	0.754	0.870

B. Paired AUROC differences: GLANCE minus comparator

Audited model	Best attention/geometry		crop−mask	
	ΔAUROC	95% CI	ΔAUROC	95% CI
Qwen2.5-VL-7B	$+0.167$	$[+0.110, +0.212]$	-0.002	$[-0.012, +0.009]$
Qwen2.5-VL-3B	$+0.105$	$[+0.051, +0.159]$	-0.005	$[-0.017, +0.007]$
LLaVA-OneVision-7B	$+0.128$	$[+0.066, +0.189]$	$+0.001$	$[-0.010, +0.011]$
Qwen3-VL-32B	$+0.133$	$[+0.073, +0.192]$	$+0.007$	$[-0.006, +0.019]$
Gemma-3-12B-it	$+0.033$	$[-0.032, +0.096]$	-0.011	$[-0.029, +0.007]$
Qwen3.6-27B	$+0.189$	$[+0.126, +0.219]$	$+0.010$	$[-0.003, +0.022]$

Reading. Cyan shading identifies GLANCE. Bold marks the highest point estimate per column, not significance. CI endpoints are not ranked; B reports paired comparisons.

Selection. Best denotes family maxima: 12 attention, 12 gradient-attribution, and 6 geometry readouts. B re-selects the best of the 18-readout attention–geometry union in each bootstrap draw. Region-blind directions are frozen; continuous contrasts are post-hoc (Appendix D).

Coverage. All entries use column n items except gradients: 600/602 (32B), 519/568 (27B).

Intervals. GLANCE AUROC: item bootstrap; descriptive panel margin: source-image-group; crop−mask margin: item-paired. Intervals are unadjusted; approximate within-family sensitivity is in Appendix F.4.

In a post-hoc comparison on the four full supplementary pools, GLANCE improves T2 AUROC over the fixed visible-only mean by 0.210 – 0.257, with all paired 95% source-image-group intervals above zero (Table 29B). No probe is fitted.

4.3 Predictive Increment from Masked Support

In matched probes, adding masked support to visible supports raises AUROC by 0.124–0.243 across eight cohorts, all with positive refitting-sensitivity intervals (Figure 3a). Adding visible supports to masked support gains 0.023–0.064: seven refitting intervals are positive; single-layer MEGA 3B spans $[-0.009, +0.091]$. Adding masked support to a reference probe containing visible supports,

Table 2: GLANCE under alternative selection and human validity checks. A: any parsed answer change versus unchanged (T2), without new-edit crop/placebo admission. B: follow versus preserve (T1) on edits where at least two of three raters select the target.

Model	Seed	n (images)	Class counts	AUROC	95% interval
A. T2: previously unused items with source images absent from the main sample					
Originally correct items; counts = changed / unchanged					
Qwen2.5-VL-7B	0	186 (185)	79/107	0.752	[0.679, 0.820]
Qwen2.5-VL-7B	1	186 (185)	73/113	0.732	[0.656, 0.804]
Qwen2.5-VL-3B	0	186 (183)	67/119	0.789	[0.717, 0.854]
Qwen2.5-VL-3B	1	186 (183)	71/115	0.780	[0.708, 0.846]
B. T1: admission-cohort subsets with majority target selection					
From class-balanced audit samples; counts = follow / preserve					
Qwen2.5-VL-7B	—	70 (70)	42/28	0.809	[0.700, 0.904]
Qwen2.5-VL-3B	—	69 (69)	42/27	0.842	[0.742, 0.924]

Populations. A is post-hoc; seeds share items/images from the same source and editing pipeline. B conditions the class-balanced admission-cohort audit on majority target selection. Endpoints differ; panels are not compared.
Intervals. Unadjusted source-image-group bootstrap, conditional on observed validity labels in B. Full pools and audit details: Tables 29, 30, and Appendix G.3.

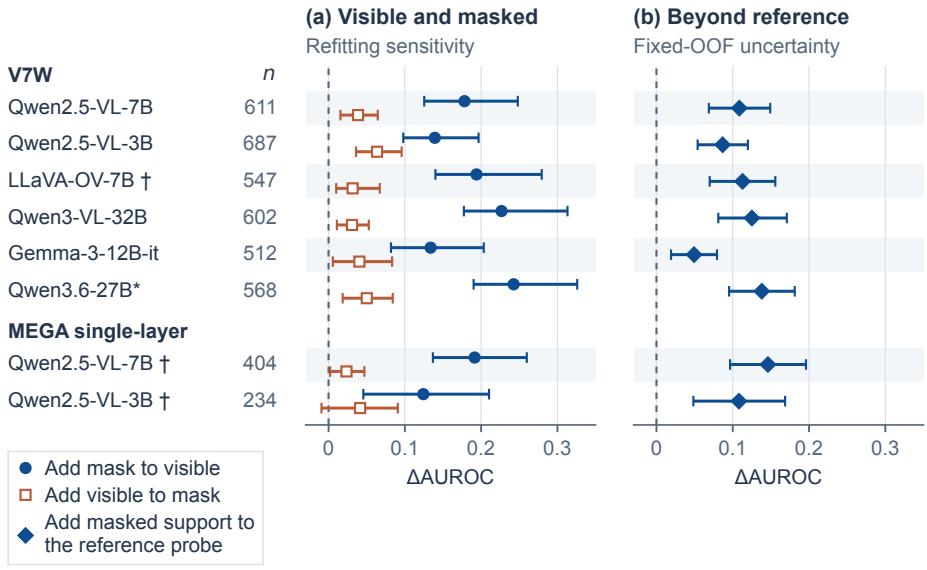


Figure 3: AUROC increments in matched probes. (a) Masked versus visible additions, with refitting sensitivity. (b) Masked support beyond visible, geometry, and blank inputs, with fixed-OOF intervals excluding refitting. Intervals: unadjusted 95%, grouped by source image. †: amended check in (b); MEGA: post hoc (Appendix E.1). LLaVA-OV: LLaVA-OneVision; *: non-thinking.

region geometry, and blank-image support improves AUROC by 0.049–0.146 across eight cohorts, with positive fixed-OOF intervals (Figure 3b; Appendix E.1).

Replacing the scoring mask on R with an equal-area, disjoint mask lowers AUROC by 0.089–0.254 across six restricted pools, all paired intervals positive. The displaced readout retains AUROC 0.576–0.612 but still includes the original full-image and R-crop supports (Appendix E).

Weighting matters less: removing masked support costs 0.101–0.234 even with the best visible-only weights on the tested grid, whereas in-sample tuning gains at most 0.0165 (Appendix E.3). Together, the matched probes, spatial controls, and component comparisons support masked-view access as the most consistent source of additional predictive information in the tested settings. The six main

T1 comparisons do not establish an advantage for GLANCE's exact weighting over crop−mask (Table 1B).

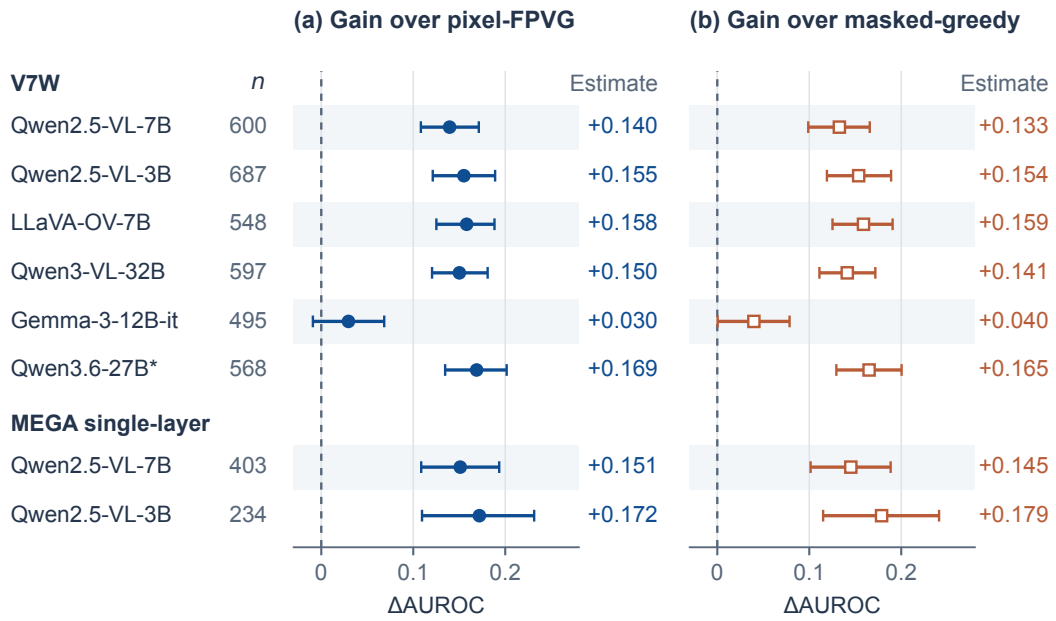


Figure 4: Paired AUROC gains over (a) pixel-FPVG and (b) masked-greedy on eight strictly parsed pools. Intervals: unadjusted 95%, grouped by source image. Both Gemma comparisons include zero after the post-hoc, approximate within-family Bonferroni adjustment (Appendix F.4). *: non-thinking.

4.4 Continuous Support versus Binary Answer Agreement

On common strictly parsed items and views, GLANCE exceeds masked-greedy in eight cohorts: 0.040–0.165 on six V7W cohorts and 0.145 and 0.179 on single-layer MEGA, all paired intervals positive, with Gemma's lower bound at $+0.0004$ (Figure 4; Table 23). Pixel-FPVG contrast intervals are positive in seven; Gemma gives $+0.030$ $[-0.009, +0.068]$. Post-hoc Bonferroni sensitivity across these 16 contrasts leaves both Gemma intervals spanning zero and the other 14 positive (Appendix F.4). GLANCE uses teacher-forced likelihoods; both binary rules use greedy generation. Within the likelihood channel, we take the highest-scoring option in each view's candidate distribution and apply the same agreement rule. GLANCE exceeds this rule by 0.177 $[0.126, 0.228]$ and 0.138 $[0.089, 0.189]$ AUROC on cached Qwen2.5-VL-7B/3B subsets ($n = 250$ and 269). The teacher-forced and greedy versions of the binary rule disagree on only 2 and 3 items, respectively. This comparison changes both discretization and aggregation (Table 24; Appendix F.3).

4.5 Edit Validity: Human-Rated Subsets and Judge Filtering

Preservation can be correct if the edited image still supports the original answer. Majority selection of the target does not establish that the original answer is no longer supported. Among edits where at least two of three raters select the target, GLANCE reaches 0.809 (7B, 70 items) and 0.842 (3B, 69 items; Table 2B). These class-balanced two-model audits describe ranking within the majority-target subsets, not the effect of validity filtering across all six models. In the same samples, AUROC is 0.805 and 0.845 for valid follow versus invalid preserve. The valid-versus-invalid preserve comparisons yield 0.542 and 0.553, with wide intervals and only 17 and 18 invalid preserve items (Appendix G.2). Adding masked support retains positive refitting intervals, $[0.121, 0.531]$ and $[0.194, 0.611]$; the reverse increment is not established under full refitting (Tables 30 and 31). A 202-item pool admitted by one author without model crop/placebo checks gives follow-versus-preserve AUROC 0.734 (Appendix G.1).

A blind VLM judge agrees with one author at $\kappa = 0.471$ (Appendix C.3) but filters response-dependently: among human-valid edits it passes 27/42 follow but 8/28 preserve items for 7B and 38/42 versus 4/27 for 3B (Appendix G.3). This response-dependent selection makes judge-pass subsets selected populations rather than independent confirmation of semantic validity. Discrimination is lower in these subsets, and most paired-contrast intervals include zero (Appendix G.4).

4.6 Scope of Prediction: Answer Change versus Target Identity

Among changed, parsed responses, we test target versus third-option discrimination in twelve cells (six V7W, two single-layer MEGA, four supplementary pools). Ten intervals include 0.5; LLaVA-OneVision gives 0.323 $[0.223, 0.423]$ and 32B 0.374 $[0.278, 0.479]$, below 0.5, each with 33 third-option responses. Supplementary-pool follow-versus-preserve AUROC is 0.748–0.784 versus 0.470–0.534 for target versus third option. An exact-edit VLM audit passes only 6/33, 13/49, 12/33, and 4/32 third-option edits in 32B, Gemma, LLaVA, and 27B, respectively (Appendix H).

4.7 Sensitivity Analyses

Mask fill. GLANCE retains above-chance discrimination under all three alternative fills—blur, noise, and SDXL inpainting—with AUROC 0.711–0.842 across six cohorts (Qwen3-VL-32B: 200-item subsample). Same-item reanalysis retains the advantage over masked-greedy, and the masked-support increment remains positive under every fill(Appendix I.1). Only 2 of the 18 differences from gray are separated from zero, in opposite directions: $+0.017$ for Qwen2.5-VL-3B with blur and -0.034 $[-0.068, -0.002]$ for Qwen3.6-27B with inpainting (Appendix I.1).

Region, edit realization, and placebo geometry. The views depend on R: with boxes from one localizer–model pair, the 123 items whose predicted box has IoU below 0.3 reverse the ranking (0.362 $[0.262, 0.464]$ against 0.825 with the reference box), whereas the 214 items with IoU in $[0.8, 0.95)$ give 0.794 against 0.788 (Appendix I.3). Under two admitted edit seeds, labels change for 21 of 163 (7B) and 27 of 179 (3B) common items with identical scores, while seed-to-seed AUROC differences span zero (Appendix I.2): the score ranks tendencies, not single-edit outcomes. Items whose admission placebo is strictly disjoint from R retain AUROC 0.700–0.834 (Appendix C.5).

5 Discussion

GLANCE gives evidence-removal measurements a predictive use: ranking later regional-edit responses from original-image views. Matched probes and controls support masked-view access more strongly than the exact weighting in the tested settings. This behavioral result does not identify internal mechanisms or establish correct semantic updating. Direct answer selection lowers accuracy, and the tested GRPO gate yields no consistent gain (Appendix J).

T1 conditions on post-edit admission; T2 without crop/placebo admission and the class-balanced human audits cover two models (Sections 4.2 and 4.5). Shared edit assets, one SDXL pipeline, and predominantly Qwen-family backbones limit generalization; cohorts and edit seeds are not independent replications. Black-bar artifacts and content regeneration under inpainting remain limitations (Appendices C.6 and I.1). Crop–question compatibility remains unaudited, and a separate preregistered human validation was not completed (Appendix G.3). Position insensitivity is not established (Appendix H.5).

6 Conclusion

GLANCE ranks VLM responses to regional edits using original-image support measurements, reaching AUROC 0.707–0.832 across six model-specific admission cohorts. Matched probes show that masked support adds 0.124–0.243 AUROC beyond visible support. The evidence is stronger for the predictive value of masked-view access than for uniquely optimal weights. These findings concern response ranking under the tested edits, not reliable target selection or general semantic correctness.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B define the target rule, admission, the frozen GLANCE score, and the evaluation, and Appendix C describes the data construction, audited models, raters, and compute. The supplementary archive `GLANCE-reproducibility.zip` contains the frozen per-item scores and admission records, question and option prompts, region coordinates, edit mappings, judge prompts, implementation snapshots, and pinned CPU dependencies. From cached measurements, `replay.py` recomputes the GLANCE AUROCs, the region-blind panel, and the attention/geometry margins in Table 1, as well as the region-aware family, geometry-matching, admission, and quality tables. `replay_probes.py` refits the matched probes behind Figure 3; further entry points recompute its refitting-sensitivity intervals and reference probes and the mask-fill ablation. All run on a CPU without model inference. The frozen edited images are archived separately because of their size, and the supplementary archive records their hashes. The supplementary and frozen-edit archives are available from the anonymous repository (GLANCE-4843). The validated replay does not cover full GPU reruns (Appendix B.6).

AI USE STATEMENT

Generative AI assistants, including Codex, Claude, and ChatGPT, helped implement the method and write and run experiment and data-processing code, including data cleaning and reformatting. They provided feedback on experimental and analysis plans and supported qualitative analysis and interpretation of results. For mathematical content, AI assistance was limited to organizing statements and algebraic identities. We did not use AI to develop theoretical models or conceptual frameworks, or to propose or refine research hypotheses. This work contains no formal proofs, so providing proof ingredients and writing proofs were not applicable tasks.

The experimental pipeline also uses generative models as instruments: a VLM writes editing phrases and SDXL synthesizes regional edits, the audited VLMs generate answers, and a Qwen3.6-27B judge assesses edit quality (Appendices C and G). Claude judged the qualitative inpainting spot check in Appendix I.1. Human and model judgments are identified separately.

AI assistants also supported literature search, summarization and reference checks, manuscript organization, drafting and language revision, LaTeX editing, and figure and table preparation. An author manually checked the experimental code and the processed data, and reviewed the full manuscript and references. The authors take responsibility for the final manuscript, code, and reported results.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9690–9698, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Agarwal_Towards_Causal_VQA_Revealing_and_Reducing_Spurious_Correlations_by_Invariant_CVPR_2020_paper.html.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Kamran Alipour, Arijit Ray, Xiao Lin, Michael Cogswell, Jurgen P. Schulze, Yi Yao, and Giedrius T. Burachas. Improving users' mental model with attention-directed counterfactual edits. *Applied AI Letters*, 2(4):e47, 2021. doi: 10.1002/ail2.47.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

Rémi Cadene, Corentin Dancette, Hédi Ben-Younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/51d92be1c60d1db1d2e5e7a07da55b26-Abstract.html.

Jiayi Chen, Benteng Ma, Zehui Liao, Winston Chong, Yasmeen George, and Jianfei Cai. VIHD: Visual intervention-based hallucination detection for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, Lecture Notes in Computer Science, pp. 643–653. Springer, 2026. doi: 10.1007/978-3-032-38059-3_61.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do models explain themselves? Counterfactual simulatability of natural language explanations. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7880–7904, 2024.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408.

Sihao Ding, Santosh Vasa, and Aditi Ramadwar. Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. *arXiv preprint arXiv:2510.00047*, 2025. URL https://arxiv.org/abs/2510.00047. NeurIPS 2025 Workshop on Regulatable Machine Learning.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Gemma Team et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Michal Golovanevsky, William Rudman, Michael A. Lepori, Amir Bar, Ritambhara Singh, and Carsten Eickhoff. Pixels versus priors: Controlling knowledge priors in vision-language models through visual counterfacts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24837–24852, 2025. doi: 10.18653/v1/2025.emnlp-main.1262. URL https://aclanthology.org/2025.emnlp-main.1262/.

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 (CVPR)*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6841–6858, 2025. doi: 10.18653/v1/2025.naacl-long.349. URL https://aclanthology.org/2025.naacl-long.349/.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

Mengyuan Jin, Zehui Liao, and Yong Xia. V-Loop: Visual logical loop verification for hallucination detection in medical visual question answering. *arXiv preprint arXiv:2601.18240*, 2026.

Adam Karvonen, Euan Ong, Subhash Kantamneni, and Samuel Marks. Would this change your answer? Evaluating explanations of LLM behavior in the wild with counterfactual experiments. *arXiv preprint arXiv:2608.16747*, 2026. URL https://arxiv.org/abs/2608.16747.

Hyounghun Kim, Abhay Zala, and Mohit Bansal. CoSIm: Commonsense reasoning for counterfactual scene imagination. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 911–923, 2022. doi: 10.18653/v1/2022.naacl-main.66.

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

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. VIEScore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024. doi: 10.18653/v1/2024.acl-long.663.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.09664.

Tiep Le, Vasudev Lal, and Phillip Howard. COCO-Counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

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, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755, Cham, 2014. Springer International Publishing. doi: 10.1007/978-3-319-10602-1_48.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Bayar Menzat, Maximilian Süss, Ruizhi Wang, Benno Steinegger, Thomas Lukasiewicz, and Oana-Maria Camburu. Counterfactual tests for measuring chain-of-thought faithfulness in visual language models. *arXiv preprint arXiv:2609.06704*, 2026. URL https://arxiv.org/abs/2609.06704. Accepted to Findings of EMNLP 2026.

OpenGVLab. InternVL3-8B-hf. Hugging Face model card, 2025. URL https://huggingface.co/OpenGVLab/InternVL3-8B-hf. Accessed 2026-09-17.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. URL https://bmva-archive.org.uk/bmvc/2018/contents/papers/1064.pdf.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Qwen Team. Qwen3-VL-32B-Instruct. Hugging Face model card, 2025. URL https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct. Accessed 2026-09-14.

Qwen Team. Qwen3.6-27B. Hugging Face model card, 2026. URL https://huggingface.co/Qwen/Qwen3.6-27B. Accessed 2026-09-14.

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3129–3144. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.206. URL https://aclanthology.org/2023.findings-emnlp.206/.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? Uncovering the illusion of visual re-examination. In *International Conference on Machine Learning (ICML)*, 2026.

Pratham Singla, Shivank Garg, Vihan Singh, and Paras Chopra. Do vision-language models see or guess? Measuring and reducing textual-prior reliance with a phrasing-controlled benchmark. *arXiv preprint arXiv:2606.10400*, 2026.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12935–12948, 2024. doi: 10.18653/v1/2024.findings-acl.767. URL https://aclanthology.org/2024.findings-acl.767/.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

An Vo, Khai-Nguyen Nguyen, Mohammad Reza Taesiri, Vy Tuong Dang, Anh Totti Nguyen, and Daeyoung Kim. Vision language models are biased. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DG4S2OlGQA.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025. doi: 10.1609/aaai.v39i8.32852.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *Conference on Language Modeling (COLM)*, 2024a.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416, 2024b. doi: 10.18653/v1/2024.findings-acl.441.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/2576_ECCV_2022_paper.php.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. 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. 11–20, 2019. doi: 10.18653/v1/D19-1002.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, 2024. doi: 10.1109/CVPR52733.2024.01243.

Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. BEAF: Observing BEfore-AFter changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73247-8_14.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67:220105, 2024. doi: 10.1007/s11432-024-4251-x. URL https://doi.org/10.1007/s11432-024-4251-x.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014. doi: 10.1007/978-3-319-10590-1_53.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. VL-Uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

APPENDIX CONTENTS

A AI Use Statement **17**

B Method Details and Reproducibility **17**
B.1 Target Rule and Response Categories . . . 17
B.2 Scoring Implementation . . . 17
B.3 Selection of the Score on the Development Subset . . . 17
B.4 What Is Available Before Editing . . . 18
B.5 Original-Answer Contrast, Target Identity, and Cost . . . 18
B.6 Reproduction Materials and Verification . . . 18

C Data, Models, and Populations **19**
C.1 Data Sources and Construction . . . 19
C.2 Audited Models and Admission Funnels . . . 20
C.3 Human Raters and the VLM Judge . . . 21
C.4 Analysis Populations . . . 21
C.5 Placebo Construction and Overlap . . . 23
C.6 Black-Bar Defect in the Original Editor . . . 23
C.7 Case Study . . . 24
C.8 Compute . . . 25

D Comparators **27**
D.1 Region-Blind Signals . . . 27
D.2 Per-Cohort Results Against the Region-Blind Panel . . . 27
D.3 Region-Aware Comparison Families . . . 28

E Masked-Support Evidence **32**
E.1 Matched Probes . . . 32
E.2 Spatial Controls for the Masked View . . . 34
E.3 Score Family and Coefficients . . . 35
E.4 Readout Scale . . . 35

F Continuous Versus Binary Readouts **36**
F.1 Strictly Parsed Comparisons . . . 36
F.2 Continuous Ranking Within Binary Outcomes . . . 39
F.3 Binary Agreement Within the Likelihood Channel . . . 39
F.4 Multiplicity Sensitivity . . . 41

G Populations and Edit Validity **41**
G.1 Populations and Admission Rules . . . 41
G.2 Human-Valid Subsets . . . 43
G.3 Judge Acceptance on Human-Valid Edits . . . 46
G.4 Judge-Pass Subsets . . . 47

H Target Selection and Option Position **48**
H.1 Originally Correct Pools . . . 48
H.2 Twelve-Cell Summary . . . 49
H.3 Edit Quality Behind Third-Option Answers . . . 49
H.4 Sensitivity to Repeated Edits . . . 50
H.5 Option Position . . . 52

I Component Sensitivity **53**
I.1 Alternative Mask Fills . . . 53
I.2 Edit Realization . . . 54
I.3 Region Specification . . . 54

J Other Uses of the Score **56**
J.1 Answer Selection . . . 57
J.2 Training Reward . . . 57
J.3 Audit Triage . . . 57

A AI USE STATEMENT

The complete AI Use Statement appears immediately before the References on page 10.

B METHOD DETAILS AND REPRODUCIBILITY

This appendix gives the implementation details that Section 3 leaves out.

B.1 TARGET RULE AND RESPONSE CATEGORIES

The edit target a' is the earliest option whose text differs from that of the correct answer a, so with distinct option texts any original answer outside position A targets A, and target positions are not randomized. The edited image I' changes pixels only within R, and the displaced-placebo image I^{pl} is generated from the same original image with the same editing prompt. The question and option order are held fixed across all views and edits. Admission depends on the audited model, so a shared edit asset can be admitted for one model and rejected for another. The edited-crop check records the model's response to crop input and does not establish semantic edit validity or region suitability, which the validity audits examine (Appendix G).

Indeterminate covers every response that is neither the target nor the original answer, including other options and parsing failures, and a parsing failure does not show that the model selected a different answer. Indeterminate outcomes remain in the admission funnel and are excluded from the binary follow-versus-preserve comparison.

B.2 SCORING IMPLEMENTATION

Under letter scoring, each candidate completion is an option letter, and the option meanings remain in the question prompt. The mean token log score $\ell_{v,j}$ of Equation 2 is computed over the backend's actual scored suffix after the prompt boundary, including termination tokens retained by its chat template. Admission and labels use the generated response r_M, whereas scoring keeps a fixed without requiring it to maximize the distribution in every view. Analyses that score full option texts are identified separately and compared only within their scoring convention.

Answer generation uses the backend's greedy default. The crop helper truncates box coordinates to integer pixels and resizes the extracted crop to twice its width and height using Pillow's default resize operation, and masks and blank images use RGB $(128, 128, 128)$. The Hugging Face (HF) backend passes pixel bounds $1280 \times 28 \times 28$ and $4 \times 28 \times 28$ to the processor, while LLaVA and Gemma retain their native processing. LLaVA uses native image tiling and ignores Qwen-style pixel-cap arguments, so visual-token budgets are not matched across models. The Qwen3.6-27B production loader uses native processor defaults with thinking disabled.

B.3 SELECTION OF THE SCORE ON THE DEVELOPMENT SUBSET

The score was selected on a V* development subset (Wu & Xie, 2024), disjoint from the V7W evaluation assets, by ranking six candidate readouts by the lower of their AUROCs on Qwen2.5-VL-7B and Qwen2.5-VL-3B, subject to exceeding two baselines on both models. The candidates were crop−mask, full−mask, crop−blank, the equal-weight visible average minus mask, crop−full, and crop support alone, and the baselines were full-image support and one minus blank-image support. The equal-weight average of the two visible supports minus the masked support ranked first and was frozen thereafter. The development subset is small, with 60 and 58 labeled items of which 13 and 12 are preserve responses, and its edits were mostly recolorings rather than SDXL inpainting, so it supports a rule fixed in advance rather than a claim of optimality. This rule addresses opposite rankings of the single-view contrasts: crop−mask performed better on 7B, whereas full−mask performed better on 3B. Averaging attained the highest cross-model minimum AUROC within the finite candidate panel and exceeded both baselines on each model.

The development-set comparison used full option-text scoring, and subsequent letter-scoring analyses retained the selected formula. The later within-family comparisons are also model- and pool-dependent: differences change sign across pools and no member is consistently best. Appendix E

examines sensitivity within the linear family $s_{\alpha,\beta} = \alpha g_{\text{crop}} + (1-\alpha) g_{\text{full}} - \beta g_{\text{mask}}$, whose canonical member is $(\alpha, \beta) = (\frac{1}{2}, 1)$.

B.4 What Is Available Before Editing

For a fixed source-item bank, the inputs to S_M and membership in the originally correct population C_M depend only on original-image information; part of the supplementary bank was selected using earlier edit outcomes (Appendix C.1). T1 additionally depends on post-edit checks: G_M requires the edited-crop and placebo responses, and membership in $\mathcal{D}_M$ also requires a determinate edited-full response, so T1 AUROC describes these conditional populations. The any-change target T2 uses none of these post-edit admission conditions. It evaluates any answer change within the originally correct population and reports label and score coverage separately. Both AUROCs measure behavioral ranking and are not probabilities of semantic edit success on a new image. All views retain the original prompt and option order. Cropping can remove referents needed by relational or global questions, and the crop check records the model's response to that input rather than whether the question is self-contained. No independent crop–question compatibility audit was performed.

B.5 Original-Answer Contrast, Target Identity, and Cost

Equation 3 indexes the original answer a, without distinguishing a particular alternative as a'. Alternative options affect candidate normalization symmetrically. Holding I, q, R, a fixed and changing only the target designation therefore leaves S_M unchanged. On $\mathcal{D}_M$, any departure from a is a follow response because other destinations have been excluded. T3 instead retains parsed changed answers and asks whether the destination is a'. Without an explicit target term, the score can still correlate with the destination, so T3 AUROC need not equal chance. A target-conditioned score would be a different predictor with its own evaluation.

Computing GLANCE requires the original image, question and options, answer a to be scored, region R, and access to candidate-completion scores from the audited model. With $K = |\mathcal{A}|$ candidates, the local implementation scores each candidate separately in each view: K candidate evaluations per view and $3K$ in total. If full-image scores for the same prompt, candidates, and scoring convention are already cached, the additional count is $2K$. The full view preserves joint regional and surrounding context, and its inclusion comes from the development-subset selection described above. In the six main T1 cohorts, GLANCE and crop$-$mask are not separated in AUROC, and that alternative needs $2K$ candidate evaluations if neither view is cached. Runtime depends on batching, caching, and image preprocessing and was not benchmarked under a standardized protocol (Appendix C.8 lists representative measurements). Region localization, original-answer generation, counterfactual construction, and admission checks incur separate costs. Model parameters remain frozen during scoring, and score computation needs neither an edited image nor follow/preserve labels.

B.6 Reproduction Materials and Verification

The accompanying `GLANCE-reproducibility.zip` supplies frozen per-item scores, admission records, question/option prompts, region coordinates, edit mappings, judge prompts, and implementation snapshots. Its `replay.py` recomputes the GLANCE AUROCs, the region-blind panel, and the margins over the strongest attention and geometry readouts in Table 1, the region-aware family and geometry-matching tables (Appendix D.3), and Tables 26 and 33; its `replay_probes.py` refits the eight cohorts in Table 15 and checks their fixed out-of-fold (OOF) intervals. Further entry points recompute the refitting-sensitivity intervals of Figure 3a from all archived draws, the reference probes of Figure 3b, and the mask-fill ablation of Appendix I.1. The archive pins the CPU dependencies for these checks. The separate `GLANCE-frozen-edits.zip` preserves edited-full, edited-crop, and placebo images for the six V7W admission cohorts, and original-image retrieval instructions and asset hashes accompany the mappings. The reproducibility archive also provides the backend-specific code and a reference scoring entry that reports differences from cached scores. The validated CPU replay does not cover full GPU reruns, the original dependency locks, or unrecorded checkpoint revisions.

For Qwen3.6-27B, the production drivers and logs, recovered after the original runs, accompany the caches, and all eight gate and panel shards match their archived checksums. The execution

notes require a newer Transformers environment with offline model loading, because an older installation did not recognize the checkpoint architecture. GPU replay verifies numerical agreement on 96 fixed items: the initial 24 and 72 additional items stratified by input size and follow/preserve response, spanning 64–832 image tokens. Full, crop, mask, blank, and GLANCE scores match the cache with maximum absolute error 0, within the frozen 10^{-5} tolerance, and all four-view parsed answers, admission fields, and labels also match. Repeating 12 expansion items in the same process gives identical scores and generated token IDs. The check covers the 27B admission cohort in that environment on an A800 with 80 GB memory and is neither a population rerun nor a cross-environment comparison. The archived caches lack scores for third-option and parse-failure items (Appendix H scores the third-option items separately), and inference-time image hashes and raw generations from the original runs remain unavailable.

The later V7W probe manifests apply a pre-existing exclusion of one source image unsuitable for academic display, which removes one follow item from each Qwen2.5-VL cohort and from the LLaVA cohort, giving 611/687/547 probe items against the admission-cohort counts 612/688/548. The LLaVA determinate quality audit uses the same 547-item pool. Estimates computed before the exclusion keep their original denominators, and the exclusion is not applied retroactively to those cached statistics.

C Data, Models, and Populations

This appendix gives the data sources, audited models, raters, analysis populations, and edit assets behind Section 4.1.

C.1 Data Sources and Construction

Table 3 lists the sources and the counting unit of each row. The main evaluation uses Visual7W (V7W) (Zhu et al., 2016), whose telling questions carry human answer-grounding boxes. We keep what, where, and who questions with exactly one grounding box, form four options from the correct answer and the three decoys supplied by the dataset, and sample 3,000 of the 31,526 eligible items. A pre-edit geometric screen keeps 2,234 of them and excludes the 766 regions that cover at least half of the image.

The supplementary item bank, labeled H1 in the appendix tables, draws on the same source and the same human boxes. Per model it combines 200 items that previously passed original correctness but failed the edited-crop check, 200 items from the previous determinate pool, and 320 unused Visual7W items, which the same screen reduces to 625 (7B) and 634 (3B) items, each edited under two seeds.

GQA-derived multiple-choice questions (MCQs) use conversations from the GQA image-prefix portion of the LLaVA-665K mixture (Hudson & Manning, 2019; Liu et al., 2024). The construction draws distractors from answer-type pools and derives regions with Qwen2.5-VL-3B. MEGA, our second item pool, applies the same construction to the GQA and COCO image-prefix portions of the mixture, where the prefixes identify image sources. Its converter removes free-answer suffixes but passes through questions that already contain multiple-choice options, and wrapping them in another question makes the correct outer option a bare letter while the distractors come from a word pool. The source audit identified $1{,}759/2{,}397$ affected items. The analysis therefore restricts MEGA post hoc to single-layer questions (404 items for Qwen2.5-VL-7B and 234 for Qwen2.5-VL-3B) and reuses the frozen edits, responses, labels, and scores. The structural screen establishes a single option layer, while semantic edit quality remains a separate question. The V* development subset served only to select the score (Appendix B).

For each item, the audited model writes a short editing prompt from the question and the target answer, SDXL inpainting (Podell et al., 2024) generates a patch for a square context around R, and only the pixels inside the box are pasted back. The archived drivers load the ModelScope checkpoint `AI-ModelScope/stable-diffusion-xl-1.0-inpainting-0.1` without pinning a revision. The placebo repeats the procedure on a displaced box. Because the prompts come from the audited model, the Qwen2.5-VL-7B and 3B audits produced two edit families, and LLaVA-OneVision, Qwen3-VL-32B, Gemma-3-12B-it, and Qwen3.6-27B reuse the 7B family. The six main binary cohorts comprise 3,530 model–item records over 1,610 distinct cached edit assets, 922 from the 7B family and 688 from the 3B family.

Table 3: Data sources and their roles. Counts refer to the stage named in each row and are not interchangeable.

Source and counting unit	Region source	Models	Role
V7W cached candidate edit records: 2,234 per edit family	Visual7W human answer-grounding boxes	Six audited models	Main conditional cohorts (T1); target selection (T3)
Supplementary item bank (H1): 625 (7B) and 634 (3B) items, two edit seeds per model	Visual7W human answer-grounding boxes	Qwen2.5-VL-7B and 3B	Originally correct populations without post-edit admission (T2); repeated edits; alternative admission
GQA-derived MCQs: 295 and 245 items	Regions from Qwen2.5-VL-3B	Qwen2.5-VL-7B and 3B	Extension to derived regions
MEGA single-layer subset: 404 and 234 items	Regions from Qwen2.5-VL-3B	Qwen2.5-VL-7B and 3B	Extension to a second construction source
V* development subset: 60 and 58 labeled items	V* Bench target-object boxes (union)	Qwen2.5-VL-7B and 3B	Selecting the score

Table 4: Analysis pools and audited models. Counts are determinate admitted items under each model's own admission with letter scoring; the six V7W cohorts form the main evaluation, and the remaining pools are extensions with the roles listed.

Data pool	Audited model	Language backbone	n	Role
V7W	Qwen2.5-VL-7B / 3B	Qwen2.5-7B / 3B	612 / 688	Initial audits; own edit families
V7W cached edits	LLaVA-OneVision-7B	Qwen2-7B	548	Audited-model extension; shared assets
V7W cached edits	Qwen3-VL-32B-Instruct	Qwen3-32B	602	Audited-model extension; shared assets
V7W cached edits	Gemma-3-12B-it	Gemma-3	512	Audited-model extension; shared assets
V7W cached edits	Qwen3.6-27B	Integrated Qwen text backbone	568	Non-thinking audited-model extension
GQA-derived MCQs	Qwen2.5-VL-7B / 3B	Qwen2.5-7B / 3B	295 / 245	Derived-region extension
MEGA single-layer subset	Qwen2.5-VL-7B / 3B	Qwen2.5-7B / 3B	404 / 234	Structurally filtered extension

The language-backbone column names the language model or integrated text component of each audited model. Coverage is partial across dataset–model combinations; each estimate in the paper uses its listed cohort.

C.2 Audited Models and Admission Funnels

The six audited models are Qwen2.5-VL-7B/3B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2025), Qwen3-VL-32B-Instruct (Qwen Team, 2025), Qwen3.6-27B (Qwen Team, 2026), and Gemma-3-12B-it (Gemma Team et al., 2025). Each supplies its own admission responses, edited-full labels, and likelihood scores. The language-backbone column of Table 4 identifies the language model or integrated text component. LLaVA-OneVision-7B combines SigLIP, an MLP connector, and Qwen2-7B (Li et al., 2025). Gemma-3-12B-it combines SigLIP-400M with the Gemma-3 language backbone, extending coverage beyond the Qwen series (Gemma Team et al., 2025). Qwen3.6-27B has an integrated text component with unspecified separate initialization.

Admission requires a correct original answer, the target on the edited crop, and an unchanged answer on the displaced placebo, followed by restriction to determinate edited full-image responses. The LLaVA-OneVision, Qwen3-VL-32B, Gemma, and Qwen3.6-27B audits reuse the V7W edit assets of the Qwen2.5-VL-7B audit and query each audited model on them. For Qwen3.6-27B, the 2,234 cached records yield 2,073 originally correct answers, 642 target responses on the edited crop, 600 stable placebos, and 568 determinate edited-image responses (423 follow and 145 preserve). The Qwen3-VL-32B run covers all 2,234 cached edits and yields 602 determinate items (443 follow, 159

preserve). Gemma reuses these edits and the same ordered admission rule: 2,234 records yield 1,891 originally correct answers, 611 edited-crop target responses, and 512 determinate admitted responses (344 follow, 168 preserve), with 2 letter-parse failures across 8,936 view-level responses (2,234 edits, four views each). Both failures occur on the edited full image and are excluded from the 512-item determinate cohort. Each model's admission yields a different conditional population.

The Qwen3.6-27B pipeline stops after a failed admission check, so not every candidate record carries all four responses. The LLaVA-OneVision-7B, Qwen3-VL-32B, and Gemma caches retain all four response fields. The supplementary pools also compute all four and their originally correct populations have a parsed edited-image letter and three-view scores for every item (611, 611, 618, and 618). The strict-parsing subsets used for greedy readouts and the frozen probe lists are analysis populations, not admission stages (Table 5).

C.3 Human Raters and the VLM Judge

Three raters independently answered the edited images for class-balanced samples of the two Qwen2.5-VL cohorts (90 and 89 items) under shuffled option positions. For this audit, we classify an edit as target-valid when at least two of the three raters select the designated target (70 and 69 items). One of the three raters is an author; each rated separately, and none saw the audited models' responses. Across all 90 and 89 items, Fleiss' κ for whether a rater's answer selects the target is 0.424 and 0.496, with mean pairwise agreement 0.785 and 0.820 and prevalence- and bias-adjusted kappa (PABAK) 0.570 and 0.640. Target selections account for about 0.75 and 0.77 of votes, respectively; we report these proportions alongside raw agreement and κ. A separate human admission pool for Qwen2.5-VL-7B combines two rounds annotated by one author into 202 determinate responses, admitted by region relevance and an independently chosen edited answer rather than by the model's crop and placebo checks.

The VLM judge (Qwen3.6-27B, non-thinking) shares a Qwen language backbone with five of the six audited models and makes two calls that never see each other's output. One call answers the edited full image alone and independently selects a unique answer. The other sees the original image, the edited image, and R, and assesses region relevance, edited-evidence readability, and preservation of necessary information outside the region. Neither call receives the target, the audited model's response, a label, a score, or a human judgment. An offline join assigns pass when the independent answer matches the frozen target and all quality judgments pass, and uncertain and unavailable outcomes remain separate. Of 432 held-out human-reference comparisons with one author, 406 are comparable and 26 are unknown or noncomparable, giving $\kappa = 0.471$ on comparable cases. Re-presenting 100 audit units in a fixed alternative presentation changes 12 of the judgments, so we use the judge as a presentation-sensitive stratification variable.

The exact-edit judge audit covers all 602 Qwen3-VL-32B and 512 Gemma determinate items, with $329/602 = 0.547$ and $279/512 = 0.545$ passing, and all four third-option pools in Table 37. On Qwen3.6-27B, only the 32 third-option edits are audited. Before this extension, an edit had been audited only if it entered an earlier audited pool, which required admission by another model's machine gate. When edits are grouped by the number of earlier audited pools that contained them ($0, 1, 2, \geq 3$), admission-cohort pass rates are $0.209, 0.368, 0.650, 0.641$ for 32B and $0.138, 0.380, 0.691, 0.721$ for Gemma, so restricting quality estimates to previously covered edits would overstate pass rates for these populations. Across five independent inference runs on 32 audit units at temperature zero, 31 units retain the same judgment throughout and one changes, a per-unit flip rate of 0.031 (Wilson 95% interval $[0.006, 0.157]$). Both the presentation and the repeat-inference checks quantify variability in VLM judgments rather than agreement with semantic ground truth.

C.4 Analysis Populations

Table 5 lists the eight admission cohorts with the analysis sets derived from them, and Table 6 lists the other analysis populations with the appendix that uses each. Throughout, a cohort is one of these eight model-specific admission populations, and a pool is any other analysis population, including a subset derived from a cohort.

Table 5: The eight admission cohorts and the analysis sets derived from them. Each audited model supplies its own admission responses and labels, so each row is a different population on shared edit assets. The derived sets are restrictions used by particular analyses, not further filtering stages of the main results, and equal counts do not imply identical memberships.

	Admission cohort	Derived sets (n)		Changed responses
Audited model (data)	n (follow/preserve)	Strict parsing	Probe list	n (follow/third)
Qwen2.5-VL-7B (V7W)	612 (435/177)	600	611	473 (435/38)
Qwen2.5-VL-3B (V7W)	688 (494/194)	687	687	565 (494/71)
LLaVA-OneVision-7B (V7W)	548 (408/140)	548	547	441 (408/33)
Qwen3-VL-32B (V7W)	602 (443/159)	597	602	476 (443/33)
Gemma-3-12B-it (V7W)	512 (344/168)	495	512	393 (344/49)
Qwen3.6-27B (V7W)	568 (423/145)	568	568	455 (423/32)
Qwen2.5-VL-7B (MEGA SL)	404 (184/220)	403	404	194 (184/10)
Qwen2.5-VL-3B (MEGA SL)	234 (104/130)	234	234	111 (104/7)

Admission requires a correct original answer, the target on the edited crop, and an unchanged answer on the displaced placebo, and the cohort keeps determinate edited-image responses. Strict parsing additionally requires parsed greedy answers on all three original-image views and is used for the binary-rule comparisons (Appendix F). Probe lists exclude one source image unsuitable for display (Appendix B.6). Changed responses pair follow with parsed third-option answers for the target-selection test T3 (Appendix H). MEGA SL denotes the single-layer MEGA subsets, and Qwen3.6-27B runs in non-thinking mode.

Table 6: Other analysis populations, with the appendix that uses each. Every row has its own denominator, and the rows overlap in items and edit assets, so they are not independent replications.

Population	Selection	n	Appendix
GQA-derived MCQs (7B and 3B)	Audited-model admission on regions proposed by Qwen2.5-VL-3B	295; 245	D
Restricted scoring masks (six cohorts)	Displaced mask strictly disjoint from the region, with exactly matched area	387–529	E
Same-channel cached subsets (7B and 3B)	Strict-parsing items with cached teacher-forced distributions	250; 269	F.3
Originally correct pools (supplementary bank; 7B and 3B, two edit seeds)	Correct original answer only, with no crop or placebo admission	611 (7B) and 618 (3B) per seed	G.1, H
Alternative admission (supplementary bank; seeds pooled)	Crop and placebo checks by the audited model, or by LLaVA-OneVision-7B	948; 878	G.1
Human admission pool (7B, seed 0)	One author, two rounds: relevant region and an independently selected edited answer matching the target	202 (131/71)	G.1
Three-rater audit (7B and 3B cohorts)	Class-balanced samples; a majority selects the target	90 and 89 rated; 70 and 69 valid	G.2, G.3
VLM-judge pass subsets (five cohorts)	Qwen3.6-27B answers the target and passes all quality checks	279–419	G.4
Third-option quality audit (four cohorts)	Exact-edit judge audit of the edits behind third-option answers	32–49	H
LLaVA-OneVision-7B on supplementary edits (two seeds)	Admission by LLaVA's own checks; changed responses for T3	162; 155 changed	H
Repeated edits (7B and 3B)	Admitted under both edit seeds	163 and 179	I.2
Disjoint-placebo subsets (six cohorts)	Admission placebo strictly disjoint from the region	391–535	C.5
Option rotation (7B and 3B)	Four cyclic option orders on sampled admitted items	250 each; 203 and 193 in all orders	H.5
Localization strata (7B; Qwen3-VL-32B boxes)	Predicted-box IoU below 0.3, or in $[0.8, 0.95)$	123; 214	I.3
Answer selection (7B and 3B)	All multiple-choice items, with no admission	1,200 each	J

Table 7: Sensitivity to strictly disjoint placebo placement in the six V7W admission cohorts. The reconstructed placebo box intersects the designated region in 22.2 to 25.5% of each cohort; on the strictly disjoint subsets GLANCE retains AUROC 0.700 to 0.834. The subsets are nested and differ in class composition, so the change from the full cohorts is not an effect of repairing the placebo.

Audited model	Overlap / cohort n (%)	Disjoint subset (follow/preserve)	Disjoint AUROC [95% CI]
Qwen2.5-VL-7B	147/612 (24.0%)	310/155	0.810 [0.764, 0.849]
Qwen2.5-VL-3B	153/688 (22.2%)	359/176	0.792 [0.749, 0.831]
LLaVA-OneVision-7B	140/548 (25.5%)	289/119	0.802 [0.752, 0.848]
Qwen3-VL-32B	140/602 (23.3%)	321/141	0.803 [0.755, 0.849]
Gemma-3-12B-it	121/512 (23.6%)	240/151	0.700 [0.646, 0.752]
Qwen3.6-27B	135/568 (23.8%)	305/128	0.834 [0.790, 0.874]

Overlap means positive intersection area between the designated region and the reconstructed placebo box after editor rounding and clipping to image bounds. Original scores and labels are frozen. Brackets use $4{,}000$ source-image-group bootstrap draws, unadjusted. Geometry is reconstructed for all items and checked against archived image support; direct comparisons to locally available originals cover 288 distinct edit assets. This audit concerns admission placebos, not the displaced masking control W.

C.5 Placebo Construction and Overlap

For each image, the placebo helper retains the designated box's integer width w and height h. It hashes the item identifier with separate horizontal and vertical suffixes, converts each leading hash fragment to a value in $[0, 1]$, and scales them to the admissible top-left coordinate range. If the candidate overlaps the original box according to the helper's coordinate test, it relocates to an opposite image corner. The code applies no object detector, semantic foreground exclusion, or final disjointness assertion, so large boxes can defeat the geometric avoidance heuristic. We call this placement displaced, which names the intended rule rather than a guaranteed separation.

Both edits use the same short target-description prompt, generated from the question and intended answer. The inpainting helper extracts a square context around the chosen box, resizes it to the configured resolution, masks the box, generates a patch, and pastes only the box pixels back into the original image. Its archived class defaults use 25 denoising steps and guidance 7.5, and the driver exposes resolution and seed settings. These defaults document the archived implementation, not a verified configuration log for every original edit. Frozen images and hashes preserve the actual stimuli used in the reported audits.

Placebo stability therefore checks a second edit response under this geometric construction. It does not certify that the displaced box lacks answer-relevant content, so a legitimate answer change caused by that content can lead to rejection, as can edit artifacts.

An asset audit reconstructs this placement for all 3,530 records in the six V7W admission cohorts. After rounding and clipping to image bounds, positive intersection area occurs in 22.2–25.5% of each cohort (Table 7). Across 1,610 distinct edit assets, edited and placebo images agree pixelwise outside the union of the two boxes. Direct comparison with available originals additionally checks 288 assets: all changes stay inside their respective boxes, and all 58 overlapping cases contain placebo changes inside R. The remaining assets have reconstruction and union-support checks, without direct original-image comparisons.

Strictly disjoint subsets retain AUROC 0.700–0.834 with the original scores and labels frozen. Disjointness does not establish semantic irrelevance, and this audit covers neither the separate displaced masking control W (Appendix E) nor the regenerated supplementary-bank edits.

C.6 Black-Bar Defect in the Original Editor

The original editor extracts a square context window capped by the image's shorter side. When the rounded region's longer side exceeds that cap, cropping the generated patch outside the window inserts black pixels. We flag this geometric condition and independently measure the proportion of exactly black RGB pixels inside the clipped region. Among $1{,}610$ unique packaged edits, 215 (13.35%) meet the geometric condition. Of these, 194 have more than 5% black pixels and 21 do not, while 3 geometrically unaffected edits exceed that threshold, so the two criteria disagree on 24

Table 8: Sensitivity to the black-bar editing defect in the six V7W admission cohorts. Affected status is geometric; frozen scores and labels are retained. Removing affected edits moves point estimates in both directions by small amounts, and discrimination stays above chance in every subset, including the affected items alone.

Audited model	Subset	n (follow/preserve)	AUROC [95% CI]
Qwen2.5-VL-7B	full cohort	612	0.8119 [0.7739, 0.8483]
	unaffected	543	0.8100 [0.7682, 0.8488]
	affected	69 (41/28)	0.7944 [0.6828, 0.8960]
Qwen2.5-VL-3B	full cohort	688	0.7982 [0.7600, 0.8348]
	unaffected	596	0.8029 [0.7602, 0.8422]
	affected	92 (54/38)	0.7451 [0.6419, 0.8407]
LLaVA-OneVision-7B	full cohort	548	0.7970 [0.7507, 0.8427]
	unaffected	476	0.7926 [0.7396, 0.8443]
	affected	72 (45/27)	0.7638 [0.6460, 0.8642]
Qwen3-VL-32B	full cohort	602	0.7890 [0.7430, 0.8328]
	unaffected	533	0.7813 [0.7314, 0.8284]
	affected	69 (43/26)	0.8095 [0.6722, 0.9197]
Gemma-3-12B-it	full cohort	512	0.7068 [0.6584, 0.7546]
	unaffected	446	0.6926 [0.6403, 0.7435]
	affected	66 (36/30)	0.7963 [0.6831, 0.8947]
Qwen3.6-27B	full cohort	568	0.8323 [0.7939, 0.8696]
	unaffected	495	0.8301 [0.7849, 0.8707]
	affected	73 (43/30)	0.8349 [0.7227, 0.9328]

Brackets are 95% source-image-group bootstrap intervals (4,000 draws, seed 20260917), unadjusted. Shared edit assets recur across cohorts, so counts cannot be added. Full-cohort intervals come from this run's draws and can differ slightly from the separately computed intervals in Table 11 and Appendix D.

edits. The package covers every item in the six scored cohorts, but not every pre-admission edit, and single-layer MEGA lacks complete matching frozen assets for this check.

Affected proportions are $69/612$, $92/688$, $72/548$, $69/602$, $66/512$, and $73/568$ for Qwen2.5-VL-7B, Qwen2.5-VL-3B, LLaVA-OneVision-7B, Qwen3-VL-32B, Gemma-3-12B-it, and Qwen3.6-27B, respectively. Table 8 retains full, unaffected, and affected estimates. Removing affected edits changes point estimates in both directions, while discrimination remains above chance in each remaining subset. These are nested descriptive sensitivities, not replacement main results or tests of an AUROC difference.

The three-rater samples contain $17/90$ affected edits on 7B and $13/89$ on 3B. Among those affected edits, a majority judged 14 and 8 target-valid, leaving 3 and 5 invalid, respectively, so geometric affected status and human invalidity identify different sets. All images and results reported here come from this original editor.

C.7 Case Study

Figure 5 shows items that passed the Qwen2.5-VL-7B admission checks; item (c) gave a third-option response and lies outside the 612-item follow/preserve cohort. The selection rules were fixed before any item was retrieved: high and low mean the top and bottom quartiles of GLANCE in this cohort, cases (a), (b), and (d) come from edits for which a majority of three raters selected the target, and each case takes the first item under its rule. In pair (a), masking the region removes most of the follow item's support for its original answer (0.955 on the full view, 0.167 when masked), whereas the preserve item keeps its support (0.950 when masked). Pair (b) has almost the same visible support (full 0.945 and 0.956, crop 0.937 and 0.939) but masked support of 0.200 and 0.952, so the masked term sets the order of their scores. Item (c) scores in the 99th percentile, and after the edit the model gives neither the original answer nor the target. The score does not read the target (Section 3.2), and this item is one of the changed responses that the target-selection test in Section 4.6 evaluates. Pair (d) reverses the ranking: the preserve item scores in the 89th percentile with low masked support (0.256), and the follow item scores in the 1st percentile with masked support above its full-view support (0.728 against 0.611). Pairs like (d) are the discordant pairs that lower the AUROC. In the

Table 9: Hardware and representative wall-clock measurements. Each row is one measured job, not a standardized benchmark; total project compute is not reported.

Job	Scope	Hardware	Wall clock
Supplementary-pool edit construction (two SDXL edits and four answers per item)	625 + 634 items, two edit seeds, two models	4 × A800 80 GB (data sharding)	1 h 50 min
Qwen3-VL-32B admission responses (continuation run)	remaining 1,755 of 2,234 cached records	4 × A800 80 GB	about 2.5 h
Qwen3-VL-32B four-view letter scoring	per item, four candidates (16 forwards)	1 × A800 80 GB	about 180 s (about 11 s per forward pass)
Qwen3.6-27B replay: five readouts and four responses	96 items	1 × A800 80 GB	156.8 s per item (48.5–580.4 s, increasing with image tokens)
Gradient-attribution baseline	3,530 model–item records, six cohorts	1 × A800 80 GB	3 h 22 min

Each audited model ran on a single GPU for inference; multiple GPUs were used only to shard items. A GLANCE score needs $3K$ teacher-forced candidate evaluations for K candidates; edit generation, response checks, gradient back-propagation, and probe fitting are separate costs.

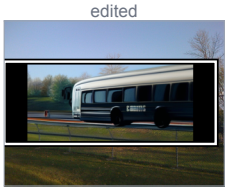
sign and sheep examples, content outside R still supports the original answer; these cases illustrate response ranking rather than confirmed failures to update.

C.8 Compute

Table 9 lists the hardware and the wall-clock measurements recorded in the run logs.

(a) Consistent with the ranking: a high-scoring follow and a low-scoring preserve





What is black?
answer: Train · target: Bus
full 0.955 crop 0.958 mask 0.167
GLANCE 0.789 (97th percentile)
edited answer: Bus **follow**





What does the orange sign say?
answer: Sidewalk Closed · target: Children Playing
full 0.850 crop 0.821 mask 0.950
GLANCE −0.114 (3rd percentile)
edited answer: Sidewalk Closed **preserve**

(b) Similar visible support, different masked support





What are the two people outside the shop holding?
answer: An umbrella · target: Cups of fresh coffee
full 0.945 crop 0.937 mask 0.200
GLANCE 0.741 (93rd percentile)
edited answer: Cups of fresh coffee **follow**





What is mostly white?
answer: Sheep · target: Seagulls
full 0.956 crop 0.939 mask 0.952
GLANCE −0.004 (8th percentile)
edited answer: Sheep **preserve**

(c) A high score whose edited answer is a third option



Who is with the bears?
answer: A woman · target: A wolf
full 0.916 crop, mask not stored
GLANCE 0.851 (99th percentile)
edited answer: A man **third option**

(d) Opposite to the ranking: a high-scoring preserve and a low-scoring follow





What is the woman player wearing?
answer: A white shirt · target: A tube top
full 0.952 crop 0.963 mask 0.256
GLANCE 0.701 (89th percentile)
edited answer: A white shirt **preserve**





What kind of meat is under the cheese?
answer: Bacon · target: Ham
full 0.611 crop 0.347 mask 0.728
GLANCE −0.249 (1st percentile)
edited answer: Ham **follow**

Figure 5: Rule-selected items from the Qwen2.5-VL-7B audit illustrate score interpretation, not pattern frequency. Selection rules were fixed before retrieval. Panels show (a) outcomes consistent with the ranking, (b) similar visible support but different masked support, (c) a third-option response outside the binary comparison, and (d) outcomes opposite to the ranking. Percentiles use the 612-item cohort; boxes mark R. A majority of the three raters selected the target for (a), (b), and (d). Item (c), shown on its edited image, has no human rating and stores only full-view support and GLANCE. The edited image in (a) contains original-editor black bars (Appendix C.6).

D COMPARATORS

Table 1 compares GLANCE with nine region-blind signals and with three families of region-aware readouts. This appendix defines both groups and reports the per-cohort details.

D.1 REGION-BLIND SIGNALS

The region-blind panel contains nine local signal slots that do not use the specified region R. Their names describe the implemented operations. Write $p_v(j)$ for the normalized candidate support in view v, $\ell_{\mathrm{full},a}$ for the original answer's mean token log score, and $\bar{p}_{\mathrm{blur}}$ for the average distribution over the Gaussian-blurred views. Table 10 fixes each signal's direction so that higher values predict follow.

Here $H_\epsilon(p) = -\sum_j p(j)\log(p(j)+\epsilon)$, with the implementation's fixed numerical stabilizer $\epsilon = 10^{-12}$. The blur entropy is computed after averaging candidate distributions. The self-verification prompt states the original answer text and scores Yes and No, whose normalized weights are not calibrated correctness probabilities. The swap uses another pool item selected by a fixed cyclic offset, without semantic matching or guaranteed distinct source images, so its per-item values depend on cohort size and order. The blank control still supplies a same-size gray image. None of these slots requires a crop or region mask, and their signs are retained from the frozen panel.

The panel size does not represent independent signals: max-softmax and original-answer support, for example, coincide on 547 of 548 LLaVA-OneVision rows. We retain both slots and the original score directions. Exact duplicate columns do not change the maximum AUROC over the panel, and nearly identical columns remain separate candidates.

For the best-of-panel comparison, the point difference is $\widehat{\Delta} = \widehat{\mathrm{AUROC}}(S) - \max_{b\in\mathcal{P}} \widehat{\mathrm{AUROC}}(b)$, where $\mathcal{P}$ is the retained panel. Paired bootstrap intervals use $B = 4{,}000$ draws and percentile 95% bounds, with the maximum over $\mathcal{P}$ recomputed inside every draw. The Qwen2.5-VL detection intervals and the MEGA single-layer analysis resample items, and the later audits additionally report source-image cluster intervals, preserving all items from each sampled image. Cross-model pools share source assets and admission depends on the audited model, so we report cell-specific estimates without treating the cells as independent replications.

D.2 PER-COHORT RESULTS AGAINST THE REGION-BLIND PANEL

Every analyzed item was answered correctly on the original image, elicited the target on the edited crop, and retained the original answer under the displaced placebo. All estimates use the letter-scoring convention and each audited model's own responses and scores. The six V7W cohorts span AUROC 0.7068–0.8323, with positive within-cohort panel margins (Table 11). LLaVA-OneVision extends the comparison to a VLM line using a different vision encoder and connector, while retaining a Qwen-family language backbone (Qwen2-7B) (Li et al., 2025), and Gemma supplies the non-Qwen language-backbone test. Model-specific admission also changes the candidates available to the whole-image swap comparator, and margins compare signals within each conditional population.

Table 10: The nine region-blind comparison signals and their implemented operations. Each signal is oriented so that a higher value predicts follow; none reads the designated region R.

Display name	Implemented score
max-softmax	$\max_j p_{\mathrm{full}}(j)$
negative option entropy	$-H_\epsilon(p_{\mathrm{full}})$
answer mean logprob	$\ell_{\mathrm{full},a}$
blur-perturb (−entropy)	$-H_\epsilon(\bar{p}_{\mathrm{blur}})$
blur-perturb (−disagreement)	Negative fraction of blurred-view argmaxes differing from the full-view argmax
self-verify (Yes weight)	Normalized Yes weight from one Yes/No verification prompt
whole-image swap ($1 - g_{\mathrm{swap}}$)	$1 - p_{\mathrm{swap}}(a)$
original-answer support (g_{full})	$p_{\mathrm{full}}(a)$
blank-image ($1 - g_{\mathrm{blank}}$)	$1 - p_{\mathrm{blank}}(a)$

Table 11: Follow-versus-preserve AUROC on the six V7W admission cohorts against the strongest of the nine region-blind signals. Every cohort has a positive margin with an interval above zero.

Audited model	n	GLANCE AUROC [95% CI]	Strongest region-blind slot	Slot AUROC	Margin [95% CI]
Qwen2.5-VL-7B	612	0.812 [0.772, 0.848]	whole-image swap	0.549	+0.263 [+0.195, +0.302]
Qwen2.5-VL-3B	688	0.798 [0.762, 0.834]	option entropy	0.547	+0.251 [+0.188, +0.284]
LLaVA-OneVision-7B	548	0.797 [0.749, 0.840]	whole-image swap	0.513	+0.284 [+0.210, +0.328]
Qwen3-VL-32B	602	0.789 [0.743, 0.833]	self-verify	0.533	+0.256 [+0.174, +0.302]
Gemma-3-12B-it	512	0.707 [0.658, 0.754]	option entropy	0.520	+0.187 [+0.119, +0.242]
Qwen3.6-27B	568	0.832 [0.791, 0.870]	option entropy	0.546	+0.287 [+0.223, +0.336]

Brackets are 95% item-bootstrap intervals. The strongest slot is the point-estimate winner within each cohort, and margin intervals re-select the strongest of the nine slots inside every resample. Slot names abbreviate Table 10: whole-image swap is $1 - g_{\mathrm{swap}}$, option entropy is $-H_\epsilon(p_{\mathrm{full}})$, and self-verify is the normalized Yes weight of one verification prompt. Max-softmax and original-answer support (g_{full}) largely coincide in the audited caches, so the slot count does not represent independent signals. Intervals are unadjusted across models.

Table 12: Detection on the additional data pools, each with its own model-specific admission and letter scoring. Paired margins over the retained nine-slot region-blind panel have intervals above zero on all four pools.

Data pool	Audited model	n	GLANCE AUROC	Margin [95% CI]
GQA-derived MCQs	Qwen2.5-VL-7B	295	0.838	+0.291 [+0.218, +0.339]
GQA-derived MCQs	Qwen2.5-VL-3B	245	0.717	+0.182 [+0.080, +0.236]
MEGA single-layer subset	Qwen2.5-VL-7B	404	0.774	+0.267 [+0.186, +0.298]
MEGA single-layer subset	Qwen2.5-VL-3B	234	0.745	+0.215 [+0.119, +0.268]

AUROC is the point estimate; brackets are 95% paired item-bootstrap intervals with the strongest of the nine slots re-selected inside each draw, unadjusted for multiplicity. Max-softmax and original-answer support (g_{full}) largely coincide in the audited caches; the overlap audit covers cached pools that include pools not analyzed here, and subset-specific overlap rates are unavailable.

The Qwen3-VL-32B cohort obtains AUROC 0.7890 $[0.7428, 0.8328]$ with panel margin $+0.2556$ $[+0.1744, +0.3024]$, and its source-image cluster margin interval is $[+0.1764, +0.3046]$. Self-verify (Yes weight) is its strongest slot at 0.533. On Qwen3.6-27B, GLANCE reaches AUROC 0.832, while the local panel spans 0.462–0.546, with negative option entropy as its strongest slot. The item-bootstrap margin is $+0.287$ $[+0.223, +0.336]$, and grouping by source image gives a margin interval of $[+0.220, +0.338]$ and a GLANCE AUROC interval of $[0.790, 0.871]$. Gemma-3-12B-it is the weakest model by both AUROC and margin. Its AUROC is 0.7068 $[0.6583, 0.7542]$, or $[0.6574, 0.7548]$ under source-image clustering over 503 groups, and its margin is $+0.1869$ $[+0.1185, +0.2416]$, or $[+0.1195, +0.2416]$ when clustered. Its nine slots span 0.4317–0.5200, with negative option entropy highest. Encoder, language component, training, and population change together, so the lower magnitude cannot be isolated to its language backbone.

The derived-region analyses extend detection to the GQA construction and the MEGA single-layer subsets, with positive paired margins over the retained panel whose intervals exclude zero (Table 12).

D.3 Region-Aware Comparison Families

The strongest of the nine region-blind signals reaches AUROC 0.51–0.55, so exceeding them does not show that GLANCE improves on alternatives that also read the region R. We therefore compare GLANCE with three pre-registered families of region-aware readouts on the six Table 1 cohorts, using identical item sets, labels, and GLANCE scores. Each family's primary endpoint is $\Delta_{\mathrm{fam}} = \mathrm{AUROC}(\mathrm{GLANCE}) - \max_s \mathrm{AUROC}(s)$ over its slots, with the strongest slot re-selected inside each of $4{,}000$ source-image-group bootstrap draws and 95% percentile intervals unadjusted across models. Plans, code, and inputs were frozen before full computation, and each plan committed to reporting the endpoint regardless of direction.

Families. *Attention on the region* reads, at the position preceding the answer letter, the softmax attention over image tokens and aggregates the fraction falling in R as a share, an area-normalized

Table 13: Three pre-registered region-aware comparison families on the six V7W admission cohorts. Each row gives, for one cohort and family, the strongest slot at the point estimate, its AUROC, and the margin $\Delta_{\mathrm{fam}} = \mathrm{AUROC}(\mathrm{GLANCE}) - \max_s \mathrm{AUROC}(s)$, with the strongest slot re-selected in each of $4{,}000$ source-image-group bootstrap draws (95% percentile intervals). GLANCE exceeds the attention and gradient-attribution families in every cohort and the geometry family in five; on Gemma-3-12B-it the geometry interval includes zero. The joint attention ∪ geometry rows (18 slots) are descriptive: a geometry slot is always the strongest, so their point estimates coincide with the geometry rows.

Audited model	Family (slots)	Strongest slot (orientation)	AUROC	Δ_{fam} [95% CI]
Qwen2.5-VL-7B	attention (12)	attn. share, all layers (+)	0.586	+0.226 [+0.166, +0.273]
	gradient attr. (12)	$\lvert g \cdot x \rvert$ mass (+)	0.599	+0.213 [+0.159, +0.263]
	geometry (6)	centrality (+)	0.645	+0.167 [+0.110, +0.214]
	attn. ∪ geometry (18)	centrality (+)	0.645	+0.167 [+0.110, +0.212]
Qwen2.5-VL-3B	attention (12)	attn. share, all layers (+)	0.622	+0.176 [+0.117, +0.231]
	gradient attr. (12)	$\lvert g \cdot x \rvert$ mass (+)	0.628	+0.170 [+0.112, +0.222]
	geometry (6)	centrality (+)	0.694	+0.105 [+0.051, +0.159]
	attn. ∪ geometry (18)	centrality (+)	0.694	+0.105 [+0.051, +0.159]
LLaVA-OneVision-7B	attention (12)	attn. density, mid layers (−)	0.542	+0.255 [+0.171, +0.299]
	gradient attr. (12)	$\lvert g \cdot x \rvert$ mass (+)	0.542	+0.255 [+0.181, +0.301]
	geometry (6)	centrality (+)	0.669	+0.128 [+0.066, +0.189]
	attn. ∪ geometry (18)	centrality (+)	0.669	+0.128 [+0.066, +0.189]
Qwen3-VL-32B	attention (12)	attn. share, all layers (+)	0.562	+0.227 [+0.150, +0.268]
	gradient attr. (12)†	$\lvert g \cdot x \rvert$ mass (+)	0.577	+0.211 [+0.143, +0.260]
	geometry (6)	centrality (+)	0.657	+0.133 [+0.073, +0.193]
	attn. ∪ geometry (18)	centrality (+)	0.657	+0.133 [+0.073, +0.192]
Gemma-3-12B-it	attention (12)	attn. density, mid layers (−)	0.625	+0.082 [+0.011, +0.144]
	gradient attr. (12)	$\lvert g \cdot x \rvert$ mass (+)	0.600	+0.107 [+0.033, +0.168]
	geometry (6)	centrality (+)	0.674	+0.033 [−0.032, +0.096]
	attn. ∪ geometry (18)	centrality (+)	0.674	+0.033 [−0.032, +0.096]
Qwen3.6-27B	attention (12)	attn. share, all layers (+)	0.601	+0.231 [+0.164, +0.287]
	gradient attr. (12)†	$\lvert g \cdot x \rvert$ mass (+)	0.568	+0.260 [+0.179, +0.299]
	geometry (6)	aspect ratio (+)	0.644	+0.189 [+0.126, +0.222]
	attn. ∪ geometry (18)	aspect ratio (+)	0.644	+0.189 [+0.126, +0.219]

Orientation $(+)$: a higher readout predicts follow (attn. abbreviates attention); $(-)$: a higher readout predicts preserve. †Gradient attribution covers 600/602 (Qwen3-VL-32B) and 519/568 (Qwen3.6-27B) items after backward-pass out-of-memory failures, and the missing items resemble the retained ones in label rate and GLANCE (Appendix D.3). Intervals are unadjusted for multiplicity.

density, and a sink-robust percentile rank, over all layers or the middle third, in both orientations (12 slots). The capture leaves the forward pass bit-identical to scoring. *Gradient attribution on the region* back-propagates the answer-span mean log-probability, the same quantity GLANCE scores, to the first text-decoder layer's input at image-token positions and aggregates $\mathrm{ReLU}(\sum g \cdot x)$, $\sum \lvert g \cdot x \rvert$, and $\lVert g \rVert_2$ in R as mass and area-normalized enrichment, in both orientations (12 slots). *Region geometry* reads only the box and image size: area fraction, centrality (one minus the normalized distance of the box center from the image center), and aspect ratio, in both orientations (6 slots).

Results. Table 13 gives the primary endpoints. GLANCE exceeds the attention family in all six cohorts ($+0.08$ to $+0.25$) and the gradient-attribution family in all six ($+0.11$ to $+0.26$), and the strongest gradient slot is the absolute input-gradient mass in every cohort. Geometry is the strongest comparator: centrality alone reaches 0.64–0.69, GLANCE exceeds the family in five cohorts ($+0.10$ to $+0.19$), and on Gemma-3-12B-it the interval includes zero ($+0.033$ $[-0.032, +0.096]$). Taking the maximum over attention and geometry jointly leaves every point estimate and selected slot unchanged, because a geometry readout is always the strongest at the point estimate, and the same holds when gradient attribution is added on the items where it is available. The joint intervals differ slightly because the strongest slot is re-selected in every draw. The margins in Table 1 are therefore GLANCE's smallest advantage over any of the region-aware readouts we constructed.

Information beyond GLANCE. Matched logistic probes (Section 4.3) test whether each family's strongest readout adds out-of-fold AUROC to GLANCE, with a pre-declared noise floor of 0.007. Attention adds at most 0.004 in magnitude in every cohort. Gradient attribution stays within the floor in four cohorts and lands on opposite sides in two (-0.008 on Qwen3-VL-32B, $+0.012$ on Gemma-3-12B-it). Geometry differs: its increments are positive in all six cohorts, exceed the floor in four, and have intervals above zero on Qwen2.5-VL-3B ($+0.018$ $[+0.002, +0.034]$) and Gemma-3-12B-it ($+0.042$ $[+0.011, +0.073]$). Region geometry carries a small amount of information that GLANCE does not, most on the cohort where the two are not separated.

Matching on geometry. Because geometry both predicts the label and is correlated with the admission gate, we ask how much of GLANCE's advantage survives within geometry strata. Items are stratified by centrality, area fraction, and aspect ratio (terciles or quartiles, two- and three-way), and GLANCE and each feature are compared only between a follow and a preserve item of the same stratum. A configuration is reportable only if its positive control passes: the matched geometry feature must move toward 0.5. Table 14 reports the primary three-way tercile configuration. In 21 of 24 configurations the positive control passes, and all 63 resulting margins have intervals above zero. On Gemma-3-12B-it the matched margin over centrality is $+0.101$ $[+0.013, +0.192]$ in the primary configuration and $+0.10$ to $+0.15$ across all four configurations: geometry alone approaches GLANCE on that cohort, but within geometry strata GLANCE still separates follow from preserve. The three failing configurations involve features whose raw AUROC already sits near 0.5 (area on LLaVA-OneVision-7B at 0.534 and aspect on Qwen3-VL-32B in one quartile configuration at 0.585), where the binary control is sensitive to sampling noise, and their margins are not reported.

Scope and disclosure. The attention family was pre-registered on 2026-09-19, with endpoint and code frozen before any full-cohort computation, and its results did not inform any other analysis in this paper. The gradient-attribution family covers 600 of 602 Qwen3-VL-32B items and 519 of 568 Qwen3.6-27B items after backward-pass out-of-memory failures on long sequences. The missing items resemble the retained ones in follow rate (0.714 versus 0.748 on 27B) and are not distinguishable from them by GLANCE (AUROC 0.540). Gradient-derived quantities are reproducible only to about 10^{-4} on this hardware because the bf16 backward pass is not bit-deterministic. Forward quantities are exact. The region-aware margins in Table 1 are recomputed from the archived run records of the three families. The replay archive contains the per-item region-aware readouts and recomputes these margins, the three family tables, and the geometry-matching table, checking each against the archived records. The three families and the matching analysis share the frozen items, labels, and scores of the main cohorts, so they are not independent replications.

Table 14: GLANCE after matching on region geometry (centrality × area × aspect ratio, terciles) on the six V7W admission cohorts. Comparisons are pooled within-stratum concordances between one follow and one preserve item of the same stratum; each margin subtracts the feature's matched concordance from GLANCE's, with 95% item-bootstrap intervals (4,000 draws). Every reported margin has an interval above zero, so GLANCE separates follow from preserve within geometry strata, including on Gemma-3-12B-it, where geometry alone approaches GLANCE. A configuration whose positive control fails (the matched geometry feature does not move toward 0.5) reports no margin.

Audited model (GLANCE raw→matched; pairs)	Control	Feature	Margin [95% CI]
Qwen2.5-VL-7B (0.812→0.781; 2795)	pass	area	+0.297 [+0.211, +0.381]
		centrality	+0.233 [+0.151, +0.312]
		aspect ratio	+0.238 [+0.152, +0.319]
Qwen2.5-VL-3B (0.798→0.760; 3611)	pass	area	+0.261 [+0.185, +0.335]
		centrality	+0.179 [+0.099, +0.261]
		aspect ratio	+0.174 [+0.092, +0.252]
LLaVA-OneVision-7B (0.797→0.767; 2063)	fail (area)	—	not reported
Qwen3-VL-32B (0.789→0.771; 2608)	pass	area	+0.254 [+0.161, +0.343]
		centrality	+0.194 [+0.110, +0.276]
		aspect ratio	+0.208 [+0.124, +0.294]
Gemma-3-12B-it (0.707→0.695; 1980)	pass	area	+0.178 [+0.087, +0.271]
		centrality	+0.101 [+0.013, +0.192]
		aspect ratio	+0.136 [+0.045, +0.226]
Qwen3.6-27B (0.832→0.810; 2206)	pass	area	+0.324 [+0.241, +0.407]
		centrality	+0.215 [+0.131, +0.296]
		aspect ratio	+0.237 [+0.153, +0.317]
LLaVA-OneVision-7B, configurations whose positive control passes			
two-way, quartiles (0.797→0.789; 3637)	pass	area	+0.257 [+0.168, +0.339]
		centrality	+0.241 [+0.158, +0.323]
		aspect ratio	+0.217 [+0.137, +0.300]
three-way, quartiles (0.797→0.780; 933)	pass	area	+0.251 [+0.141, +0.360]
		centrality	+0.269 [+0.167, +0.366]
		aspect ratio	+0.211 [+0.111, +0.311]

Across the 21 of 24 configurations (six cohorts × two stratifications × terciles or quartiles) whose positive control passes, all 63 margins have intervals above zero. On LLaVA-OneVision-7B the tercile configurations fail the positive control because area's raw AUROC (0.534) already sits near chance and its matched value crosses to 0.445; the quartile configurations pass. Matched concordance is not an AUROC and is not comparable to Table 1.

E MASKED-SUPPORT EVIDENCE

This appendix gives the details behind Section 4.3, including its view and coefficient analyses.

E.1 MATCHED PROBES

We distinguish access to a view from the formula used to combine the views by fitting matched logistic-regression (LR) probes. LR-visible receives $(g_{\mathrm{crop}}, g_{\mathrm{full}})$, LR-mask receives g_{mask}, and LR-all3 receives all three supports from the unedited original image. The probes share image-grouped five-fold splits across 20 repeats, training-fold standardization, and fixed regularization. The archived solver uses full-batch gradient descent on mean logistic loss plus $\frac{0.01}{2}\|w\|_2^2$, with an unpenalized intercept, zero initialization, and step size 0.5. It stops when the gradient infinity norm falls below 10^{-7}, with a maximum of $10{,}000$ steps. Writing repeat-mean out-of-fold AUROC as A, as in Section 3, we define $\Delta_{\mathrm{visible}} = A_{\mathrm{all3}} - A_{\mathrm{visible}}$, the increment from adding masked support to the visible supports, and $\Delta_{\mathrm{mask}} = A_{\mathrm{all3}} - A_{\mathrm{mask}}$, the reverse addition. These supervised probes diagnose predictive access to the inputs within the admitted follow/preserve task, and the fixed GLANCE score is not fitted to response labels from the evaluation cohorts.

Under fixed out-of-fold intervals, both additions improve the matched probes in all eight cohorts (Table 15), and the MEGA probes are refitted on the single-layer question subsets. On Qwen3-VL-32B, LR-visible, LR-mask, and LR-all3 attain 0.5495, 0.7455, and 0.7763, with increments $+0.2267$ $[+0.1721, +0.2830]$ and $+0.0308$ $[+0.0105, +0.0518]$. On Gemma-3-12B-it, the corresponding AUROCs are 0.5665, 0.6599, and 0.7003, with increments $+0.1338$ $[+0.0820, +0.1860]$ and $+0.0404$ $[+0.0103, +0.0723]$. On Qwen3.6-27B in non-thinking mode, the three probes reach 0.585, 0.778, and 0.828, with $\Delta_{\mathrm{visible}} = +0.243$ $[+0.184, +0.301]$ and $\Delta_{\mathrm{mask}} = +0.050$ $[+0.023, +0.078]$.

The intervals in Table 15 resample source-image groups while holding fitted models, predictions, and split assignments fixed, so they exclude variability from refitting the learning procedure. Figure 3a reports refitting-sensitivity intervals for the same eight cohorts; these intervals are obtained by repeating splitting and fitting within each of $1{,}000$ source-image-group resamples. Under refitting, all eight masked-input increments and seven of the eight reverse increments keep intervals above zero, the exception being the single-layer MEGA 3B reverse increment at $[-0.009, +0.091]$, and Appendix G.2 reports the same analysis for the three-rater subsets.

Reference input with geometry and blank-image support. A further pre-registered probe asks whether masked support still adds information when the reference input already carries region geometry and an image-free prior. The reference probe receives the two visible supports, the three geometry features of Appendix D.3 (area fraction, centrality, and aspect ratio), and blank-image support g_{blank}, the original-answer support on a same-size gray image. The extended probe adds g_{mask}, and both use the matched-probe protocol above. A pre-registered replication check required the three original probes to reproduce their registered values on the full probe lists, and five cohorts passed. LLaVA-OneVision-7B and the two single-layer MEGA cohorts (the latter added post hoc) failed only on the interval endpoints: their fitted values and point estimates matched, but their archived intervals were drawn from random streams shared with earlier cohorts and differ from the recomputed ones by up to 2.2×10^{-3}. Before any extended probe was computed for these three cohorts, we amended the check to restore the recorded stream positions, under which all three reproduce their archived values and intervals exactly, and then fitted their extended probes with the same protocol. In the five cohorts that passed the original check, the reference probe already exceeds the visible-only probe by 0.071–0.139 AUROC, and adding masked support raises it by a further 0.049–0.138, with every interval above zero (Table 16). In the three cohorts that passed the amended check, the corresponding increments are 0.009–0.103 and 0.108–0.146, the latter again with every interval above zero. The extended probe exceeds the all-view probe by 0.001–0.054 in the five cohorts, with intervals above zero only on Qwen2.5-VL-3B and Gemma-3-12B-it, the two cohorts where geometry also adds to GLANCE (Appendix D.3), and is not separated from it in the other three. These intervals hold the fitted predictions fixed and exclude refitting; g_{blank} is an operational control for an image-free prior, and the geometry block has three features.

Table 15: Predictive increments from adding one view to matched logistic probes on the eight admission cohorts. Adding masked support to the two visible views raises out-of-fold AUROC by 0.124 to 0.243 with every interval above zero; the reverse increment, adding the visible views to masked support, is 0.023 to 0.064. Brackets are 95% fixed-OOF intervals over source-image groups and exclude refitting variability.

Cohort	n	Add masked to visible [95% CI]	Add visible to masked [95% CI]
Qwen2.5-VL-7B	611	+0.178 [+0.124, +0.232]	+0.039 [+0.017, +0.062]
Qwen2.5-VL-3B	687	+0.139 [+0.092, +0.186]	+0.064 [+0.035, +0.092]
LLaVA-OneVision-7B	547	+0.194 [+0.133, +0.252]	+0.032 [+0.006, +0.060]
Qwen3-VL-32B	602	+0.227 [+0.172, +0.283]	+0.031 [+0.011, +0.052]
Gemma-3-12B-it	512	+0.134 [+0.082, +0.186]	+0.040 [+0.010, +0.072]
Qwen3.6-27B	568	+0.243 [+0.184, +0.301]	+0.050 [+0.023, +0.078]
Qwen2.5-VL-7B (MEGA SL)	404	+0.191 [+0.133, +0.245]	+0.023 [+0.003, +0.044]
Qwen2.5-VL-3B (MEGA SL)	234	+0.124 [+0.057, +0.194]	+0.041 [+0.003, +0.081]

Each increment is the mean over 20 grouped five-fold repeats of the OOF-AUROC difference between the all-view probe and its corresponding reduced-input probe, using identical splits. $4{,}000$ paired resamples of source-image groups condition on the fitted OOF predictions. The six V7W rows use the frozen probe lists (Appendix C.4); MEGA SL denotes the post-hoc single-layer subsets refit with Qwen2.5-VL.

Table 16: Masked support beyond a reference input of visible supports, region geometry, and blank-image support. Adding masked support raises out-of-fold AUROC by 0.049 to 0.138 in the five cohorts that passed the pre-registered replication check and by 0.108 to 0.146 in the three that passed the amended check, with every interval above zero.

Cohort	n	Reference	Reference + mask	Add masked [95% CI]	Reference + mask − all three [95% CI]
Qwen2.5-VL-7B	611	0.703	0.812	+0.109 [+0.069, +0.149]	+0.001 [−0.009, +0.012]
Qwen2.5-VL-3B	687	0.734	0.821	+0.087 [+0.054, +0.120]	+0.020 [+0.004, +0.037]
Qwen3-VL-32B	602	0.659	0.784	+0.125 [+0.081, +0.171]	+0.008 [−0.012, +0.028]
Gemma-3-12B-it	512	0.705	0.755	+0.049 [+0.019, +0.079]	+0.054 [+0.020, +0.088]
Qwen3.6-27B	568	0.700	0.838	+0.138 [+0.095, +0.181]	+0.010 [−0.005, +0.026]
Amended replication check					
LLaVA-OneVision-7B	547	0.692	0.805	+0.113 [+0.070, +0.156]	+0.022 [−0.001, +0.045]
Qwen2.5-VL-7B (MEGA SL)	404	0.642	0.788	+0.146 [+0.097, +0.196]	+0.007 [−0.011, +0.025]
Qwen2.5-VL-3B (MEGA SL)	234	0.615	0.723	+0.108 [+0.048, +0.169]	−0.007 [−0.032, +0.019]

The reference probe receives g_{crop}, g_{full}, area fraction, centrality, aspect ratio, and g_{blank}; the extended probe adds g_{mask}. Entries are repeat-mean OOF AUROCs under the matched-probe protocol of Table 15, and brackets are unadjusted 95% fixed-OOF intervals from $4{,}000$ source-image-group resamples that exclude refitting. Cohorts share edit assets. LLaVA-OneVision-7B and the two single-layer MEGA cohorts (the latter run post hoc) failed the pre-registered check only on archived interval endpoints and passed the amended check of Appendix E.1, which was decided before their extended probes were fitted. Qwen3.6-27B runs in non-thinking mode.

Table 17: Scoring masks restricted to disjoint, equal-area pixel footprints: GLANCE with the designated region (S_R) against its displaced-mask counterpart (S_W) on the restricted pools. All six paired differences favor the designated region with intervals above zero, while S_W itself stays at 0.576 to 0.612, so displacement lowers discrimination without removing it.

Audited model	n	S_R	S_W	Paired difference [95% CI]
Qwen2.5-VL-7B	459	0.807	0.609	+0.198 [+0.134, +0.265]
Qwen2.5-VL-3B	529	0.791	0.608	+0.183 [+0.120, +0.244]
LLaVA-OneVision-7B	404	0.800	0.586	+0.214 [+0.136, +0.294]
Qwen3-VL-32B	458	0.800	0.576	+0.224 [+0.162, +0.286]
Gemma-3-12B-it	387	0.701	0.612	+0.089 [+0.012, +0.165]
Qwen3.6-27B	428	0.833	0.579	+0.254 [+0.186, +0.324]

Absolute columns report AUROC. Intervals use $4{,}000$ paired source-image-group resamples, unadjusted. W coordinates are cached and reconstruction-checked for Qwen3.6-27B; the other models use the archived placement code and original metadata. These are restricted pools, distinct from the full admission cohorts.

Table 18: Input and readout comparisons after restricting placebo geometry to strictly disjoint boxes, on the common probe-eligible, strictly parsed items. Refitted probes retain masked-input increments with all six fixed-OOF intervals above zero, and GLANCE exceeds pixel-FPVG in five cohorts, with the Gemma-3-12B-it interval including zero.

Audited model	n	All three − visible [95% CI]	GLANCE − pixel-FPVG [95% CI]
Qwen2.5-VL-7B	457	+0.193 [+0.133, +0.254]	+0.144 [+0.109, +0.178]
Qwen2.5-VL-3B	534	+0.157 [+0.104, +0.211]	+0.149 [+0.111, +0.186]
LLaVA-OneVision-7B	407	+0.200 [+0.133, +0.266]	+0.151 [+0.117, +0.189]
Qwen3-VL-32B	460	+0.256 [+0.197, +0.316]	+0.163 [+0.131, +0.195]
Gemma-3-12B-it	383	+0.107 [+0.051, +0.162]	+0.031 [−0.011, +0.074]
Qwen3.6-27B	433	+0.251 [+0.188, +0.316]	+0.172 [+0.135, +0.209]

Each row uses the common P-disjoint, probe-eligible, strictly parsed set. Probes are refitted within this set with the original learner and splits, and their intervals condition on the newly fitted OOF predictions. All intervals use $4{,}000$ paired source-image-group draws, unadjusted.

E.2 Spatial Controls for the Masked View

The scoring mask W and admission-placebo edit P use different operations, but their shared placement heuristic can return boxes intersecting the designated region R. We write S_R for GLANCE and S_W for its displaced-mask counterpart. For the cohorts 7B, 3B, LLaVA, 32B, Gemma, and 27B, reconstructing raw boxes, rounding, patch dimensions, and clipping identifies $147/612$, $153/688$, $140/548$, $140/602$, $121/512$, and $135/568$ scoring-mask overlaps, and pixel areas differ in 5, 6, 3, 4, 4, and 5 items. All 27B reconstructed boxes match its archived raw coordinates, while the other cohorts rely on the archived implementation and original metadata.

Restricting W to disjoint, positive, exactly area-matched pixel footprints and retaining the one-image source exclusion (Appendix B.6) gives Table 17. All six paired intervals favor the designated region. This supports location sensitivity under the restricted control, and mask-induced distribution shift and background dependence remain possible contributors. The restricted subsets differ from the full cohorts, so their estimates are not effects of repairing the control.

For P, we intersect the disjoint-placebo subset with the probe and strict greedy-response manifests. Table 18 uses identical items within each row. Refitted probes retain masked-input increments of 0.107–0.256, with all six fixed-OOF intervals positive, and reverse increments are 0.026–0.057, also with positive intervals. GLANCE exceeds pixel-FPVG in five subsets, and the Gemma interval includes zero. The common-item and W-restricted populations differ.

The probes are refitted within each restricted population using the original optimizer, regularization, training-fold standardization, and twenty grouped five-fold splits. All $1{,}800$ fits converge. Source-image-group bootstrap holds these newly fitted OOF predictions fixed, excluding refitting and annotation uncertainty. The `upgrade/geometry/` folder of the reproducibility archive includes every fixed-score absolute AUROC and contrast, all three probe AUROCs, folds, OOF predictions,

Table 19: Score-family decomposition on the Qwen3-VL-32B and Gemma-3-12B-it cohorts. Differences are canonical GLANCE minus the row score with 95% paired item-bootstrap intervals. Removing the masked view (crop−full, crop, full) costs 0.10 to 0.27 AUROC on both models, whereas crop−mask is not separated from GLANCE on either.

	Qwen3-VL-32B ($n = 602$)		Gemma-3-12B-it ($n = 512$)	
Score	AUROC	Canonical − row [95% CI]	AUROC	Canonical − row [95% CI]
Canonical GLANCE	0.7890	—	0.7068	—
crop−mask	0.7818	+0.0072 [−0.0058, +0.0194]	0.7176	−0.0107 [−0.0286, +0.0066]
full−mask	0.7783	+0.0107 [−0.0012, +0.0234]	0.6757	+0.0311 [+0.0094, +0.0530]
−mask	0.7552	+0.0339 [+0.0153, +0.0535]	0.6722	+0.0346 [+0.0057, +0.0636]
crop−full	0.6099	+0.1791 [+0.1152, +0.2428]	0.5626	+0.1443 [+0.0766, +0.2119]
crop (g_{crop})	0.6142	+0.1749 [+0.1088, +0.2433]	0.6056	+0.1013 [+0.0404, +0.1622]
full (g_{full})	0.5179	+0.2712 [+0.1959, +0.3457]	0.5169	+0.1900 [+0.1260, +0.2568]
Displaced mask (S_W)	0.6011	+0.1880 [+0.1311, +0.2475]	0.5947	+0.1122 [+0.0450, +0.1760]

Letter scoring on 602 and 512 determinate items. Differences subtract the row score from canonical before rounding; intervals use 4,000 paired item resamples (seed 0), unadjusted.

geometry, and source exclusions. These post-hoc analyses use plans frozen before the restricted estimates were computed.

E.3 Score Family and Coefficients

Among the same-region contrasts, crop−mask and GLANCE are not separated in any of the six V7W cohorts (Table 1, Panel B). For the Qwen3-VL-32B and Gemma cohorts, Table 19 decomposes the score family: the AUROC ranges of the formulas that mask the designated region and of the visible-only formulas stay disjoint; this summarizes point estimates, not paired uncertainty. Point-estimate differences change sign across cohorts, and no family member consistently attains the highest AUROC. What the high-AUROC members share is access to masked-view support.

We examine coefficient sensitivity within $s_{\alpha,\beta} = \alpha g_{\text{crop}} + (1-\alpha) g_{\text{full}} - \beta g_{\text{mask}}$, whose canonical setting is $(0.5, 1.0)$. The grid spans $\alpha \in [0, 1]$ and $\beta \in [0, 2]$ in steps of 0.05, and a plateau is defined in advance as AUROC at least canonical AUROC minus 0.01. Across the eight admission cohorts, the plateau occupies 43.6%–59.1% of grid points, and its projected ranges reach $\alpha = 1.0$ and $\beta = 2.0$ in every cohort, meaning that each endpoint works with some value of the other coefficient. The projections are not a Cartesian set: a pair inside both ranges can still fall outside the plateau.

The AUROC loss from removing the mask term exceeds the gain from optimizing the coefficients (Figure 6). At $\beta = 0$, even the best visible-only convex mixture on the grid falls below canonical GLANCE by 0.101–0.234 AUROC, whereas the within-sample grid maximum exceeds canonical by only 0.0007–0.0165. The former is an ablation that also searches the remaining visible coefficient, whereas the latter is a descriptive maximum selected on the evaluated data, without confidence intervals. Positive mask weight alone is not sufficient: the lowest projected plateau mask weights range from 0.35 to 0.90. The search reaches $\beta = 2.0$, and the MEGA-7B single-layer optimum lies near that limit at $\beta = 1.95$, leaving behavior beyond the grid unresolved. On Qwen3-VL-32B and Gemma, the maxima occur at $(\alpha, \beta) = (0.40, 0.90)$ and $(0.95, 0.95)$. Across the matched probes and coefficient analyses, masked-view support contributes consistently, while the gains from visible supports and weight tuning vary with the model and population.

E.4 Readout Scale

Applying the same three-view contrast to log or logit supports instead of probability supports changes AUROC by -0.008 to $+0.007$ in the six V7W cohorts (Table 20). Every logit-version interval and five of the six log-version intervals include zero, and this post-hoc comparison has no prespecified equivalence margin.

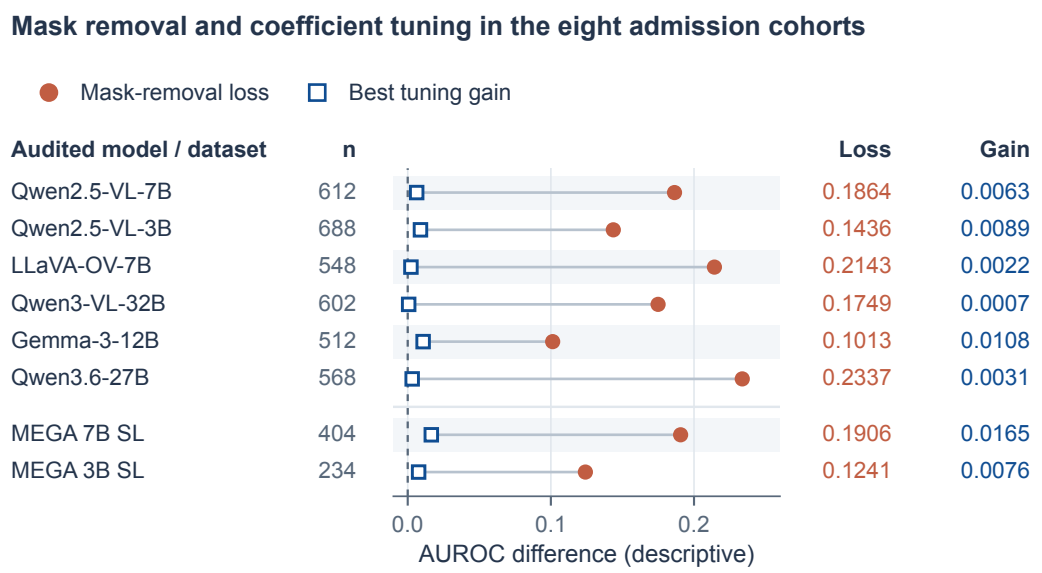


Figure 6: Mask-removal loss exceeds the observed tuning gain within each admission cohort. Orange points give canonical AUROC minus the best visible-only mixture at $\beta = 0$, and blue squares give the same cohort's full-grid maximum minus canonical. These are descriptive values selected on the evaluated data, without confidence intervals. The rows include Qwen3-VL-32B, Gemma-3-12B-it, and the single-layer MEGA subsets, and Qwen3.6-27B uses non-thinking mode.

Table 20: Readout scale on the six V7W admission cohorts (post hoc). Applying the same three-view contrast to log or logit supports instead of probability supports changes AUROC by -0.008 to $+0.007$. Every logit-version interval includes zero, and five of the six log-version intervals do.

Audited model	GLANCE	Log version	Logit version	Log − GLANCE [95% CI]
Qwen2.5-VL-7B	0.8119	0.8072	0.8066	−0.0046 [−0.0155, +0.0062]
Qwen2.5-VL-3B	0.7982	0.7971	0.7977	−0.0011 [−0.0127, +0.0101]
LLaVA-OneVision-7B	0.7970	0.7998	0.8012	+0.0029 [−0.0036, +0.0098]
Qwen3-VL-32B	0.7890	0.7808	0.7957	−0.0082 [−0.0168, −0.0007]
Gemma-3-12B-it	0.7068	0.6996	0.7140	−0.0073 [−0.0224, +0.0070]
Qwen3.6-27B	0.8323	0.8317	0.8283	−0.0006 [−0.0116, +0.0099]

The log version is $\frac{1}{2}(\log g_{\text{crop}} + \log g_{\text{full}}) - \log g_{\text{mask}}$, and the logit version applies the same contrast to $\text{logit}\, g_v$ clipped to $[10^{-4}, 1 - 10^{-4}]$, a bound no item reaches. Logit-minus-GLANCE differences range from -0.0053 to $+0.0071$. Intervals are paired item bootstraps (4,000 draws, seed 0), unadjusted for multiplicity, and the Qwen3-VL-32B log interval narrowly excludes zero.

F Continuous Versus Binary Readouts

This appendix gives the details behind Section 4.4.

F.1 Strictly Parsed Comparisons

We compare continuous support with discrete answer agreement on the same original-image views. Let F, C, M denote greedy answers from the unedited full image, crop, and masked image. The local pixel-FPVG adaptation uses $\mathbf{1}[F = C \wedge F \neq M]$, while masked-greedy uses $\mathbf{1}[F \neq M]$. This pixel-level adaptation of FPVG's agreement rule replaces the original object-feature deletion intervention with crop and mask inputs (Reich et al., 2023). All scores are evaluated on the same strictly parsed subset, $600/611$ fixed items for Qwen2.5-VL-7B and $687/687$ for Qwen2.5-VL-3B, with the admission and follow/preserve labels frozen.

On the strictly parsed items of the two Qwen2.5-VL cohorts, GLANCE exceeds pixel-FPVG by $+0.140$ $[+0.108, +0.171]$ on 7B and $+0.155$ $[+0.121, +0.189]$ on 3B. Adding single-layer MEGA

Table 21: Continuous supports against binary agreement rules on the strictly parsed Qwen2.5-VL pools. Every score in a column uses exactly the same items. Among the continuous scores that include the masked view, GLANCE and crop−mask rank highest, and both binary rules rank below every such score in each pool. Bold marks the highest AUROC in each column.

Score	Qwen2.5-VL-7B ($n = 600$)	Qwen2.5-VL-3B ($n = 687$)
GLANCE	0.811	0.798
crop−mask	**0.813**	**0.803**
full−mask	0.787	0.769
−mask	0.777	0.741
masked-greedy $\mathbf{1}[F \neq M]$	0.679	0.644
pixel-FPVG $\mathbf{1}[F = C \wedge F \neq M]$	0.672	0.643
crop−full	0.656	0.656

Continuous scores use the letter-scoring convention anchored on the original answer; the binary rules use parsed greedy answers F, C, M from the full, crop, and masked views. The pools contain 591 and 678 original-image groups.

Table 22: Continuous versus binary readouts on the six remaining strictly parsed pools (the Qwen2.5-VL cohorts are in Table 21). GLANCE exceeds pixel-FPVG with an interval above zero in five pools; on Gemma-3-12B-it the interval includes zero. Within each row all AUROCs and the difference use exactly the same items.

Pool	n/N	Image groups	GLANCE	Pixel-FPVG	Masked-greedy	GLANCE − pixel-FPVG [95% CI]
LLaVA-OneVision-7B	548/548	539	0.7970	0.6389	0.6380	+0.1580 [+0.1251, +0.1883]
Qwen3-VL-32B	597/602	582	0.7879	0.6377	0.6465	+0.1501 [+0.1204, +0.1809]
Gemma-3-12B-it	495/512	486	0.6986	0.6691	0.6589	+0.0295 [−0.0091, +0.0684]
Qwen3.6-27B	568/568	556	0.8323	0.6634	0.6672	+0.1689 [+0.1347, +0.2016]
Qwen2.5-VL-7B (MEGA SL)	403/404	403	0.7732	0.6221	0.6280	+0.1511 [+0.1087, +0.1934]
Qwen2.5-VL-3B (MEGA SL)	234/234	234	0.7450	0.5731	0.5663	+0.1719 [+0.1095, +0.2314]

n/N denotes strictly parsed items / fixed-list items; image groups are the source images resampled in the 4,000 paired bootstrap draws (unadjusted intervals). The LLaVA row retains the fixed list from before the one-image source exclusion (Appendix B.6). GLANCE uses letter scoring; the binary rules use parsed greedy answers. Qwen3.6-27B runs in non-thinking mode with `max_new_tokens=8`, matching its own gate-generation configuration; all other pools use 64. MEGA SL denotes the single-layer MEGA subsets.

7B/3B, LLaVA-OneVision-7B, Qwen3.6-27B, Qwen3-VL-32B, and Gemma extends the comparison to eight strictly parsed pools (Tables 21 and 22). Seven advantages have intervals above zero, including $+0.1501$ $[+0.1204, +0.1809]$ on Qwen3-VL-32B, and Gemma gives $+0.0295$ $[-0.0091, +0.0684]$. The strictly parsed single-layer MEGA 7B denominator is 403, distinct from the 404 items in its admission cohort.

Masked-greedy and pixel-FPVG are not separated in any of the eight strictly parsed pools: their paired differences span zero (-0.010 to $+0.009$), while GLANCE exceeds masked-greedy in every pool, with the Gemma interval's lower bound at $+0.0004$ (Table 23). These differences use the same source-image-group draws as the pixel-FPVG contrast above, which Figure 4 plots.

Parsing leaves 11 unknown items on Qwen2.5-VL-7B, 1 on single-layer MEGA 7B, 5 on Qwen3-VL-32B, and 17 on Gemma, all in the masked view. Qwen3-VL-32B retains $597/602$ items ($438/159$ follow/preserve, 582 image groups), and Gemma retains $495/512$ ($328/167$, 486 groups). These unknowns concern views of the unedited image, distinct from indeterminate edited-full responses. Setting their binary outcomes to either extreme bounds the fixed-list pixel-FPVG positive rate at $[0.406, 0.424]$ on Qwen2.5-VL-7B, $[0.376, 0.376]$ on Qwen2.5-VL-3B, $[0.183, 0.186]$ on single-layer MEGA 7B, $[0.344, 0.352]$ on Qwen3-VL-32B, and $[0.367, 0.400]$ on Gemma. The last two bounds use $207/602$ and $188/512$ known positives, respectively, with their unknowns retained in the fixed-list denominator.

Table 23: GLANCE against the simplest binary region rule, masked-greedy $\mathbf{1}[F \neq M]$, on the eight strictly parsed pools of Figure 4. GLANCE exceeds masked-greedy in every pool, with the Gemma-3-12B-it lower bound at $+0.0004$ (Panel A), whereas masked-greedy and pixel-FPVG are not separated in any pool (Panel B). Paired differences resample source-image groups $4{,}000$ times (seed 20260911, reset per pool), the same draws as the pixel-FPVG contrast; intervals are unadjusted.

A. AUROC and the paired difference GLANCE − masked-greedy

Audited model	n (F/P)	GLANCE	Masked-greedy	GLANCE − masked-greedy [95% CI]
Qwen2.5-VL-7B	600 (424/176)	0.811	0.679	+0.1327 [+0.0989, +0.1658]
Qwen2.5-VL-3B	687 (493/194)	0.798	0.644	+0.1537 [+0.1192, +0.1890]
LLaVA-OneVision-7B	548 (408/140)	0.797	0.638	+0.1589 [+0.1253, +0.1906]
Qwen3-VL-32B	597 (438/159)	0.788	0.647	+0.1413 [+0.1110, +0.1718]
Gemma-3-12B-it	495 (328/167)	0.699	0.659	+0.0398 [+0.0004, +0.0786]
Qwen3.6-27B	568 (423/145)	0.832	0.667	+0.1651 [+0.1295, +0.2004]
Qwen2.5-VL-7B (MEGA SL)	403 (183/220)	0.773	0.628	+0.1452 [+0.1014, +0.1885]
Qwen2.5-VL-3B (MEGA SL)	234 (104/130)	0.745	0.566	+0.1786 [+0.1151, +0.2411]

B. Paired difference masked-greedy − pixel-FPVG on the same draws

Audited model	Masked-greedy − pixel-FPVG [95% CI]
Qwen2.5-VL-7B	+0.0068 [−0.0062, +0.0188]
Qwen2.5-VL-3B	+0.0013 [−0.0125, +0.0138]
LLaVA-OneVision-7B	−0.0009 [−0.0160, +0.0118]
Qwen3-VL-32B	+0.0088 [−0.0054, +0.0219]
Gemma-3-12B-it	−0.0102 [−0.0299, +0.0075]
Qwen3.6-27B	+0.0038 [−0.0111, +0.0167]
Qwen2.5-VL-7B (MEGA SL)	+0.0059 [−0.0033, +0.0169]
Qwen2.5-VL-3B (MEGA SL)	−0.0067 [−0.0231, +0.0093]

F/P denotes follow/preserve counts on the strictly parsed items. Differences keep four decimals because they are small relative to three-decimal rounding. Qwen3.6-27B runs in non-thinking mode; MEGA SL denotes the single-layer MEGA subsets. Pixel-FPVG AUROCs on the same items appear in Tables 21 and 22.

Table 24: Same-channel binary comparison on the cached Qwen2.5-VL subsets. Switching the binary rule from greedy answers (pixel-FPVG) to teacher-forced argmax choices (b_{TF}) changes AUROC by at most 0.003, so the answer channel accounts for little of the gap between GLANCE and the binary rule; the gap itself ($+0.177$ and $+0.138$) reflects binarization together with the agreement rule's aggregation. Full strictly parsed pools and cached subsets are separate populations; paired differences share items within each row.

A. Absolute AUROC [95% CI]

Model	Sample	n (F/P)	GLANCE	TF binary	Pixel-FPVG
7B	Full	600 (424/176)	0.811 [0.773, 0.847]	—	0.672 [0.635, 0.707]
7B	Cached	250 (179/71)	0.818 [0.756, 0.874]	0.642 [0.581, 0.699]	0.642 [0.581, 0.699]
3B	Full	687 (493/194)	0.798 [0.760, 0.835]	—	0.643 [0.608, 0.677]
3B	Cached	269 (196/73)	0.803 [0.740, 0.860]	0.665 [0.608, 0.717]	0.663 [0.607, 0.715]

B. Paired differences [95% CI]

Model	Sample	n (F/P)	Δ_{mag}	Δ_{ch}	GLANCE − pixel-FPVG
7B	Full	600 (424/176)	—	—	+0.140 [+0.110, +0.170]
7B	Cached	250 (179/71)	+0.177 [+0.126, +0.228]	+0.000 [−0.008, +0.008]	+0.177 [+0.126, +0.227]
3B	Full	687 (493/194)	—	—	+0.155 [+0.121, +0.188]
3B	Cached	269 (196/73)	+0.138 [+0.089, +0.189]	+0.003 [−0.005, +0.012]	+0.140 [+0.091, +0.192]

Models are Qwen2.5-VL-7B and 3B; F/P denotes follow/preserve; "Full" is the full strictly parsed pool and "Cached" its cache-covered subset. TF binary applies the pixel-FPVG rule to teacher-forced argmax choices, $b_{\mathrm{TF}} = \mathbf{1}[\hat{a}_{\mathrm{full}} = \hat{a}_{\mathrm{crop}} \wedge \hat{a}_{\mathrm{full}} \neq \hat{a}_{\mathrm{mask}}]$; $\Delta_{\mathrm{mag}} = \mathrm{AUROC}(S) - \mathrm{AUROC}(b_{\mathrm{TF}})$ and $\Delta_{\mathrm{ch}} = \mathrm{AUROC}(b_{\mathrm{TF}}) - \mathrm{AUROC}(b)$. Unadjusted 95% intervals resample source-image groups $4{,}000$ times (seed 20260919, reset per population); tied scores receive half credit. Full-pool rows are recomputed with this seed, so their intervals can differ in the third decimal from Appendix F and Figure 4.

F.2 Continuous Ranking Within Binary Outcomes

To test whether continuous support separates items assigned the same binary outcome, we stratify each strict-parsing pool by $b = \mathbf{1}[F = C \wedge F \neq M]$. Figure 7 retains all sixteen strata. GLANCE ranks follow above preserve in all eight $b = 0$ groups, with unadjusted intervals above chance. For $b = 1$, only the two Qwen2.5-VL cohorts and Qwen3.6-27B have intervals above chance, whereas those for LLaVA, Qwen3-VL-32B, Gemma, and both single-layer MEGA cohorts span chance.

The protocol was fixed before computing these post-hoc stratum estimates. Each of 4,000 bootstrap draws resamples source-image groups from the complete strict pool before evaluating both strata, and the generator resets to seed 20260917 for each cohort. All draws are estimable. Different interval positions relative to chance do not test a difference between the two stratum AUROCs. Continuous teacher-forced supports and greedy answers also use different measurement channels, so the comparison does not isolate numerical binarization.

F.3 Binary Agreement Within the Likelihood Channel

For each original-image view, define $\hat{a}_v = \arg\max_j p_v(j)$ and $b_{\mathrm{TF}} = \mathbf{1}[\hat{a}_{\mathrm{full}} = \hat{a}_{\mathrm{crop}} \wedge \hat{a}_{\mathrm{full}} \neq \hat{a}_{\mathrm{mask}}]$. This substitutes teacher-forced argmax choices for greedy answers in the pixel-FPVG rule. Ties use the smallest option index, and no maximum ties occur in any cached view. Unique cohort-and-item joins recover 250 7B and 269 3B items, and their original-answer probability components match the registered crop, full, and masked supports exactly.

Table 24 separates the cached subsets from their full strict pools. Full/subset source-image counts are $591/248$ for 7B and $678/269$ for 3B. We report $\Delta_{\mathrm{mag}} = \mathrm{AUROC}(S) - \mathrm{AUROC}(b_{\mathrm{TF}})$ and $\Delta_{\mathrm{ch}} = \mathrm{AUROC}(b_{\mathrm{TF}}) - \mathrm{AUROC}(b)$, alongside their total. AUROC assigns tied scores half credit, and every bootstrap draw is estimable. These are nested, cache-covered subsets whose composition is reported without a representativeness test.

Both Δ_{mag} intervals lie above zero, meeting the prespecified criterion. Teacher-forced and greedy binary outputs disagree on only $2/250$ and $3/269$ items, and both Δ_{ch} intervals span zero. Within these two cached pools, continuous original-answer support therefore retains information lost by

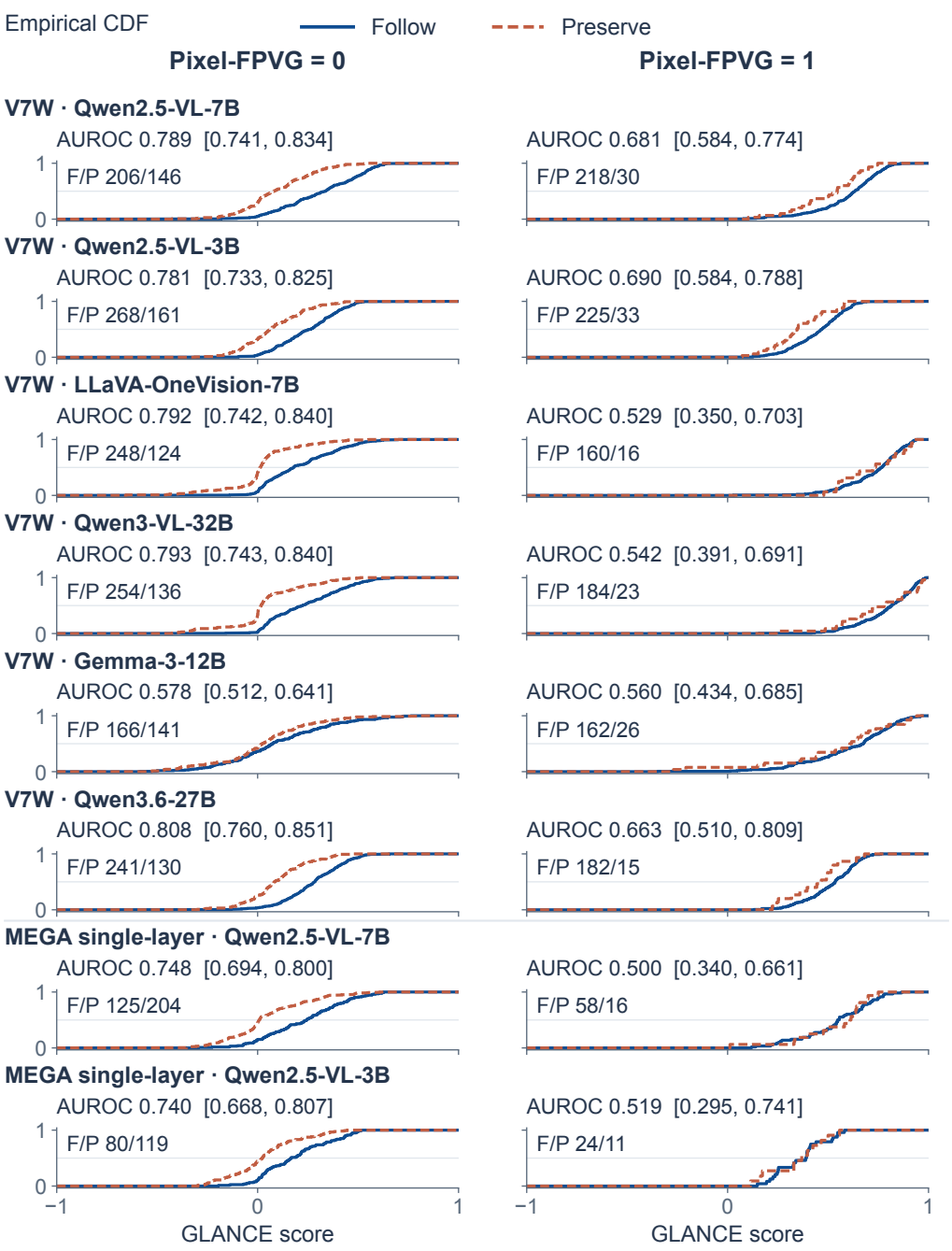


Figure 7: Continuous scores within each binary pixel-FPVG outcome. GLANCE ranks follow above preserve with intervals above chance in all eight FPVG=0 strata, while five of the eight FPVG=1 intervals include chance. This post-hoc diagnostic retains all eight strictly parsed pools and both strata. Curves are empirical GLANCE distributions, and F/P gives follow/preserve counts. Each panel reports within-stratum AUROC with an unadjusted 95% source-image-group bootstrap interval (4,000 draws, seed 20260917), and each complete pool is resampled before taking its two strata.

same-channel argmax agreement. The comparison also changes the aggregation rule rather than thresholding the GLANCE score itself, and same-channel results are unavailable for the other cohorts.

F.4 Multiplicity Sensitivity

As a post-hoc, approximate sensitivity analysis, we widen the intervals of three claim families to family-wise Bonferroni intervals. The families were fixed before the extended resampling: the margin over the attention and geometry panel on the six V7W cohorts ($m = 6$), GLANCE against masked-greedy and pixel-FPVG on the eight strictly parsed pools ($m = 16$), and the fixed-OOF masked-support increment Δ_{visible} on the eight cohorts ($m = 8$). Each family keeps its original percentile-bootstrap scheme: source-image groups for the first two, and fixed out-of-fold predictions for the third, which excludes refitting variation. After the 30 registered intervals were reproduced at 4,000 draws, resampling was extended along the same random streams to 50,000 draws, and each member's interval is taken at the $\alpha/(2m)$ and $1 - \alpha/(2m)$ quantiles with $\alpha = 0.05$. Lower bounds stay above zero for 5 of 6 panel margins (Gemma-3-12B-it: $[-0.056, +0.116]$), 14 of 16 binary-rule comparisons (Gemma-3-12B-it against masked-greedy: $[-0.018, +0.096]$; against pixel-FPVG: $[-0.027, +0.085]$), and all 8 masked-support increments (smallest lower bound $+0.030$, single-layer MEGA 3B). These are approximate within-family simultaneous intervals whose coverage depends on the underlying percentile-bootstrap approximation. They do not provide a paper-wide error guarantee: other intervals remain unadjusted. Shared edit assets do not invalidate the Bonferroni adjustment, but the cohorts are not independent replications.

G Populations and Edit Validity

This appendix gives the details behind the population comparisons of Section 4.2 and the validity analyses of Section 4.5.

G.1 Populations and Admission Rules

Table 25 reports discrimination across the populations discussed in Section 4.2: the any-change target on originally correct populations without post-edit admission, and follow versus preserve under alternative admission rules.

Admission defines the population on which GLANCE is evaluated. The audited-model gate requires an originally correct answer, a target-answer response on the edited crop, and displaced-placebo stability. The external-model gate retains original correctness but uses LLaVA-OneVision-7B's crop response and placebo stability for admission. Human admission, by one author, requires a region judged relevant (yes or partial) and an independently selected edited answer matching the target, and it adds no machine-crop or placebo filter. All three rules retain the audited model's edited-full follow/preserve labels and GLANCE's letter scores from the unedited original image, so their AUROCs describe distinct conditional populations and do not rank admission rules. The external-gate study has different model coverage and sampling composition and retains the same edit-generation process, audited-model labels, and original-image scores (Tables 26 and 27).

The originally correct pools also separate what GLANCE ranks from what the edit does to the crop (Table 28). GLANCE ranks whether the audited model answers the target on the edited crop, the first post-edit admission check, at 0.612–0.643, below its ranking of any answer change on the edited image (0.753–0.792), and all four differences have intervals below zero. The crop response is the audited model's own answer rather than a semantic validity check, and the endpoint differences describe the ranking without decomposing it into causes.

The supplementary bank combines items that failed an earlier edited-crop check, items from an earlier determinate pool, and previously unused Visual7W items (Appendix C), so part of it was selected using earlier edit outcomes. Table 29 splits the four originally correct pools by source. On previously unused items, the any-change AUROC is 0.740–0.791, and 0.732–0.789 on the 186 items per model whose source images are also absent from the 3,000-item main sample; the earlier crop-failure and determinate strata give 0.733–0.750 and 0.749–0.825, without tests between strata. On the full pools, GLANCE exceeds the visible-only average by 0.210–0.257 AUROC; its difference from crop$-$mask is $+0.011$ and $+0.019$ for 7B, with intervals above zero, and is not separated for 3B,

Table 25: Discrimination across populations and prediction endpoints. Panel A evaluates any parsed answer change versus an unchanged answer on originally correct populations without crop or placebo admission of the new edits. Panel B evaluates follow versus preserve under alternative admission rules. Rows describe scope; they do not rank admission rules.

A. Any change vs. unchanged (T2), originally correct population without crop/placebo admission

Population	Model / edit seed	n	AUROC [95% CI]
Originally correct	Qwen2.5-VL-7B, seed 0	611	0.753 [0.714, 0.791]
Originally correct	Qwen2.5-VL-7B, seed 1	611	0.760 [0.722, 0.797]
Originally correct	Qwen2.5-VL-3B, seed 0	618	0.792 [0.756, 0.826]
Originally correct	Qwen2.5-VL-3B, seed 1	618	0.775 [0.739, 0.812]

B. Follow vs. preserve (T1) under alternative admission

Population	Model / assets	n	AUROC [95% CI]
Audited-model gate (Table 1)	Six V7W cohorts	512–688	0.707–0.832
Originally correct, no crop/placebo gate	Qwen2.5-VL-7B, seed 0	538	0.748 [0.706, 0.789]
Originally correct, no crop/placebo gate	Qwen2.5-VL-7B, seed 1	547	0.756 [0.715, 0.796]
Originally correct, no crop/placebo gate	Qwen2.5-VL-3B, seed 0	560	0.784 [0.745, 0.820]
Originally correct, no crop/placebo gate	Qwen2.5-VL-3B, seed 1	562	0.781 [0.743, 0.818]
Human-admitted pool (one author, two rounds pooled)	Qwen2.5-VL-7B, seed 0	202	0.734 [0.659, 0.805]
External pass, audited-model reject	LLaVA-OneVision-7B checks; 7B/3B, two seeds pooled	133	0.742 [0.653, 0.824]

Panel A: membership requires only a correct original answer; every originally correct item has a parsed edited-full letter and three-view scores, and T2 measures whether the parsed answer changes, not whether the model responds correctly to a valid semantic change. Panel B: external admission uses LLaVA-OneVision-7B crop and placebo responses while labels and scores remain the audited model's, and the two VLMs are not independent. Intervals: item bootstraps, except the external row, which keeps both edit seeds of an item together. Items repeated across seeds and nested subsets are not independent replications. Populations differ in composition and model coverage.

Table 26: Two machine admission rules on the supplementary assets, pooled over both edit seeds. Both keep the audited model's follow/preserve labels and original-image GLANCE scores, and because the populations differ, the AUROCs do not rank the rules. The one-author human admission pool is in Table 25.

Admission rule	Crop and placebo checks	Model / seed coverage	n	GLANCE AUROC [95% CI]
Audited-model	The audited model's own responses	Qwen2.5-VL-7B and 3B, seeds 0 and 1	948	0.830 [0.798, 0.861]
External-model	LLaVA-OneVision-7B responses	Same shared edit assets	878	0.833 [0.797, 0.867]

Intervals resample model:item clusters, keeping both seeds of an item together.

Table 27: Per-seed estimates under audited-model and external (LLaVA-OneVision-7B) admission on the supplementary assets. Both rules retain the row's audited-model follow/preserve labels and shared original-image scores; all eight estimates lie within 0.801 to 0.868.

		Audited-model admission		External admission	
Audited model	Edit seed	n	AUROC	n	AUROC
Qwen2.5-VL-7B	0	219	0.801	212	0.835
Qwen2.5-VL-7B	1	227	0.838	202	0.837
Qwen2.5-VL-3B	0	249	0.868	237	0.841
Qwen2.5-VL-3B	1	253	0.834	227	0.835

Table 28: Three endpoints on the four originally correct supplementary pools, ranked by the same frozen GLANCE score. GLANCE ranks whether the audited model answers the target on the edited crop at 0.612 to 0.643, below its ranking of any answer change on the edited image (0.753 to 0.792), and every difference interval lies below zero. The endpoints share items, so the differences describe the ranking rather than decompose it into causes.

		AUROC			
Pool	F/third/unch.	Crop answers target [95% CI]	Any change [95% CI]	Follow vs. rest	Crop − any change [95% CI]
7B, seed 0	198/73/340	0.632 [0.586, 0.676]	0.753 [0.714, 0.791]	0.702	−0.121 [−0.162, −0.080]
7B, seed 1	192/64/355	0.612 [0.567, 0.657]	0.760 [0.722, 0.797]	0.718	−0.148 [−0.188, −0.109]
3B, seed 0	218/58/342	0.641 [0.597, 0.685]	0.792 [0.756, 0.826]	0.738	−0.151 [−0.193, −0.110]
3B, seed 1	215/56/347	0.643 [0.600, 0.687]	0.775 [0.739, 0.812]	0.747	−0.132 [−0.175, −0.089]

Models are Qwen2.5-VL-7B (611 originally correct items per seed) and Qwen2.5-VL-3B (618). F/third/unch. counts follow, third-option, and unchanged responses on the edited image. "Crop answers target" is the audited model's own edited-crop response, not a semantic validity check, and follow vs. rest scores follow against third-option and unchanged responses. On the admitted determinate subsets of the same pools, GLANCE ranks follow above preserve at 0.801 to 0.868 (Table 27). Intervals are item bootstraps (4,000 draws, seed 0), unadjusted. The analysis plan was fixed before computation, and the two seeds share original images.

and its differences from full−mask and from mask-only support change sign across pools. These T2 contrasts on originally correct pools differ from the T1 comparison on admission cohorts in Table 1: the positive 7B results do not establish a consistent advantage over crop−mask across both models.

G.2 Human-Valid Subsets

Under class-balanced sampling of machine-admitted items, a majority of three raters independently selected the edited target for $70/90 = 0.778$ of the 7B items and $69/89 = 0.775$ of the 3B items. Among edits judged invalid, preserve accounted for $17/20$ and $18/20$, respectively. Within these response-balanced samples, $17/45$ and $18/45$ preserve items and $3/45$ and $2/44$ follow items are invalid edits; these are sample proportions, not invalidity rates of the six-model populations, and a majority not selecting the target does not by itself make preservation correct. When an edit does not realize its target, preserving the original answer can be appropriate, so discrimination may partly track properties associated with edit success, which the present audits do not identify.

We join each three-rater item to the judge's exact edited and original RGB pixels, retaining the original majority-target rule, and all $90/89$ identities match for 7B/3B. In the rated 7B sample, GLANCE AUROC is 0.802 $[0.707, 0.890]$ across 90 items and 0.809 $[0.700, 0.904]$ on the 70 majority-valid items. The corresponding 3B values are 0.839 $[0.749, 0.915]$ across 89 items and 0.842 $[0.742, 0.924]$ on 69 valid items. These nested, response-balanced samples show the estimates after excluding majority-invalid edits but do not test a filtering effect. Table 30 reports the matched probes and the fixed readouts on the valid edits.

In the same samples, GLANCE separates valid follow items from invalid preserve items with AUROC 0.805 $[0.672, 0.922]$ on 7B (42 and 17 items) and 0.845 $[0.723, 0.939]$ on 3B (42 and 18). The valid-versus-invalid preserve comparisons yield 0.542 $[0.347, 0.725]$ and 0.553 $[0.361, 0.732]$, with wide intervals. Median GLANCE is 0.424, 0.133, and 0.144 for valid follow, valid preserve, and invalid preserve items on 7B and 0.388, 0.085, and 0.099 on 3B. With 17 and 18 invalid preserve items these estimates are descriptive; raw scores are not pooled across models, and intervals use 4,000 source-image-group resamples, unadjusted.

To check the probe increments beyond fixed out-of-fold predictions, we retain the three-rater valid samples and their original pooled-OOF estimator, then resample source-image groups and repeat the complete fitting procedure. All copies of a resampled source remain in one fold, and visible-only, mask-only, and all-view probes share the same draws and splits. All $1{,}000$ draws per model are estimable. Table 31 reports the original points with empirical refitting sensitivity intervals. Both masked-input increments remain positive, with intervals $[0.121, 0.531]$ on 7B and $[0.194, 0.611]$ on 3B, whereas both reverse increments span zero.

Table 29: Any-change ranking (T2) on the four originally correct supplementary pools, split by the source of each item (post hoc). GLANCE also ranks answer changes within the previously unused and new-source-image subsets. Panel B contrasts GLANCE with other readouts on the full pools; these T2 contrasts differ from the Table 1 comparison on admission cohorts.

A. T2 GLANCE AUROC by source stratum

Pool	Stratum	n (images)	Change/unch.	AUROC [95% CI]
7B, seed 0	All originally correct	611 (603)	271/340	0.753 [0.714, 0.791]
	Earlier crop-check failures	200 (199)	53/147	0.750 [0.677, 0.817]
	Earlier determinate items	200 (199)	134/66	0.749 [0.673, 0.816]
	Previously unused	211 (210)	84/127	0.763 [0.695, 0.824]
	Previously unused, new source image	186 (185)	79/107	0.752 [0.679, 0.820]
7B, seed 1	All originally correct	611 (603)	256/355	0.760 [0.722, 0.797]
	Earlier crop-check failures	200 (199)	51/149	0.746 [0.673, 0.815]
	Earlier determinate items	200 (199)	125/75	0.789 [0.723, 0.851]
	Previously unused	211 (210)	80/131	0.740 [0.669, 0.806]
	Previously unused, new source image	186 (185)	73/113	0.732 [0.656, 0.804]
3B, seed 0	All originally correct	618 (604)	276/342	0.792 [0.757, 0.828]
	Earlier crop-check failures	200 (198)	54/146	0.744 [0.671, 0.811]
	Earlier determinate items	200 (200)	140/60	0.825 [0.752, 0.888]
	Previously unused	218 (215)	82/136	0.791 [0.728, 0.851]
	Previously unused, new source image	186 (183)	67/119	0.789 [0.717, 0.854]
3B, seed 1	All originally correct	618 (604)	271/347	0.775 [0.739, 0.812]
	Earlier crop-check failures	200 (198)	53/147	0.733 [0.658, 0.801]
	Earlier determinate items	200 (200)	136/64	0.799 [0.731, 0.858]
	Previously unused	218 (215)	82/136	0.776 [0.710, 0.837]
	Previously unused, new source image	186 (183)	71/115	0.780 [0.708, 0.846]

B. Full-pool paired T2 differences: GLANCE minus readout [95% CI]

Pool	crop−mask	full−mask	$-g_{\text{mask}}$	$\frac{1}{2}(g_{\text{crop}} + g_{\text{full}})$
7B, seed 0	+0.011 [+0.002, +0.020]	+0.003 [−0.007, +0.013]	−0.015 [−0.042, +0.011]	+0.210 [+0.163, +0.256]
7B, seed 1	+0.019 [+0.010, +0.028]	−0.006 [−0.017, +0.004]	−0.034 [−0.060, −0.008]	+0.257 [+0.212, +0.301]
3B, seed 0	+0.002 [−0.008, +0.013]	+0.023 [+0.010, +0.036]	+0.016 [−0.014, +0.046]	+0.254 [+0.207, +0.300]
3B, seed 1	+0.004 [−0.007, +0.015]	+0.020 [+0.006, +0.033]	+0.031 [+0.002, +0.060]	+0.210 [+0.161, +0.259]

Post-hoc analysis; the plan and script were frozen before computation. Only the Qwen2.5-VL-7B and 3B supplementary assets are covered, and the two edit seeds share items and source images, so seeds are not independent replications. The new-source-image subset still uses the same data source and editing pipeline. T2 counts any parsed answer change, not a valid semantic update. Intervals use 4,000 source-image-group resamples (seed 0), unadjusted; strata are not tested against each other.

Table 30: Edits judged target-valid by a majority of three raters (Qwen2.5-VL-7B and 3B). The masked-input increment is positive with an interval above zero in both subsets, and the continuous readout exceeds pixel-FPVG on 3B, while the 7B difference includes zero.

A. Matched probes: absolute OOF AUROC and masked-input increment

Model	Follow/preserve	Visible	Masked	All three	All three − visible [95% CI]
7B	42/28	0.472	0.795	0.763	+0.291 [+0.151, +0.421]
3B	42/27	0.324	0.777	0.799	+0.475 [+0.355, +0.590]

B. Fixed continuous and binary readouts on the same items

Model	GLANCE	Pixel-FPVG	Paired difference [95% CI]
7B	0.809	0.762	+0.047 [−0.042, +0.138]
3B	0.842	0.689	+0.153 [+0.065, +0.240]

Earlier class-balanced samples, restricted post hoc by the original majority rule. Panel A averages 20 grouped five-fold OOF AUROCs. Intervals use 4,000 paired source-image-group draws, unadjusted for multiplicity; probe intervals condition on fitted OOF predictions and exclude refitting and annotation uncertainty. Pixel-FPVG uses strictly parsed greedy responses, available for all 70/69 items.

Table 31: Human-valid probe increments with refitting sensitivity intervals on the three-rater target-valid samples. Both masked-input increments stay positive across 1,000 source-group resamples that repeat splitting and fitting, whereas both reverse increments include zero.

Model	Aggregation	All three − visible [95% CI]	All three − masked [95% CI]
Qwen2.5-VL-7B	pooled OOF	+0.291 [+0.121, +0.531]	−0.032 [−0.090, +0.026]
Qwen2.5-VL-7B	within fold	+0.247 [+0.091, +0.481]	−0.030 [−0.100, +0.014]
Qwen2.5-VL-3B	pooled OOF	+0.475 [+0.194, +0.611]	+0.023 [−0.056, +0.122]
Qwen2.5-VL-3B	within fold	+0.474 [+0.160, +0.569]	+0.009 [−0.065, +0.084]

Original-sample points with percentile 95% intervals from 1,000 source-group resamples per model; every resample repeats the twenty grouped five-fold fits, and duplicate source copies stay in the same fold. All 2,000 draws are estimable. These empirical sensitivity intervals include the specified fitting variation but remain conditional on the selected samples and cached measurements; they are distinct from the fixed-OOF intervals in Table 30.

Table 32: Human-majority validity and VLM-judge decisions on the same edits (Qwen2.5-VL-7B and 3B). Even among edits that a majority of three raters judged valid, the judge passes follow responses far more often than preserve responses, under both treatments of uncertain judgments.

A. Joint counts

Model	Human valid	Response	Pass	Fail	Uncertain
7B	Yes	follow	27	13	2
7B	Yes	preserve	8	14	6
7B	No	follow	1	2	0
7B	No	preserve	0	16	1
3B	Yes	follow	38	3	1
3B	Yes	preserve	4	21	2
3B	No	follow	0	1	1
3B	No	preserve	0	15	3

B. Judge acceptance within human-valid edits

Model	Uncertain	Follow pass/n, rate [95% CI]	Preserve pass/n, rate [95% CI]	Difference [95% CI]
7B	counted as fail	27/42, 0.643 [0.492, 0.770]	8/28, 0.286 [0.153, 0.471]	+0.357 [+0.131, +0.571]
7B	excluded	27/40, 0.675 [0.520, 0.799]	8/22, 0.364 [0.197, 0.570]	+0.311 [+0.055, +0.562]
3B	counted as fail	38/42, 0.905 [0.779, 0.962]	4/27, 0.148 [0.059, 0.325]	+0.757 [+0.578, +0.905]
3B	excluded	38/41, 0.927 [0.806, 0.975]	4/25, 0.160 [0.064, 0.347]	+0.767 [+0.598, +0.921]

Rates have Wilson 95% intervals; follow-minus-preserve differences have source-image-group bootstrap intervals (4,000 draws, seed 20260919 reset per model and population). All intervals are unadjusted. "Counted as fail" treats uncertain judgments as non-pass; "excluded" removes them from the denominator.

The 3B visible-only probe ranks below chance under the pooled-OOF estimator (0.324), so we check its coefficients and whether cross-fold aggregation causes the reversal. Its crop coefficient is positive in 56 fits and negative in 44, and its full-image coefficient is positive in 68 and negative in 32. Computing concordance within each test fold and weighting it by the fold's positive–negative pair count gives 0.351 instead of 0.324, and the within-fold masked-input increment is 0.474, close to the pooled estimate, so cross-fold aggregation alone does not explain the weak visible-only predictor. On 7B, the corresponding increment changes from 0.291 to 0.247. Both estimators are retained as specified before this sensitivity analysis. Resampling reduces the number of distinct training sources and can shift the fitting regime: mean resampled pooled increments are $0.320/0.417$, compared with original points $0.291/0.475$. The intervals describe this specified resampling and refitting sensitivity without establishing nominal population coverage. The `upgrade/human-refit/` folder of the reproducibility archive preserves every draw and source-fold assignment, absolute probe intervals, coefficient diagnostics, and a replay of all 600 original fits.

G.3 Judge Acceptance on Human-Valid Edits

Table 32 reports every joint count and both treatments of uncertain judgments. With uncertain counted as non-pass, judge acceptance of valid follow versus valid preserve differs by 0.357 $[0.131, 0.571]$ on 7B and 0.757 $[0.578, 0.905]$ on 3B. Excluding uncertain leaves positive differences, 0.311 $[0.055, 0.562]$ and 0.767 $[0.598, 0.921]$. Under both treatments, judge filtering selects response-dependent populations even among human-valid edits. With only 28 and 27 valid preserve edits, precision is limited.

Counting uncertain as non-pass, raw agreement with human-majority validity is $0.600/0.697$ and $\kappa = 0.280/0.411$ for 7B/3B. Human-valid fractions are $70/90$ and $69/89$, while judge-pass fractions are $36/90$ and $42/89$. Excluding uncertain gives agreement $0.654/0.707$, $\kappa = 0.347/0.406$, human-valid fractions $62/81$ and $66/82$, and judge-pass fractions $36/81$ and $42/82$. Among the 20 human-invalid edits per model, the judge accepts $1/3$ follow and $0/17$ preserve for 7B, and $0/2$ follow and $0/18$ preserve for 3B, small strata that are descriptive only. These post-hoc comparisons reuse the existing three-rater samples and add no annotation, and the separately pre-registered human validation, with three raters and external calibration, was not completed.

Table 33: VLM-judge pass subsets of the five judge-covered V7W cohorts: composition and absolute discrimination (A) and paired readout contrasts within the same subsets (B). Pass subsets keep few preserve responses, and eight of the ten paired intervals include zero, so the positive point estimates preserve direction rather than establish an advantage.

A. Subset composition and absolute discrimination

Audited model	Follow	Preserve	GLANCE AUROC [95% CI]
Qwen2.5-VL-7B	313	18	0.725 [0.604, 0.829]
Qwen2.5-VL-3B	396	23	0.640 [0.526, 0.748]
LLaVA-OneVision-7B	296	21	0.734 [0.602, 0.855]
Qwen3-VL-32B	312	17	0.679 [0.510, 0.835]
Gemma-3-12B-it	252	27	0.619 [0.499, 0.735]

B. Paired differences within the same subsets

Audited model	Mask-readout contrast [95% CI]	Displaced-mask contrast [95% CI]
Qwen2.5-VL-7B	+0.138 [+0.001, +0.287]	+0.135 [−0.028, +0.305]
Qwen2.5-VL-3B	+0.106 [−0.036, +0.241]	+0.107 [−0.048, +0.260]
LLaVA-OneVision-7B	+0.125 [−0.028, +0.286]	+0.206 [+0.009, +0.407]
Qwen3-VL-32B	+0.192 [−0.028, +0.408]	+0.121 [−0.090, +0.334]
Gemma-3-12B-it	+0.056 [−0.095, +0.216]	+0.045 [−0.118, +0.206]

All rows use Qwen3.6-27B (non-thinking) as judge and the same blind two-call pass criterion (Appendix C.3). Mask-readout: $\mathrm{AUROC}(g_{\mathrm{crop}} - g_{\mathrm{mask}}) - \mathrm{AUROC}(g_{\mathrm{crop}} - g_{\mathrm{full}})$; displaced mask: $\mathrm{AUROC}(S_R) - \mathrm{AUROC}(S_W)$, with only the masked box displaced to W, which can overlap R. Intervals use item bootstraps, paired for contrasts, unadjusted. The judge audits these five admission cohorts in full except that LLaVA-OneVision-7B uses the 547-item pool after the one-image source exclusion; Qwen3.6-27B lacks a matched quality baseline for its cohort.

G.4 Judge-Pass Subsets

Table 33 reports, for the five audited models whose admission cohorts the judge covers, the follow/preserve composition and absolute GLANCE AUROC of the VLM-judge pass subsets (Panel A) and two fixed-readout paired contrasts within the same subsets (Panel B). On Qwen3-VL-32B, the 329 passes (312 follow, 17 preserve) yield mask-readout $+0.1921$ $[-0.0275, +0.4075]$ and displaced-mask $+0.1210$ $[-0.0898, +0.3344]$. On Gemma, the 279 passes (252 follow, 27 preserve) yield $+0.0561$ $[-0.0954, +0.2156]$ and $+0.0447$ $[-0.1177, +0.2061]$. All four intervals span zero, and eight of the ten paired intervals in the table cross zero. The exceptions are the Qwen2.5-VL-7B mask-readout contrast and the LLaVA displaced-mask contrast, so positive point estimates preserve direction rather than establish uniform advantages.

A composition diagnostic separates the effect of fewer preserve items from other changes. It draws subsets from each original pool with the pass pool's exact follow/preserve counts, ignoring judge decisions, using 300 outer draws without replacement and $1{,}000$ item-bootstrap draws within each subset. For the paired contrasts, the composition-matched null subsets place the Qwen3-VL-32B and Gemma estimates at the 59th/16th and 5th/11th percentiles, all within their null 95% reference intervals, and $47.7\%/36.3\%$ and $28.3\%/60.3\%$ of null subsets have paired intervals that cross zero. Reduced preserve counts therefore account for the loss of precision in the paired contrasts. They do not account for every change in the absolute discrimination: the Qwen2.5-VL-3B pass subset's AUROC of 0.640 falls below its composition range $[0.694, 0.876]$, and the panel margins of the 3B and LLaVA pass pools fall below their central composition ranges (Table 34). These descriptive departures use outer-draw point estimates and no prespecified decision threshold, and because pass subsets are nested in the original pools, the contrast combines quality selection with other population differences.

Table 34: Judge-pass pools whose panel margins fall below their composition-matched reference ranges. For these two pass pools the observed margin over the nine-slot region-blind panel lies below the central 95% range of margins obtained by drawing subsets with the same follow/preserve counts without reading judge outcomes, so class composition alone does not account for the change.

Pass pool	Observed margin	Central 95% composition range
V7W, Qwen2.5-VL-3B	+0.081	[+0.083, +0.304]
V7W, LLaVA-OneVision-7B	+0.107	[+0.112, +0.369]

Ranges contain the central 95% of composition-matched outer-draw point estimates; margins retain the nine local signal slots. Max-softmax and original-answer support (g_{full}) coincide on every row in five of eleven audited cache pools, which include pools not analyzed here, with a minimum overlap of 97.66% elsewhere; pass-subset overlap rates are unavailable.

H Target Selection and Option Position

This appendix gives the details behind Section 4.6 and the option-position analysis of Section 5. GLANCE carries predictive information about follow versus preserve responses. To test whether that information also identifies the designated target answer, we examine responses that have already changed. T1 compares follow with preserve, and T2 compares any change, follow or a third option, with an unchanged answer. T3 compares follow with a third option within changed responses, where a third option must be parsed and differ from both the target and the original answer, so parsing failures remain indeterminate and do not enter T3. The three targets therefore contain different sets of items. All tests use the fixed GLANCE score with letter scoring on three views of the unedited original image, anchored to the original answer.

H.1 Originally Correct Pools

Across Qwen2.5-VL-7B and 3B with two edit seeds on the supplementary item bank (H1 in the tables), the four originally correct pools yield T3 AUROCs 0.470–0.534, each with an interval spanning chance. T1 AUROCs are 0.748–0.784, and T2 (any change versus unchanged) AUROCs are 0.753, 0.760, 0.792, and 0.775 for 7B s0/s1 and 3B s0/s1. A coverage check confirms that every originally correct item in these four pools has a parsed edited-full letter and three-view scores, so T2 excludes no item beyond original correctness. Admission here requires original correctness, scored views, and parsed responses, with no crop-target or placebo filter, and these broader pools yield the most precise T3 estimates.

Machine admission on the same bank adds the edited-crop and placebo checks, but each such pool contains only 7–17 third-option responses, and T3 interval half-widths of approximately 0.13–0.22 provide insufficient precision to support the negative finding alone. Together with the broader pools, these estimates leave target-selection discrimination unresolved under every examined admission rule. These nested populations add no cells to the twelve-cell summary below.

Table 35: Originally correct pools on the supplementary (H1) assets under two edit seeds, without crop or placebo admission. Follow versus preserve (T1) stays at 0.748 to 0.784, whereas target versus third option (T3) stays at 0.470 to 0.534 with every interval including 0.5.

Audited model	Edit seed	Originally correct n	T1 n	T1 AUROC	T3 n / third	T3 AUROC [95% CI]
Qwen2.5-VL-7B	0	611	538	0.748	271 / 73	0.487 [0.410, 0.562]
Qwen2.5-VL-7B	1	611	547	0.756	256 / 64	0.505 [0.427, 0.587]
Qwen2.5-VL-3B	0	618	560	0.784	276 / 58	0.470 [0.391, 0.549]
Qwen2.5-VL-3B	1	618	562	0.781	271 / 56	0.534 [0.447, 0.621]

T1 compares follow with preserve; T3 compares follow with parsed third-option responses, and "T3 n / third" gives the T3 denominator and its third-option count, so the two targets share no denominator. Brackets are 95% item-bootstrap intervals ($B = 4{,}000$, seed 0), without source-image clustering or multiplicity correction; edit seeds and the bootstrap seed serve different purposes.

Table 36: Eight further target-selection (T3) cells; with the four supplementary-asset cells of Table 35 they form the twelve-cell summary. Under original correctness, an edited-crop target response, and displaced-placebo stability, GLANCE ranks follow above preserve (T1) in every cell but does not consistently rank the designated target above a third option: six T3 intervals include 0.5 and two, LLaVA-OneVision-7B and Qwen3-VL-32B, lie below it.

Assets	Audited model	T1 AUROC	T3 n (follow/third)	T3 AUROC [95% item CI]
V7W	Qwen2.5-VL-7B	0.8119	473 (435/38)	0.4791 [0.3790, 0.5727]
V7W	Qwen2.5-VL-3B	0.7982	565 (494/71)	0.5577 [0.4854, 0.6293]
V7W	LLaVA-OneVision-7B	0.7970	441 (408/33)	0.3226 [0.2233, 0.4233]
V7W	Qwen3-VL-32B	0.7890	476 (443/33)	0.3736 [0.2780, 0.4791]
V7W	Gemma-3-12B-it	0.7068	393 (344/49)	0.4794 [0.3849, 0.5740]
V7W	Qwen3.6-27B	0.8323	455 (423/32)	0.4747 [0.3652, 0.5809]
MEGA single-layer	Qwen2.5-VL-7B	0.7744	194 (184/10)	0.3288 [0.1651, 0.5189]
MEGA single-layer	Qwen2.5-VL-3B	0.7450	111 (104/7)	0.5014 [0.2562, 0.7500]

T1 and T3 have different denominators; T3 excludes parsing failures. Fixed GLANCE uses letter scoring anchored on the original answer over original-image views. Brackets are 95% item-bootstrap intervals ($B = 4{,}000$, seed 0), unadjusted. The single-layer MEGA subsets have only 10 and 7 third-option responses, so their T3 intervals are wide.

H.2 Twelve-Cell Summary

The twelve-cell summary in Figure 8 comprises the four originally correct H1 cells, six model-specific V7W cells, and two single-layer MEGA cells (Tables 35 and 36). Ten T3 intervals span 0.5 and two lie below it, LLaVA and Qwen3-VL-32B. The four H1 cells apply neither human-validity filtering nor crop/placebo gates. Each row keeps its own third-option denominator, and nested admission and edit-seed sensitivities add no independent cells.

In T3, every parsed third-option response is a negative, including those that the any-change target counts as changes; third-option items without cached scores are scored with the same frozen rule, and follow scores are the cached ones. Source-image cluster intervals are $[0.3636, 0.5848]$ for 27B and $[0.2262, 0.4288]$ for LLaVA, over 449 and 435 image groups, respectively, agreeing in direction with the item intervals. For the two Qwen2.5-VL cohorts, third-option items come from an earlier subset selected for another analysis and are relabeled as the T3 negative class, and both their item and source-image-group intervals contain 0.5.

On LLaVA-OneVision-7B and Qwen3-VL-32B, higher GLANCE scores favor third options within T3, although both models retain positive T1 discrimination. Both reverse cells have 33 third-option items. The Qwen3-VL-32B T3 AUROC is 0.3736 $[0.2780, 0.4791]$, with source-image interval $[0.2708, 0.4768]$ over 470 groups. Gemma gives 0.4794 $[0.3849, 0.5740]$, or $[0.3880, 0.5753]$ over 390 groups, with 49 third options. No common explanation for the two reverse cells was identified, and the 32B third-option items have no comparison against the region-blind panel. The single-layer MEGA subsets contain only 10 and 7 third-option responses, yielding imprecise T3 estimates of 0.3288 $[0.1651, 0.5189]$ and 0.5014 $[0.2562, 0.7500]$ that cannot support the negative finding independently.

H.3 Edit Quality Behind Third-Option Answers

GLANCE predicts follow versus preserve, while target selection gives ten intervals spanning chance and two favoring third options, with none wholly above chance. The quality audit matches the original and edited image pixels, question, ordered choices, and region before reusing a judgment and audits every remaining edit, covering all four third-option pools (Table 37). In the three pools with a determinate baseline, third-option edits pass less often, and Qwen3.6-27B has no audited determinate baseline. This shared edit-quality gap does not identify a causal explanation for the T3 reversals. The judge is Qwen3.6-27B, a checkpoint within the Qwen family, rather than a human reference or independent-family validation.

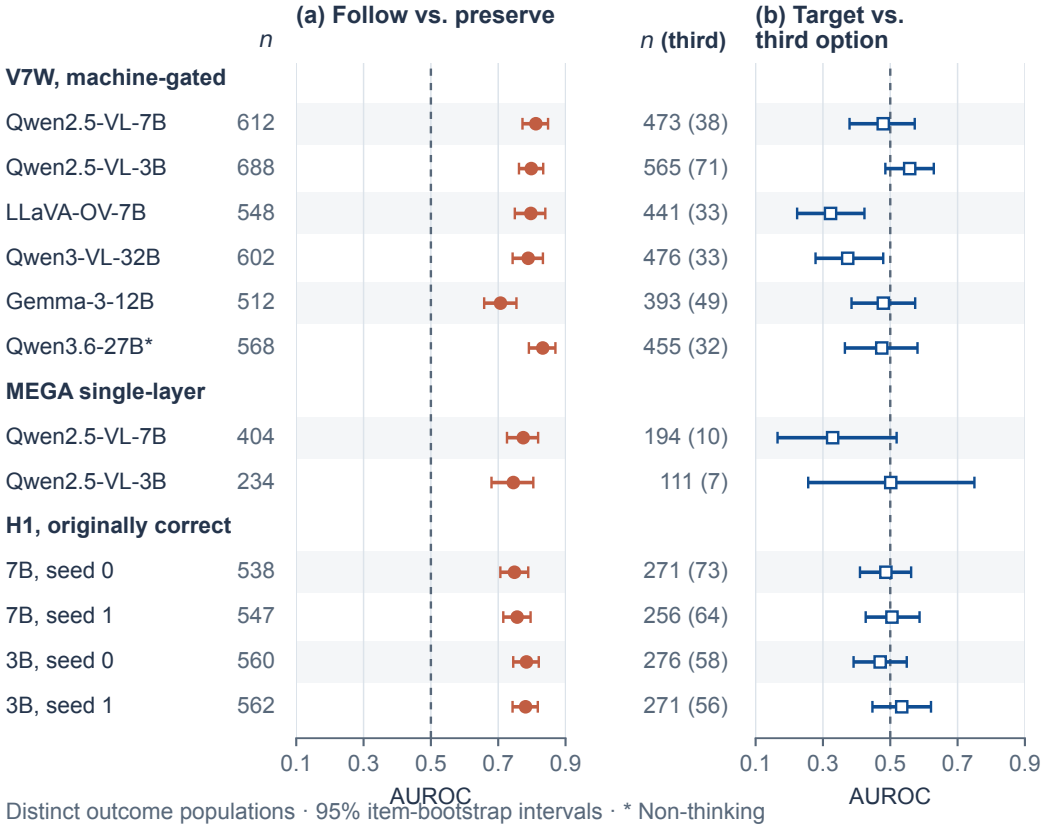


Figure 8: Regional response and target selection in twelve registered settings. GLANCE ranks follow above preserve in every setting (a) but does not consistently rank the designated target above a third option (b). (a) T1 compares follow with preserve, and (b) T3 compares the target with parsed third options among changed answers. Panels share the AUROC scale but use different outcome populations, and each reports its own n, with third-option counts in parentheses. Intervals are 95% item-bootstrap bounds, unadjusted for multiplicity. H1 rows condition on original correctness alone, and the other rows use model-specific machine admission. MEGA uses single-layer questions. Third-option edit-quality audits are in Table 37, and * denotes non-thinking mode.

Third-option edit quality remains unaudited for the two Qwen2.5-VL cohorts, MEGA, and the H1 pools, and the GQA-derived pools remain outside T3 coverage. Appendix C.3 describes how earlier audit coverage was selected and how stable repeated judgments are.

H.4 Sensitivity to Repeated Edits

Pooling both H1 edit seeds (Table 38) treats an item as a repeated observation with unchanged GLANCE and potentially changed labels. The pooled machine-gated T3 estimates are 0.5166 and 0.5123, and pooling the originally correct pools narrows the interval half-widths to 0.0648 and 0.0721. The per-seed estimates remain the main results. Source-image intervals are $[0.4317, 0.5622]$ and $[0.4044, 0.6318]$ for the 7B originally correct and machine-gated pools ($313/227$ groups), and $[0.4297, 0.5728]$ and $[0.3548, 0.6582]$ for 3B ($325/246$ groups).

LLaVA on two H1 edit sets retains positive T1 discrimination: 0.8278 $[0.7563, 0.8918]$ at seed 0 and 0.7314 $[0.6495, 0.8104]$ at seed 1, pooled to 0.7793 $[0.7176, 0.8362]$. T3 point estimates again favor third options, but only 8 and 15 third-option outcomes are available at the two seeds. The pooled item-cluster interval crosses chance, whereas the source-image interval narrowly excludes it; with so few third options, the changed-edit test does not settle the reversal on its own.

Table 37: Exact-edit VLM audit of the third-option responses in four pools. Edits that produced a third-option response pass the quality audit less often than each pool's determinate edits, with all three available rate-difference intervals below zero; the Qwen3.6-27B admission cohort has no same-instrument baseline.

	Third-option edits		Determinate edits			
Pool	pass/n	rate [95% CI]	pass/n	rate	Rate difference [95% CI]	Fisher p
Qwen3-VL-32B	6/33	0.182 [0.086, 0.344]	329/602	0.547	-0.365 [-0.491, -0.220]	< 0.0001
Gemma-3-12B-it	13/49	0.265 [0.162, 0.403]	279/512	0.545	-0.280 [-0.403, -0.145]	0.0003
LLaVA-OneVision-7B	12/33	0.364 [0.222, 0.534]	317/547	0.580	-0.216 [-0.379, -0.049]	0.0183
Qwen3.6-27B	4/32	0.125 [0.050, 0.281]	—	—	—	—

Pass requires an independently selected answer matching the target and all quality judgments passing; the judge is Qwen3.6-27B, so these are VLM assessments, not human ground truth, and its own admission cohort was not audited (dashes). Third-option rates carry Wilson 95% intervals. Differences subtract the admission-cohort pass rate from the third-option rate, with $4{,}000$ bootstrap draws (seed 20260916); intervals and Fisher tests are unadjusted for multiplicity. All four third-option pools are fully covered. The LLaVA baseline is its 547-item audit pool, distinct from the 548-item admission cohort.

Table 38: T3 sensitivity to the edit seed. Pooling both supplementary-asset seeds treats an item as a repeated observation with an unchanged GLANCE score; the pooled Qwen2.5-VL estimates stay near 0.5, while the LLaVA-OneVision-7B estimates favor third options with only 8 and 15 third-option outcomes per seed.

Audited model / population	T3 n (follow/third)	T3 AUROC [95% CI]
Qwen2.5-VL-7B / originally correct, both seeds	527 (390/137)	0.4961 [0.4313, 0.5609]
Qwen2.5-VL-7B / audited-model admission, both seeds	357 (330/27)	0.5166 [0.4038, 0.6316]
Qwen2.5-VL-3B / originally correct, both seeds	547 (433/114)	0.5012 [0.4274, 0.5717]
Qwen2.5-VL-3B / audited-model admission, both seeds	380 (358/22)	0.5123 [0.3529, 0.6552]
LLaVA-OneVision-7B / seed 0	162 (154/8)	0.3994 [0.2038, 0.6202]
LLaVA-OneVision-7B / seed 1	155 (140/15)	0.3262 [0.1780, 0.4937]
LLaVA-OneVision-7B / both seeds pooled	317 (294/23)	0.3534 [0.2111, 0.5014]

Pooled rows retain repeated items and cluster repeated edits by item ($4{,}000$ draws, seed 0); single-seed rows resample items. The pooled LLaVA T1 estimate is 0.7793 [0.7176, 0.8362], against 0.7970 on the V7W edits of the main cohort. The pooled LLaVA T3 item interval spans chance while its source-image interval [0.2091, 0.4994] narrowly excludes it, so the boundary is sensitive to the resampling unit. These nested, shared-score sensitivities add no independent cells to the twelve-cell summary.

Table 39: Detection within option-position strata on the six V7W admission cohorts. Within-stratum concordance compares only follow and preserve items sharing the same original–target letter pair, so option position cannot contribute; it differs from full-cohort AUROC by -0.008 to $+0.004$ with every interval including zero, whereas a position-only predictor reaches 0.52 to 0.57.

Audited model (n)	Quantity	Estimate [95% CI]
Qwen2.5-VL-7B (612)	Full-cohort AUROC	0.8119 [0.7721, 0.8481]
	Within-stratum concordance	0.8164 [0.7759, 0.8538]
	Difference	−0.0045 [−0.0147, +0.0057]
	Position-only AUROC	0.519
Qwen2.5-VL-3B (688)	Full-cohort AUROC	0.7982 [0.7617, 0.8337]
	Within-stratum concordance	0.7945 [0.7566, 0.8308]
	Difference	+0.0037 [−0.0035, +0.0116]
	Position-only AUROC	0.526
LLaVA-OneVision-7B (548)	Full-cohort AUROC	0.7970 [0.7493, 0.8399]
	Within-stratum concordance	0.8049 [0.7575, 0.8484]
	Difference	−0.0079 [−0.0199, +0.0020]
	Position-only AUROC	0.536
Qwen3-VL-32B (602)	Full-cohort AUROC	0.7890 [0.7423, 0.8315]
	Within-stratum concordance	0.7898 [0.7409, 0.8341]
	Difference	−0.0008 [−0.0110, +0.0096]
	Position-only AUROC	0.562
Gemma-3-12B-it (512)	Full-cohort AUROC	0.7068 [0.6583, 0.7533]
	Within-stratum concordance	0.7120 [0.6625, 0.7590]
	Difference	−0.0052 [−0.0186, +0.0084]
	Position-only AUROC	0.571
Qwen3.6-27B (568)	Full-cohort AUROC	0.8323 [0.7932, 0.8706]
	Within-stratum concordance	0.8321 [0.7915, 0.8717]
	Difference	+0.0002 [−0.0076, +0.0083]
	Position-only AUROC	0.533

Differences are full-cohort AUROC minus within-stratum concordance; brackets are 95% paired item-bootstrap intervals from this analysis (4,000 draws, seed 0), unadjusted. The position-only predictor scores each item by its stratum's in-sample follow rate and is an optimistic comparator outside the frozen panel. Qwen3.6-27B runs in non-thinking mode.

H.5 Option Position

The target rule induces four original–target strata: A→B, B→A, C→A, and D→A. With distinct option texts, targets occupy only A or B, so C and D are never selected, and the rule alone does not determine the empirical target-A frequency, which depends on the original-answer distribution. Within-stratum concordance compares only follow–preserve pairs in the same stratum and weights strata by their comparable-pair counts. Table 39 compares this quantity with full-cohort AUROC using paired item bootstrap (4,000 draws, seed 0). Differences range from -0.008 to $+0.004$, with all six intervals spanning zero, and all 24 individual-stratum AUROC intervals have lower bounds above 0.5, the smallest being 0.5785. The position-only predictor scores each item by its stratum's in-sample follow rate and reaches AUROC 0.52–0.57, an optimistic in-sample comparator outside the frozen panel. For T3, pooled within-stratum estimates are 0.4608 $[0.3586, 0.5599]$ on 7B and 0.5495 $[0.4778, 0.6233]$ on 3B, versus full-pool 0.4791 and 0.5577. C→A and D→A strata have only 113–163 items, and T3 strata contain 7–26 third-option responses, so we report pooled T3 values.

These observational strata retain the original content–position assignment and leave position–content interactions uncontrolled, so we also counterbalance the option order. For each Qwen2.5-VL cohort, we select 250 originally admitted, determinate items by original-answer-letter strata and hash rank, using pre-edit information within that pool. Four cyclic option orders retain the semantic target and cached edit. Each order regenerates the original, edited-crop, placebo, and edited-full responses, recomputes admission and labels, and rescores the original-image views. Original-order replay matches all 250 original responses per model, with maximum score difference 0, and all $2{,}000$ runs complete.

Table 40: Cyclic option rotations on the two Qwen2.5-VL cohorts. On the common subset, admitted and determinate under every order, the three rotated orders give AUROC 0.832 to 0.856 on 7B and 0.688 to 0.727 on 3B, below the original-order value in each case; the lower panel uses each order's own admitted subset.

A. Common subset

Model	n (follow/preserve)	$k = 0$ (original)	$k = 1$	$k = 2$	$k = 3$
7B	203 (151/52)	0.8669	0.8320	0.8401	0.8558
3B	193 (147/46)	0.7453	0.6884	0.7273	0.7020

B. Each order's own subset

Model	Quantity	$k = 0$ (original)	$k = 1$	$k = 2$	$k = 3$
7B	n	250	217	216	229
	follow/preserve	172/78	162/55	161/55	160/69
	AUROC	0.8334	0.8228	0.8029	0.8589
3B	n	250	206	214	224
	follow/preserve	177/73	158/48	164/50	168/56
	AUROC	0.7939	0.7268	0.7623	0.7614

Models are Qwen2.5-VL-7B and 3B. Common-subset counts use original-order labels. Each order regenerates the original, edited-crop, placebo, and edited-full responses, recomputes admission and labels, and rescores the original-image views. Source-image-group intervals for the original-minus-rotated-mean differences are given in the accompanying text.

Table 40 separates the subset admitted and determinate under all orders from each order's own subset. On the common subset, rotated-order AUROCs remain 0.832–0.856 on 7B and 0.688–0.727 on 3B, below their original-order values. The original-minus-rotated-mean difference is $+0.0243$ $[-0.0048, +0.0573]$ on 7B and $+0.0395$ $[+0.0028, +0.0791]$ on 3B. The latter indicates a decrease. The former spans zero but exceeds the prespecified 0.02 point-estimate tolerance, leaving position insensitivity unresolved. Labels agree across all orders for $197/203$ and $180/193$ common items. Whether the target occupies A changes discrimination by $+0.0089$ $[-0.0228, +0.0420]$ on 7B and $+0.0284$ $[-0.0195, +0.0736]$ on 3B, both imprecise. The original order defines the sampling population, so these differences combine reordered responses with asymmetric selection. Common subsets are nested within the selected items. Coverage is limited to these two cohorts and cyclic orders, and confidence intervals use source-image-group bootstrap with $4{,}000$ draws, seed 20260919, without multiplicity adjustment.

A further analysis on the same items separates the two components of rotation. Swapping only the two distractors gives differences not separated from zero (7B -0.003 $[-0.009, +0.004]$; 3B $+0.000$ $[-0.010, +0.011]$). Swapping the letters of the original answer and target lowers AUROC on 7B ($+0.037$ $[+0.003, +0.077]$) and is imprecise on 3B ($+0.017$ $[-0.026, +0.062]$). Each difference is original-order minus reordered AUROC on items admitted and determinate under both orders (244 per model for distractor swaps, and 222 and 205 for pair swaps on 7B and 3B). All $1{,}000$ runs complete, and the bootstrap settings match the rotation analysis. Each swap uses its own nested common subset and retains the original-order selection asymmetry. These comparisons locate the 7B sensitivity in the letters of the competing answers.

I Component Sensitivity

This appendix gives the mask-fill, edit-realization, and region analyses of Section 4.7. The placebo-geometry restriction is in Appendix C.5, and the view and coefficient analyses are in Appendix E.

I.1 Alternative Mask Fills

The main score fills R with uniform gray (Section 3.2). A pre-registered ablation replaces this fill with a Gaussian blur, matched noise, or SDXL inpainting, and recomputes GLANCE, masked-greedy, and the matched-probe increment on the six V7W admission cohorts with the formula, labels, and visible views unchanged. The plan fixed three criteria for every alternative fill: GLANCE above 0.5, GLANCE above masked-greedy, and a positive masked-support increment, each judged by its

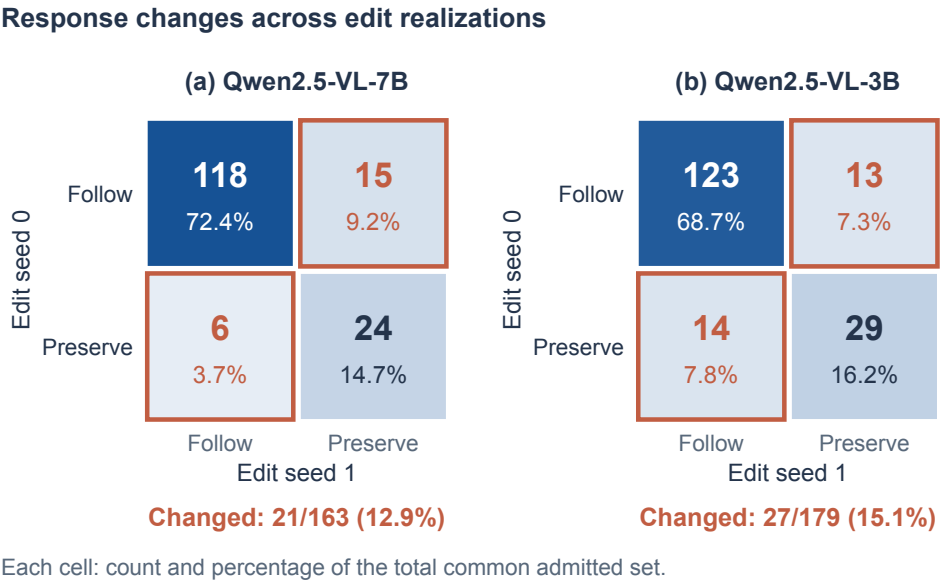


Figure 9: Response changes across edit realizations. Each matrix compares seeds 0 and 1 on the common set admitted under both edits, retaining discordant labels. Cells show raw counts and percentages of the total set on a shared color scale, and orange frames mark changed responses. The 7B and 3B sets contain 163 and 179 items (163 and 178 source-image groups), with 21/163 and 27/179 response changes. These descriptive fractions concern the doubly admitted sets. Original-image letter scores are shared across seeds.

interval. All six cohorts meet all three criteria under every fill (Table 41). Relative to gray, 16 of the 18 cohort–fill differences are not separated from zero. GLANCE rises on Qwen2.5-VL-3B with blur and falls on Qwen3.6-27B with inpainting, where it remains at 0.798.

Inpainting does not only remove content. We checked 20 randomly drawn inpainted items with a frontier VLM, Claude, rather than with human raters. Answer-relevant content was regenerated or retained in 5 items, partly retained in 4, changed into a distractor attribute in 1, and removed in 10, so the inpainting column partly measures content that the generator puts back.

I.2 Edit Realization

Changing the edit seed can change a follow/preserve label even when the question, original answer, target, and original-image scores remain fixed. We retain the common set admitted by each seed's own machine gate, with determinate edited-full responses and complete analysis scores, including label-discordant items. Labels change for $21/163 = 12.9\%$ of 7B items and $27/179 = 15.1\%$ of 3B items. Both AUROC columns of Table 42 use the same original-image GLANCE scores, with labels reconstructed from each seed's own response records. Both difference intervals contain zero, so no between-seed discrimination difference is detected, while item-level label changes remain part of what the task measures. Double admission changes the common set's composition, and these two seeds do not estimate the distribution of all edit randomness. Figure 9 shows the response transitions between seeds.

I.3 Region Specification

GLANCE's three views depend on the region R, so localizing the wrong region changes what is measured. With Qwen3-VL-32B as the localizer and Qwen2.5-VL-7B as the audited model, predicted-box scores yield AUROC 0.362 $[0.262, 0.464]$ on the 123 items with reference-box IoU in $[0, 0.3)$. On the 214 items in $[0.8, 0.95)$, AUROC reaches 0.794 $[0.718, 0.863]$, close to 0.788 using reference boxes on those same items. Low overlap reverses the ranking, while the high-overlap stratum stays close to the reference-box estimate for this localizer–model pair (Figure 10). Fixing the

Table 41: GLANCE under three alternative mask fills on the six V7W admission cohorts. Every cohort meets all three pre-registered criteria under every fill (GLANCE above 0.5, GLANCE above masked-greedy under the mixed-pool definition below, and a positive masked-support increment). Only 2 of the 18 differences from gray are separated from zero, one positive (Qwen2.5-VL-3B, blur) and one negative (Qwen3.6-27B, inpainting). Only the pixels inside R change; the formula, the labels, and the visible views are fixed.

Fill	GLANCE [95% CI]	GLANCE − gray [95% CI]	GLANCE − masked-greedy (mixed pools) [95% CI]	Masked-support increment [95% CI]
Qwen2.5-VL-7B ($n = 612$)				
gray	0.812 [0.773, 0.847]	—	+0.133 [+0.101, +0.165]	+0.180 [+0.125, +0.235]
blur	0.807 [0.768, 0.843]	−0.0047 [−0.0172, +0.0077]	+0.151 [+0.119, +0.183]	+0.174 [+0.119, +0.230]
noise	0.813 [0.776, 0.848]	+0.0010 [−0.0136, +0.0151]	+0.133 [+0.102, +0.163]	+0.181 [+0.128, +0.237]
inpainting	0.788 [0.749, 0.825]	−0.0238 [−0.0554, +0.0084]	+0.114 [+0.083, +0.144]	+0.152 [+0.100, +0.204]
Qwen2.5-VL-3B ($n = 688$)				
gray	0.798 [0.761, 0.834]	—	+0.154 [+0.117, +0.189]	+0.140 [+0.095, +0.187]
blur	0.815 [0.779, 0.849]	+0.0166 [+0.0005, +0.0344]	+0.163 [+0.127, +0.199]	+0.155 [+0.111, +0.203]
noise	0.797 [0.759, 0.833]	−0.0015 [−0.0151, +0.0127]	+0.145 [+0.108, +0.181]	+0.137 [+0.092, +0.184]
inpainting	0.790 [0.751, 0.826]	−0.0084 [−0.0385, +0.0209]	+0.152 [+0.119, +0.184]	+0.124 [+0.081, +0.168]
LLaVA-OneVision-7B ($n = 548$)				
gray	0.797 [0.751, 0.843]	—	+0.159 [+0.127, +0.192]	+0.200 [+0.142, +0.258]
blur	0.813 [0.769, 0.854]	+0.0160 [−0.0088, +0.0440]	+0.156 [+0.121, +0.192]	+0.216 [+0.157, +0.272]
noise	0.803 [0.758, 0.846]	+0.0059 [−0.0097, +0.0210]	+0.144 [+0.112, +0.176]	+0.210 [+0.153, +0.268]
inpainting	0.789 [0.747, 0.830]	−0.0077 [−0.0492, +0.0353]	+0.160 [+0.128, +0.193]	+0.195 [+0.146, +0.244]
Qwen3-VL-32B ($n = 200$)				
gray	0.810 [0.735, 0.878]	—	+0.154 [+0.099, +0.208]	+0.198 [+0.126, +0.272]
blur	0.834 [0.761, 0.896]	+0.0237 [−0.0164, +0.0678]	+0.204 [+0.144, +0.267]	+0.210 [+0.139, +0.283]
noise	0.817 [0.737, 0.886]	+0.0068 [−0.0281, +0.0415]	+0.164 [+0.112, +0.216]	+0.195 [+0.121, +0.272]
inpainting	0.819 [0.750, 0.880]	+0.0091 [−0.0529, +0.0730]	+0.142 [+0.089, +0.191]	+0.192 [+0.116, +0.266]
Gemma-3-12B-it ($n = 512$)				
gray	0.707 [0.660, 0.753]	—	+0.048 [+0.011, +0.085]	+0.134 [+0.082, +0.186]
blur	0.736 [0.689, 0.781]	+0.0289 [−0.0002, +0.0595]	+0.048 [+0.012, +0.083]	+0.162 [+0.109, +0.216]
noise	0.711 [0.663, 0.757]	+0.0044 [−0.0301, +0.0373]	+0.054 [+0.016, +0.091]	+0.134 [+0.080, +0.189]
inpainting	0.726 [0.682, 0.768]	+0.0192 [−0.0274, +0.0648]	+0.058 [+0.024, +0.092]	+0.147 [+0.095, +0.198]
Qwen3.6-27B ($n = 568$)				
gray	0.832 [0.793, 0.870]	—	+0.165 [+0.129, +0.199]	+0.243 [+0.184, +0.301]
blur	0.823 [0.782, 0.862]	−0.0095 [−0.0275, +0.0067]	+0.175 [+0.138, +0.210]	+0.232 [+0.174, +0.291]
noise	0.842 [0.803, 0.879]	+0.0093 [−0.0033, +0.0216]	+0.171 [+0.136, +0.204]	+0.252 [+0.194, +0.310]
inpainting	0.798 [0.759, 0.834]	−0.0341 [−0.0676, −0.0015]	+0.126 [+0.091, +0.159]	+0.206 [+0.149, +0.262]

Blur applies a Gaussian blur of radius $\max(2, \lfloor \min(w, h)/8 \rfloor)$ inside R; noise draws independent Gaussian pixels with the per-channel mean and variance of the pixels outside R; inpainting regenerates R with SDXL from the prompt "background" using an editor version that avoids the black-bar defect of Appendix C.6, and pastes back only the pixels inside R. Qwen3-VL-32B uses a stratified 200-item subsample fixed by the pre-registered compute rule, so its gray row differs from Table 1. GLANCE, its differences from gray, and the masked-greedy differences resample source-image groups $4{,}000$ times with the same draws for every fill. The pre-registered GLANCE − masked-greedy column compares GLANCE on all cohort items (the stated subsample for 32B) with masked-greedy on strictly parsed items; it is a mixed-pool comparison, not the same-item paired comparison in Figure 4, so gray rows can differ. The masked-support increment is the matched-probe Δ_{visible} on fixed out-of-fold predictions; for 7B, 3B, and LLaVA, it uses the full admission cohorts rather than the source-excluded probe lists in Table 15. A post-hoc re-analysis that restricts GLANCE to the same strictly parsed items keeps all 24 masked-greedy differences above zero, with Gemma-3-12B-it lower bounds of $+0.0015$ to $+0.0213$. Differences from gray keep four decimals because they are small relative to three-decimal rounding. Intervals are unadjusted; Qwen3.6-27B runs in non-thinking mode.

Table 42: Edit repeatability on the items admitted under both edit seeds (Qwen2.5-VL). Labels change for 12.9% and 15.1% of items, yet the follow-versus-preserve AUROC computed with each seed's labels differs by an amount whose interval includes zero on both models.

Model	Common n	Label agreement	Cohen's κ	AUROC, seed 0 / 1	Δ [95% CI]
7B	163	0.871	0.616	0.823 / 0.840	+0.017 [−0.072, +0.108]
3B	179	0.849	0.583	0.903 / 0.856	−0.048 [−0.116, +0.022]

Models are Qwen2.5-VL-7B and 3B. Δ is the seed-1 minus seed-0 AUROC from unrounded values, with paired source-image-group bootstrap intervals ($B = 4{,}000$, seed 20260912); the 7B/3B sets contain 163/178 source-image groups. κ measures agreement between the seed-specific labels.

localizer gives one observation per item. The strata are defined by IoU with the reference box, which is unavailable when no reference region exists.

The lowest-IoU interval lies wholly below chance, indicating reversed ranking on that stratum rather than only weaker discrimination. One possible explanation, which the observational comparison does not isolate, is that an incorrect box removes unrelated content while leaving the relevant evidence visible. Figure 10 reports same-item reference-box point estimates in each stratum, separating changed box inputs from differences in the items being compared.



Figure 10: Region specification changes discrimination in two observed IoU strata: low overlap between the predicted and reference regions reverses the ranking, while the high-overlap stratum stays close to the reference-region estimate. Qwen3-VL-32B supplies the predicted region and Qwen2.5-VL-7B is the audited model, with one observation per question. Predicted-region estimates have unadjusted 95% item-bootstrap intervals (4,000 draws, seed 0). Open squares are same-item reference-region point estimates, for which reference and paired-difference intervals are unavailable. IoU requires a reference region, and two discrete strata do not define a usable threshold.

Practical use also depends on region inputs and access to candidate likelihoods. Within the Qwen2.5-VL-7B/3B V7W admission cohorts, replacing the scoring region with boxes proposed by Qwen3-VL-32B or InternVL3-8B (OpenGVLab, 2025), while retaining labels, changes AUROC by approximately -0.13 to -0.08 relative to reference boxes, and using the audited model's own boxes gives changes of approximately -0.15 and -0.17. External localizers reduce this replacement cost without recovering reference-box performance. These comparisons retain whole-image fallbacks for parsing or geometry failures and the admission cohorts established with the reference regions, so the replacement cost is a discrimination loss on that fixed population.

J Other Uses of the Score

This appendix gives the details behind the remarks on other uses of the score in Section 5. GLANCE ranks follow above preserve among originally correct answers whose edited crops elicit the target

and whose placebos remain stable. Predicting a response to a regional edit differs from predicting arrival at the designated target after a change, and target-selection discrimination is not established in any tested cell (Appendix H). Auditing an existing answer, selecting another answer, and converting a score into a training reward therefore each require evaluation against their own objectives. The measured quantity is a conditional behavioral ranking rather than a correctness calibration or a description of how an answer is produced.

J.1 Answer Selection

Directly selecting an answer by its regional support contrast loses accuracy in the tested V7W setting. With letter scoring for every candidate, selecting the largest GLANCE score yields accuracy 0.837 on $1{,}200$ V7W multiple-choice items for Qwen2.5-VL-7B, below 0.920 from selecting the largest full-image support. On the corresponding $1{,}200$ 3B items, accuracies are 0.799 and 0.907. This comparison retains all multiple-choice items without an original-correctness gate, and both selectors operate in the same letter-score space. For this direct argmax selector, emphasizing regional responsiveness reduces answer accuracy relative to full-image support.

J.2 Training Reward

Group relative policy optimization (GRPO) (Shao et al., 2024) experiments did not establish consistent improvement from turning GLANCE into a reward. The threshold gate was precomputed from a frozen base model's candidate-text scores, using the training-item median GLANCE and rewarding correct completions only on admitted prompts. Across Qwen2.5-VL-7B/3B on multiple-choice questions from GQA and from V* and HR-Bench (Hudson & Manning, 2019; Wu & Xie, 2024; Wang et al., 2025), threshold models' mean occlusion-invariance rates did not fall below those of the base models, and the direction relative to correctness-only reinforcement learning with verifiable rewards was not consistent. Here invariance is the fraction of each model's originally correct items retaining the original option after masking, which differs from the present edit-response labels. The gate is constant across correct completions of the same prompt, so it determines which prompts receive reward without distinguishing how the same correct answer was obtained. Training budgets and evaluation coverage, particularly the small V* and HR-Bench evaluation, limit inference.

Both negative results concern objectives other than conditional follow/preserve ranking. A low-support distractor can have a contrast near zero, and full-context support that is useful for correctness can be penalized because GLANCE subtracts the support that persists after masking. These are possible reasons to avoid treating the score as a calibrated answer-quality reward. Whether a completion-dependent objective would help is left open and would require a separate learning experiment.

J.3 Audit Triage

Pre-edit scores can prioritize which items receive audit or editing effort. Precision@k is the follow fraction among the highest-scoring $k\%$ of items, and skip@r is the largest fraction removed consecutively from the low-score end while retaining a follow recall of at least r. Table 43 reports precision@20% and skip@0.90. The local comparator is the in-sample AUROC winner among the nine slots: whole-image swap for 7B and LLaVA, negative option entropy for 3B, Gemma, and 27B, and self-verify (Yes weight) for 32B. These operating points reinterpret the T1 ranking summarized by AUROC and add no independent discrimination evidence, and the single threshold rule has no external calibration. Triage leaves the answer unchanged and allocates audit effort, so these results coexist with the negative answer-selection result above.

Table 43: Retrospective pre-edit audit triage within the six V7W admission cohorts. When items are ranked by GLANCE, follow precision among the highest-scoring 20% exceeds 0.90 in every cohort, well above each cohort's follow base rate, and the fraction of items skippable at 90% follow recall exceeds the 0.10 expected from random screening in every cohort. These operating points reinterpret the same ranking as AUROC and add no independent evidence.

A. Follow precision among the highest-scoring 20%

Audited model	Follow base rate	GLANCE precision [95% CI]	Slot precision
Qwen2.5-VL-7B	0.711	0.943 [0.893, 0.975]	0.705
Qwen2.5-VL-3B	0.718	0.942 [0.899, 0.978]	0.746
LLaVA-OneVision-7B	0.745	0.918 [0.873, 0.973]	0.782
Qwen3-VL-32B	0.736	0.908 [0.850, 0.958]	0.775
Gemma-3-12B-it	0.672	0.902 [0.824, 0.951]	0.716
Qwen3.6-27B	0.745	0.965 [0.921, 0.991]	0.807

B. Items skipped at 90% follow recall

Audited model	GLANCE skip fraction [95% CI]	Preserve purity	Preserve coverage	Slot skip fraction
Qwen2.5-VL-7B	0.206 [0.178, 0.242]	0.659	0.469	0.101
Qwen2.5-VL-3B	0.209 [0.183, 0.237]	0.660	0.490	0.122
LLaVA-OneVision-7B	0.206 [0.172, 0.237]	0.646	0.521	0.115
Qwen3-VL-32B	0.216 [0.189, 0.244]	0.662	0.541	0.103
Gemma-3-12B-it	0.137 [0.106, 0.168]	0.514	0.214	0.104
Qwen3.6-27B	0.218 [0.187, 0.252]	0.661	0.566	0.099

Cohort sizes are 612, 688, 548, 602, 512, and 568 in row order. Preserve purity is the fraction of skipped items labeled preserve; preserve coverage is the fraction of all preserve items skipped. Uniform random screening has expected skip fraction 0.10 at follow recall 0.90. The slot comparator is the in-sample AUROC winner among the nine region-blind slots and is optimistic. Brackets are 95% item-bootstrap intervals (4,000 draws, seed 0), unadjusted. These retrospective operating points condition on admission and determinate edited responses and do not estimate performance in use.