FOREGROUND-MATCHED OCCLUSION CONTROLS FOR AUDITING ANNOTATED-REGION SENSITIVITY IN MEDICAL VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Occlusion audits of medical vision-language models (VLMs) read a larger answer shift from blacking out an annotated region than from a same-sized control box as use of the region, assuming both occlusions are equally strong. A black fill on air changes almost no pixels, so a control there loses to nearly any tissue occlusion. A minimum-overlap control could favour air (replayed on 229 items, its mean tissue coverage was 0.659 against 0.943 matched), and under it every model run appeared region-sensitive. We introduce a *foreground-matched* control (same size, matched tissue coverage, no overlap with any annotated box) and a proposition showing, under a monotone foreground-dose model, that a comparator covering less tissue biases the win rate upward and matching bounds that term. On SLAKE the matched control gives graded sensitivity: both MedGemma checkpoints lead (strict-win rates 0.639 and 0.612) and survive post-hoc resampling of 175 source images, while four general models are lower (0.525–0.563) and their detection mostly depends on how ties are scored. The scope is as follows: historical and matched runs are unpaired except for MedGemma-4B, whose CIPE win rate is marginally lower when matched, so we do not attribute their difference to the control; 93.0% of the 229 items sit on images the MedGemma family reports training on, so the ranking is exposure-confounded and is not read as specialisation; and our CIPE readout showed no advantage over cheaper ones in a pre-registered head-to-head. The masked boxes annotate the image, not the question target, so this is image-level sensitivity, not target localisation, for one specialist family at two scales on the 229 of 400 SLAKE items admitting a matched control, which favour sparsely annotated images. We also build **MedTrust-IN v0.3** (released upon publication): $1{,}122$ panel-defined *text-only-hard* medical-VQA items from source *test* splits, with a fail-closed build-time provenance gate. An audit should report and match what its control covers, drop and count unmatched items, and report ties at half credit beside strict wins.

1 INTRODUCTION

To test whether a medical vision-language model (VLM) uses an annotated region, an audit blacks it out and asks whether the answer moves more than when a same-sized box elsewhere is blacked out, as Sadanandan & Behzadan (2026) do for one model with the radiologist box against a random same-size region. This presumes both occlusions are equally strong. A black fill on the background air around the body changes almost no pixels, so a comparison box there loses to nearly any occlusion of tissue, whatever the model relies on; where the control lands can decide the audit (Fig. 1).

Occlusion is one way to ask whether a VLM used in medicine (OpenAI, 2025; Gemini Team, 2025; Li et al., 2023a) answers from the image or from language priors, and a model that answers from priors fails in exactly the cases where the image matters. The comparison is usually a random or blank region or the unperturbed input (§2); what the comparison region covers is left uncontrolled. Our first design, a minimum-overlap placement, kept per annotated box the random placement with least overlap, which can favour background air (replayed on MedGemma-4B's 229 items, it covered much less tissue than the matched control; §3.1) while the organ box lies on tissue by construc-

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.

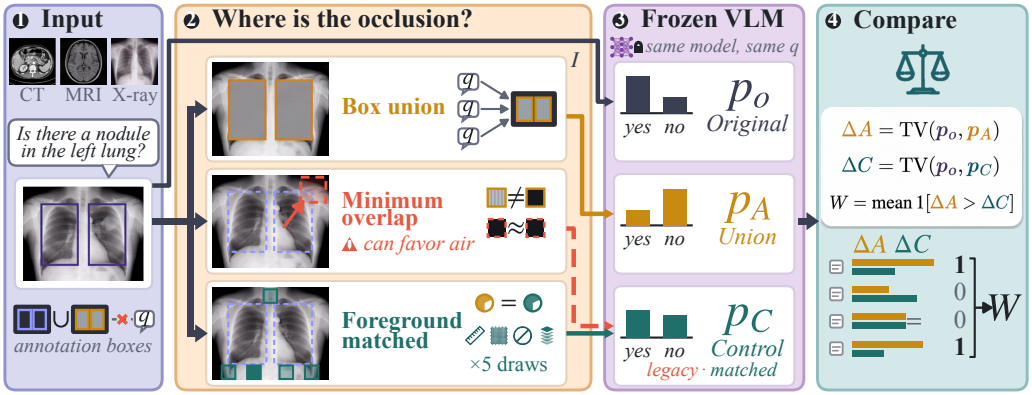

Figure 1: **An occlusion audit is a comparison, so what the control covers is part of the measurement.** One image I and closed question are held fixed; only the occluded region changes. A frozen VLM answers on the original image (p_o), with the union of all image-level annotation boxes blacked out (p_A), and with a control blacked out (p_C): the legacy minimum-overlap placement, which can land on background air (replayed on the 229 items it covered 0.659 tissue on average against 0.943 matched), or the foreground-matched one (per annotated box, a same-size box of matched tissue coverage overlapping no annotated box; five draws). W averages $\mathbf{1}[\mathrm{TV}(p_o, p_A) > \mathrm{TV}(p_o, p_C)]$ over draws (a tie scores 0 in W and $\frac{1}{2}$ in W^*) and measures image-level sensitivity, not target localisation. Under the matched control both MedGemma checkpoints lead ($W = 0.639$, 0.612; exposure-confounded) and four general models trail (0.525–0.563; Table 1); the general-model control cohorts are separate runs, aligned on only 34 QA. Schematic: the X-ray, CT and MRI images are generated illustrations, box sizes are not to scale, and the coverage pies, answer histograms and item bars carry no values; evidence is in Figures 2 and 3.

tion, and under it every model run appeared region-sensitive under CIPE, the Counterfactual Image Perturbation Effect, a per-sample KL readout we designed (Appendix Table 18).

Replacing that comparator with a *foreground-matched* one (same size as the annotated box, tissue coverage within 0.15 of it, no overlap with any annotated box) makes the effect graded. On SLAKE, occluding the annotated union moves the answer distribution more than matched tissue in 0.639 of item–control comparisons for MedGemma-27B and 0.612 for MedGemma-4B, the two strongest of six models, and both effects survive resampling whole images. General models are lower: their strict-win rates sit slightly above one half, and post-hoc half-credit tie scoring detects positive W^* effects in all four (Table 1). SLAKE's boxes annotate the image rather than the question target, so this is image-level annotated-region sensitivity, not target localisation. Proposition 1, an argument about a model rather than a measurement, says why matching is needed: if the shift grows with occluded foreground, a comparator covering less tissue pushes the win rate above one half with no annotation-specific effect.

Four readings are not supported. For the general models the difference from the legacy numbers cannot be credited to the control: their legacy and matched runs used different QA samples and readouts, only 34 shared QA are recoverable, and on them only Qwen2-VL's corrected interval for the paired contrast excludes zero (§4.1). On the 229 QA MedGemma-4B's legacy run shares, its CIPE win rate is lower with five matched draws than with one legacy draw, a marginal one-model CIPE contrast whose interval barely excludes zero and that does not isolate coverage (Appendix S4.1); MedGemma-27B never ran under the legacy control. The MedGemma lead is not evidence of specialisation: nearly all matched-control items sit on SLAKE train-split images that the family's own technical report says it trained on, the splits are not image-disjoint, and no subset of this dataset separates specialisation from exposure, so the ranking is exposure-confounded (Appendix S1). A ChestX-Det run that aligns the occlusion with the queried finding is hypothesis-generating only, because both admitting amendments were made after the values were seen (§4.1). Finally, a pre-registered head-to-head that named its failure branch in advance found CIPE significantly ahead in no model and a cheaper readout significantly ahead in two of six (§4.3).

Contributions.

1. **A reusable foreground-matched control with a stated estimand.** It needs only region boxes and an intensity threshold, so an auditor can match controls on size and tissue coverage and report both arms' coverage; it comes with the win rate W, its tie-aware form W^*, the mean shift difference D (§3), and Proposition 1, which says when a comparator can bias W and what matching bounds.
2. **A graded, exposure-confounded measurement** on the $229/400$ SLAKE items that admit a matched placement, a subset that selects sparsely annotated images, with image-clustered and tie-aware intervals (§4.1, Table 1).
3. **Four audit failure modes, each with a prescribed check**: answer extraction, image injection, metric–stratifier circularity and control-region composition; all four were observed in this audit's own pipeline (§4.2, §4.4, Appendix S6); CIPE, the readout we designed, is tested against simpler readouts and total variation is retained (§4.3).
4. **MedTrust-IN v0.3**: 1,122 source-*test*-split items that a positive-controlled panel rarely answers without the image, with per-item screening provenance, a generated datasheet and a fail-closed source-split gate; the label *panel-defined text-only-hard* does not assert that the image is sufficient or that a human adjudicated visual necessity.

2 Related Work

Medical VQA benchmarks and trust evaluation. Medical VQA is mostly scored by accuracy on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020) and OmniMed-VQA (Hu et al., 2024). Trust-oriented benchmarks add axes around accuracy: CARES (Xia et al., 2024) covers trustfulness, fairness, safety, privacy and robustness; MediConfusion (Sepehri et al., 2025) pairs images that a vision encoder confuses although a clinician would not; ProbMed (Yan et al., 2025) pairs each question with adversarial negatives; and general frameworks such as MultiTrust (Zhang et al., 2024) and HELM (Liang et al., 2023) do the same outside medicine. These benchmarks score behaviour on fixed or globally altered inputs. We ask a per-item question instead—does the answer move more when a specific annotated region is removed than when comparable tissue is?—and treat the choice of comparison region as part of the estimand.

Occlusion, perturbation and counterfactual probes. Occluding part of an input and reading the change in output is among the oldest attribution tools (Zeiler & Fergus, 2014). Meaningful perturbation (Fong & Vedaldi, 2017) and ROAR (Hooker et al., 2019) showed that the deletion itself can move the output because it produces out-of-distribution inputs, and sanity checks (Adebayo et al., 2018) showed that an attribution can look plausible while ignoring the model; counterfactual infilling replaces the blank occluder with generated content (Chang et al., 2019). ROAD (Rong et al., 2022) showed that perturbation evaluations can be driven by information leaking through the shape of the removed pixels rather than their values, and the choice of baseline input changes integrated-gradients attributions (Sturmfels et al., 2020); we study the analogous choice for the comparison region. In VQA, FPVG (Reich et al., 2023) and modality-contribution scores (Gat et al., 2021; Parcalabescu & Frank, 2023) test whether answers depend on relevant visual evidence; HEAL-MedVQA (Nguyen et al., 2025) and concurrent medical audits (Zafar et al., 2026b; Sadanandan & Behzadan, 2026; Jiang et al., 2026) carry such tests to medical VLMs: Zafar et al. (2026b) compare real with blank and shuffled images, and Sadanandan & Behzadan (2026), closest in instrument, score attention against radiologist boxes versus shifted and random boxes and, for one model, occlude the radiologist box against a random same-size region; neither describes matching what the control region covers to the annotated region. The comparison is usually a random or blank region or the unperturbed input. We argue that the comparator's *content* is a potential confound—if a box on air is a weaker intervention than one on tissue, a background comparator inflates W—and bound what foreground matching leaves.

Shortcut learning and language priors. VQA models answer from question priors (Goyal et al., 2017; Agrawal et al., 2018), and counterfactual re-splits show the same in medical VQA (Zhan et al., 2023; Wan et al., 2025). Blind evaluation finds that many frontier models lose little when the image is removed (Felizzi et al., 2025; Asadi et al., 2026; Lotfinia et al., 2026; Pokrywka et al., 2026),

many benchmark items are answerable without it (Chen et al., 2024), and radiograph classifiers can succeed through confounds rather than pathology (Geirhos et al., 2020; DeGrave et al., 2021). MedTrust-IN uses the same text-only signal as a screen, and we show that such a screen inherits its panel's answer prior. On the measurement side we document a readout shortcut: fixed single-token answer extraction (Duan et al., 2024; Biderman et al., 2024) can make an image-reading model look image-invariant.

Contamination and exposure in medical VLM evaluation. Benchmark items that leak into training inflate scores in ways accuracy alone cannot separate from ability (Chen et al., 2024). For medical VQA, Xu et al. (2026) audit provenance and find likelihood-based membership detectors unreliable on small cohorts, and model reports are a second source: MedGemma's report lists SLAKE in its training data and evaluates on SLAKE's default train/test split (Sellergren et al., 2025). We use both routes—a train-versus-test image-benefit probe anchored to a documented training corpus, and a split-membership audit of the region items—and the second is why we withdraw the specialisation reading of our region result.

3 Problem Setup

Answer distributions and occlusions. A model m receives an image I and a closed question q. One forward pass at temperature 0 gives the log-probabilities of the two answer tokens, which we renormalise to $p_o = P_m(\cdot \mid I, q)$ on $\mathcal{A} = \{\text{yes}, \text{no}\}$. For a region R, $O_R(I)$ fills R with black. On SLAKE item i, A_i is the union of every box in the image's `detection.json` (image-level annotations, not the question target) and $C_{i1}, \ldots, C_{iK}$ are K=5 seeded control placements. With $p_{A,i} = P_m(\cdot \mid O_{A_i}(I_i), q_i)$ and $p_{C,ik} = P_m(\cdot \mid O_{C_{ik}}(I_i), q_i)$,

$$\Delta_{A,i} = \mathrm{TV}(p_{o,i}, p_{A,i}), \qquad \Delta_{C,ik} = \mathrm{TV}(p_{o,i}, p_{C,ik}), \qquad \mathrm{TV}(p, p') = \tfrac{1}{2}\textstyle\sum_{a \in \mathcal{A}} |p(a) - p'(a)|, \tag{1}$$

so with two answers Δ is the absolute change in $p(\text{yes})$.

Estimands. Over the n items that admit a control,

$$W = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{k=1}^{K}\mathbf{1}[\Delta_{A,i} > \Delta_{C,ik}], \qquad W^* = W + \frac{1}{2nK}\sum_{i,k}\mathbf{1}[\Delta_{A,i} = \Delta_{C,ik}], \tag{2}$$

$$D = \frac{1}{n}\sum_{i=1}^{n}\Big(\Delta_{A,i} - \frac{1}{K}\sum_{k=1}^{K}\Delta_{C,ik}\Big). \tag{3}$$

W is the registered strict-win rate (a tie counts zero), W^* gives ties half credit, and D is the mean shift difference. The indicator is averaged over draws rather than compared with the draw-mean shift, which is what gives W a chance reference (Proposition 1(a)). The registered analysis resamples items; the post-hoc analysis resamples source images with all their questions (229 questions come from 175 images).

Two controls. Both use the fill and box sizes of the annotated boxes. The *minimum-overlap* (legacy) control draws, for each annotated box, up to 30 uniform same-size placements and keeps the first with zero (else least) overlap with that box alone, one draw per item; nothing constrains what it covers. The *foreground-matched* control places, for each annotated box b, a box c of the same size that overlaps no annotated box and no other control box, and whose foreground fraction satisfies $|\phi(c) - \phi(b)| \le 0.15$, where $\phi(r)$ is the share of pixels in r above the image minimum by 5% of the dynamic range (an intensity criterion: any non-background pixel counts, including a scanner table or burned-in text). Placement is rejection sampling; an item with fewer than five admissible draws is dropped, never given a fallback box. 229 of 400 candidates admit placement; mean coverage is 0.961 for annotated boxes and 0.943 for controls. Because both arms then black out nearly the same share of tissue, the out-of-distribution shift that deleting tissue itself causes (Hooker et al., 2019) is present in both arms, not in the annotated arm alone. The tolerance 0.15 and the 5% threshold were set once in the placement script, never changed, and are recorded in all 11 placement receipts; tightening the tolerance to 0.05 on the stored draws leaves both MedGemma W^* intervals above one half (§4.1), while the threshold, looser tolerances and the black fill itself are untested.

Panel readout (CIPE). For the model panel we also score three whole-image counterfactuals—`image_swap` (a random same-dataset image), `anatomy_mask` (a centred 25%-area square in every panel run, so it carries no region claim) and `text_only` (image dropped, "no image provided" sentinel)—by $\mathrm{CIPE} = 1 - \mathrm{clip}(\mathrm{KL}(p_o \,\|\, p_{\mathrm{cf}}), 0, 1)$ in nats. CIPE is *high* when the intervention has no effect. The panel covers 50 unique open-source VLMs (63 valid evaluations) in eleven families (e.g. Wang et al., 2024; Bai et al., 2025b; Zhu et al., 2025; Gemma Team, 2025; all in Appendix S16) plus MedGemma-4B/27B-it (Sellergren et al., 2025); per-item validation uses 21 separate n=500 model×dataset re-runs, and 4 of those 21 evaluations fail the positive control and are excluded, leaving 17 (Appendix S13).

3.1 Foreground matching of the comparator

To see why the comparator's content matters, model, for one item, the shift caused by occluding a region R as

$$\Delta(R) = h\big(\phi(R)\big) + \tau\,\mathbf{1}[R = A] + \varepsilon_R, \tag{4}$$

where h is nondecreasing (covering more tissue moves the answer at least as much), τ is the annotation-specific effect the audit is after, and the noise pair $(\varepsilon_A, \varepsilon_C)$ is exchangeable and independent of the placement, with $\xi = \varepsilon_C - \varepsilon_A$ having a continuous CDF F (symmetric, so $F(0) = \frac{1}{2}$). A black box on air is close to the identity, so a control there has $h(\phi(C)) \approx h(0)$.

Proposition 1. (a) Ties. *If (Δ_A, Δ_C) is exchangeable, then* $\Pr(\Delta_A > \Delta_C) = \frac{1}{2}\big(1 - \Pr(\Delta_A = \Delta_C)\big)$*; hence* $\mathbb{E}W \le \frac{1}{2}$ *and* $\mathbb{E}W^* = \frac{1}{2}$.

(b) Background bias. *Under* (4)*,* $\Pr(\Delta_A > \Delta_C \mid C) = F\big(\tau + h(\phi(A)) - h(\phi(C))\big)$*. If* $\phi(C) \le \phi(A)$ *this is at least* $F(\tau)$*, strictly when* $h(\phi(A)) > h(\phi(C))$ *and F is strictly increasing; with* $\tau = 0$ *and those conditions, the annotated region beats a background control with probability above* $\frac{1}{2}$.

(c) Matching. *If* $|\phi(C) - \phi(A)| \le \delta$*, h is L-Lipschitz and F has density at most* $\bar{f}$*, then* $|\Pr(\Delta_A > \Delta_C \mid C) - F(\tau)| \le \bar{f} L \delta$*. For* $\delta = 0$ *and F strictly increasing, the annotated region wins with probability above* $\frac{1}{2}$ *if and only if* $\tau > 0$.

Part (b) is the failure mode our first design was open to: a control that removes less tissue can push W above one half with no annotation-specific effect. Replayed from its seed on CPU, our legacy placement covered less tissue on the same 229 items (mean coverage 0.659 against 0.943 matched and 0.961 annotated; Appendix S4.2), so this premise held on average for our first control, the gap sitting in CT and MRI, with X-ray legacy near matched (a descriptive, unregistered split); the replay does not say how much of the legacy win rate it produced. Part (a) is why we report W^* and treat 0.5 as a conservative reference for strict W. Part (c) is why matching need only be approximate; the tolerance bounds δ per box, which bounds the union's gap only when the annotated boxes are disjoint (171 of 229 items; Appendix A), and we estimate neither L nor $\bar{f}$, so it gives no numeric bound. The proposition is an argument about a model, not a measurement. It does not say how much of the historical legacy-versus-matched difference came from the control, which the unpaired runs cannot identify (§4.1), and every other difference between A and C—position, anatomical coherence—stays inside τ, so $W > \frac{1}{2}$ is not evidence of question-target reliance. Proofs are in Appendix A.

4 Results

4.1 Region sensitivity under the foreground-matched control

Under the foreground-matched control, occluding the annotated union moves the answer distribution more than matched tissue for both MedGemma checkpoints, and the effect survives resampling whole images: both image-cluster intervals keep W^* above one half and D above zero (Table 1, Fig. 2). The separation from general models is graded rather than categorical. InternVL3-8B also reaches significance under the registered item bootstrap, and once ties receive the half credit that Proposition 1(a) calls for, all four general models show weaker positive effects; all six conditional image-cluster sign-flip tests on W^* survive Holm correction (on strict W only the two MedGemma tests do), as a post-hoc sensitivity analysis (Appendix S5). Because Δ is the absolute change in

Table 1: **Region sensitivity under the foreground-matched control is graded: both MedGemma checkpoints lead, and general-model effects are weaker.** Left: SLAKE, total variation on normalised distributions, 229 QA on 175 images, five control draws per item. W is the registered strict-win rate with its item-bootstrap Benjamini–Hochberg (BH) q; W^* gives ties half credit; brackets are image-cluster 95% intervals (post-hoc, Appendix S5). The MedGemma rows are exposure-confounded: 93.0% of the 229 items sit on SLAKE train-split images the family reports training on (Appendix S1). Right: exploratory ChestX-Det yes arm (n=168), admitted by two non-blind amendments; MedGemma-27B is excluded by construction (it trained on Chest ImaGenome for anatomy localisation), and Qwen2-VL-7B's W there is uninformative (it is image-insensitive under this readout; Appendix S2). Sources: `image_cluster_sensitivity.json`, `expclean_*.json`.

	SLAKE, foreground-matched control				ChestX-Det, yes
Model	W	q	W^* [95% CI]	D [95% CI]	W (q)
MedGemma-27B	0.639	0.0003	0.661 [0.611, 0.708]	+0.059 [+0.031, +0.089]	—
MedGemma-4B	0.612	0.0003	0.657 [0.610, 0.704]	+0.051 [+0.021, +0.085]	0.742 (0.0005)
InternVL3-8B	0.563	0.030	0.595 [0.547, 0.642]	+0.011 [+0.002, +0.020]	0.626 (0.0005)
Gemma3-4B	0.545	0.099	0.567 [0.520, 0.615]	+0.007 [−0.014, +0.031]	0.526 (0.358)
Qwen2.5-VL-7B	0.536	0.178	0.559 [0.509, 0.610]	+0.011 [−0.002, +0.025]	0.607 (0.001)
Qwen2-VL-7B	0.525	0.316	0.560 [0.511, 0.609]	−0.001 [−0.005, +0.004]	0.570 (0.016)

p(yes), D says occluding the annotated union moves p(yes) on average 0.059 (MedGemma-27B) and 0.051 (MedGemma-4B) more than matched tissue; for the general models D is at most +0.011 and its interval includes zero in three of four. The two MedGemma checkpoints are a within-family repeat, not generality across specialists. Two receipt-only checks, registered as descriptive before they ran, bear on the instrument (Appendix S5.1): as a negative control, scoring one matched draw against the other four (no annotation effect by construction) gives W^* of 0.465–0.545, none of 30 intervals excluding one half; tightening the coverage tolerance to 0.05 (194 items) leaves MedGemma-4B at $W^* = 0.673$ [0.617, 0.727].

The general-model control comparison is not paired. Four general-model legacy receipts disagree with the foreground candidate sequence on most box counts at identical ordinal IDs: they are different QA samples. Replaying the original loaders recovers only 34 shared admitted QA. On them the annotated-mask CIPE agrees exactly across runs, and the paired reduction in CIPE contrast survives four-model correction only for Qwen2-VL; most images barely move it and a handful carry it, with both signs in three of four models (Fig. 3, Appendix S4). We withdraw our first analysis's "control, not the item pool" attribution and read the foreground audit as a separate result.

What the estimate describes. Only 229 of 400 candidates admit a matched placement, and they are sparsely annotated: admitted items carry 1.66 boxes on average against 4.46 for exclusions (Appendix Table 17). The MedGemma lead is a lead in point estimates: InternVL3-8B's W^* interval overlaps both MedGemma intervals, and we test no between-model difference. It also sits on images that family disclosed training on (Appendix S1). What we can state is differential answer-distribution sensitivity to the annotated union under exposure—not where a model looks, and not question-target correspondence.

Exploratory target-aligned follow-up. Under a registration committed before any file was downloaded, we ran the same frozen control and readout on ChestX-Det (Lian et al., 2021) (finding-presence items, balanced yes and no, 13 classes), changing only the loader; its admission gate and arm-level analysis were then amended after outcomes were seen, so the run is non-blind. Only the yes arm occludes the queried finding, so arms are reported separately. On the primary yes arm MedGemma-4B has the highest W of five models (Table 1, right), a ranking not attributable to any disclosed ChestX-Det inclusion; combined over both arms its W is lower, and "clean" means absent from *disclosed* training data (Appendix S2). The run motivates a frozen-protocol replication; it does not rescue the SLAKE claim, and the SLAKE withdrawal stands.

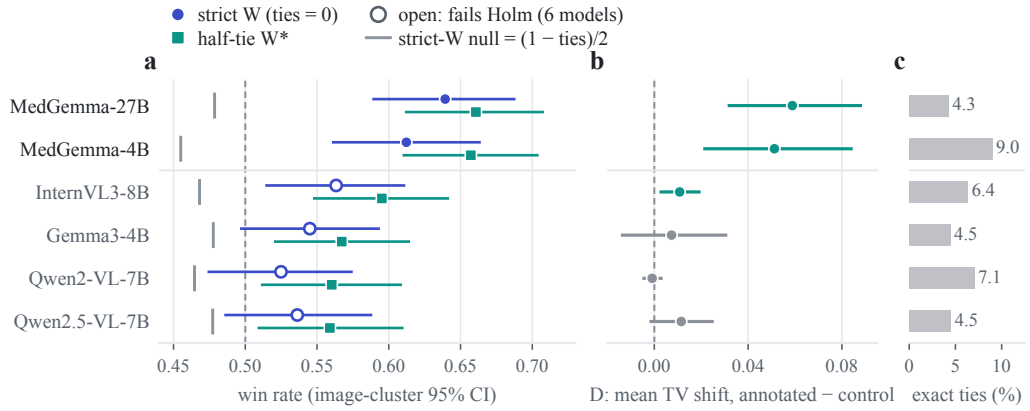

Figure 2: **Both MedGemma effects survive image clustering; general-model classifications depend on how ties are scored.** Same 229 QA on 175 images and five control draws, rows sorted by W^* (Table 1 sorts by W); intervals resample images with all their QA rows. (a) Strict W (ties $= 0$) and half-tie W^*; the grey tick is the expectation of strict W under exchangeable occlusions, $(1 - \text{tie share})/2$ (Proposition 1(a)), so the dashed 0.5 line is conservative for strict W. (b) Mean shift difference D; teal where the interval excludes zero. (c) Share of exact ties. Intervals are unadjusted; open markers fail Holm over six models: strict W passes only for the two MedGemma rows (InternVL3-8B $p_{\text{Holm}} = 0.053$), all six half-tie tests pass. This is post-hoc sensitivity, not a new confirmatory registration. Source: `image_cluster_sensitivity.json`.

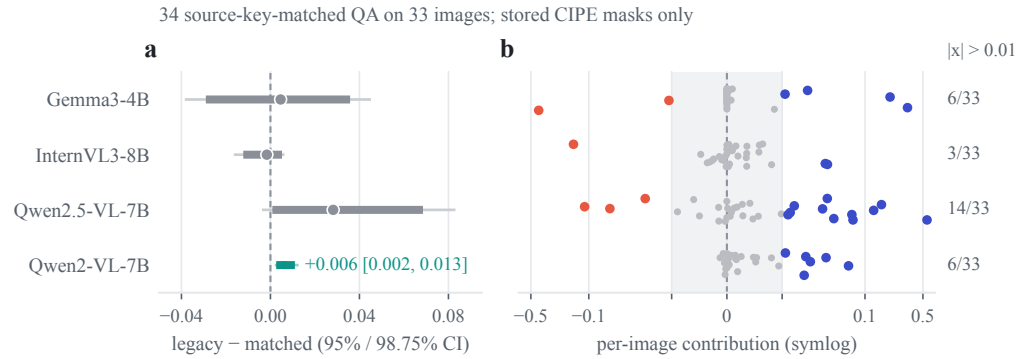

Figure 3: **The recoverable shared subset does not support a uniform control effect.** Only 34 true shared included QA (33 images) match across the historical loaders; identical ordinal IDs do not establish identity. (a) Paired stored-mask CIPE contrast $D_L - D_F$ (legacy minus matched; each is control minus annotation CIPE, not the TV-based D of Eq. 3); thick bars are image-bootstrap 95%, thin bars Bonferroni 98.75% over four models, and only Qwen2-VL, whose mean and 98.75% interval are printed, excludes zero after correction. (b) Each image's contribution on a symlog axis (indigo positive, coral negative, grey $|\cdot| \leq 0.01$), with the number of images above 0.01 in magnitude. This post-hoc intersection cannot be extrapolated to $229/400$ items or treated as a common-roster experiment. Source: `paired34_stored_mask_descriptive.json`.

4.2 Readout artefacts: the "image-invariant" cluster

A matched comparator also needs a readout that registers the answer. Our first analysis reported five models (LLaVA-1.5-7B (Liu et al., 2024a), VipLLaVA-7B (Cai et al., 2024), Idefics2-8B (Laurençon et al., 2024b), Aria-MoE-25B-3A (Li et al., 2024b), Fuyu-8B) whose CIPE was exactly 1.000 under every counterfactual, and read them as image-invariant. For the three models we could recover, a raw-logit audit shows the image is injected and causal; the zero comes from the readout. We scored the space-prefixed `yes`/ `no` tokens while these models emit `Yes`/`No`, so both scored to-

kens carry near-zero, near-equal mass, $P_m(\text{yes})$ renormalises to 0.5, and CIPE saturates whatever the image. Both obvious controls—the raw-KL signature and a black-versus-white positive control—read through the same broken readout and cannot tell this from a dead image; total variation over the full vocabulary can. A multi-surface-form readout recovers three families as weakly to moderately image-reliant, two remain unresolved, and three further saturated checkpoints were never re-audited (Appendix S29). We therefore *prescribe* running the positive control at the raw-distribution level, validating answer extraction per model, and recording an `image_injection_ok` flag that downstream consumers fail closed on—§4.4 shows what happens when they do not.

4.3 CIPE AGAINST SIMPLER READOUTS

Item-level signal is small, and a flip rate matches it. Over the 17 certified $n{=}500$ model$\times$dataset evaluations, per-item CIPE predicts answer correctness with macro-mean AUROC barely above one half, and an argmax flip rate does no worse (slightly higher, not significantly; Appendix Table 8). CIPE also ranks items almost exactly as total variation does, so the clipping and the KL asymmetry add little. The aggregate association runs counter to a simple faithfulness reading (image-invariant items are answered correctly more often); §4.4 separately shows why the apparent stratified reversal was circular (per-model tables in Appendix S16).

The pre-registered head-to-head fired its worst branch. On the one intervention with a surviving signal we registered a readout comparison that named its failure branch in advance (`PREREG_region_metric_comparison.md`). No model shows CIPE significantly ahead, and a cheaper readout significantly beats it in two of six models ($|\Delta p(\text{yes})|$ on Gemma3-4B, the flip rate on Qwen2-VL-7B; on the stored unnormalised vectors, Appendix S3). **We withdraw the distributional claim: CIPE has no demonstrated advantage over simpler readouts.** We retain total variation because it is simpler and CIPE did not significantly beat it, not because it won this comparison (Appendix S10).

No scaling claim. We regress the share of items with CIPE > 0.7 on $\log_{10}$ parameters over the 21 independently certified models, not all 40 non-saturated panel models with disclosed sizes, because only half the panel carries a raw-distribution positive control. The coefficient is negative, but significance depends on the specification: a family random intercept reaches $p{<}0.05$ and OLS with family fixed effects on the same models does not (Appendix S6). We do not claim significance.

4.4 CIRCULARITY, THE CORPUS SCREEN, AND CONTAMINATION

The apparent de-inversion is circular. We pre-registered that the construct-validation sign is a language-answerability artefact that should de-invert once items are stratified by text-only hardness, and under that stratification it does in every cell. But `text_only` is both a CIPE counterfactual and the stratifier, so it manufactures separation with no faithfulness content. Without it the effect largely collapses and on VQA-RAD it disappears; we retract the Simpson's-paradox resolution (Appendix S26).

The screen failed open, and fixing it worsened the label skew. Auditing the v0 corpus build against §4.2's prescription found 41 of 64 screening runs without an `image_injection_ok` flag, and the guard `.get(...) is False` passes when the key is absent. Where the flag was present and false, de-duplication promoted the same model's older unflagged run, so LLaVA-OneVision-7B (Li et al., 2024a)—which answers `yes` to every SLAKE item it was run on, with or without the image—still screened 296 v0 items. Re-issued as v0.2 under a fail-closed, model-level screen with a constant integer criterion, the corpus has a *worse* yes-skew and a smaller balanced subset (Appendix Table 9). Without the image the panels default to "no", so "no"-gold items are answered for free and screened out: a panel-defined text-only-hard screen inherits its panel's answer prior (Appendix S27).

Train-versus-test contamination probe. Every v0.2 item came from a source train split, so we ran the same harness on matched, disjoint train and test subsets of all three datasets: 24 (model, dataset) cells in all. One cell carries a large anomaly: SmolVLM-256M (Marafioti et al., 2025) on VQA-RAD scores far higher on train than on test, the only cell surviving BH (Appendix S12). That model's documented training mixture contains VQA-RAD's train split (100% of its question

turns) and neither other dataset, where its cells are null. The anomaly is consistent with documented training inclusion; the split contrast alone does not identify memorisation, and model-axis specificity remains untested (Appendix S12). Two screener cells, both on PathVQA, exclude zero only marginally and do not survive BH; the other 21 cells are null.

Acting on the rule: MedTrust-IN v0.3. We had pre-registered that any screener gap excluding zero triggers a rebuild from source test splits, so we re-screened the roster on the full closed yes/no test pools and rebuilt under the identical fail-closed screen. MedTrust-IN v0.3 has $1{,}122$ items, fewer than v0.2 because SLAKE's test split is small, while PathVQA grows. The yes-skew falls to 85.6%, and the label-balanced subset, the one suited to prior-sensitive analysis, grows but is almost all PathVQA, with no SLAKE items (Appendix Tables 9, 10). Test-split membership removes direct train-split selection, not exposure (Appendix S11).

5 LIMITATIONS

- *The measurement is image-level annotated-region sensitivity on SLAKE, and the family contrast is exposure-confounded.* The loader masks the union of every image-level box and does not map boxes to the question target. General-model effects are weak under strict wins but detectable with half-tie scoring; the specialist effect, observed in both MedGemma checkpoints on one dataset and the 229-item matched-placement subset, rests on *disclosed image-level training exposure* (Appendix S1). Our receipt audits split membership, not any model's training corpus, so this is disclosed exposure rather than verified memorisation.
- *The foreground argument is model-based.* Proposition 1 assumes a monotone foreground-dose shift and exchangeable noise; the matched control equalises size, coverage and non-overlap, not anatomical coherence or position; on the 58 items with overlapping annotated boxes the per-box tolerance does not bound the union's coverage gap. No item with a known annotation effect was constructed, so the instrument's power for small τ is unmeasured.
- *Control certification covers half the panel.* 21 of the 40 scaling-panel models (not the $n{=}500$ re-runs of §3) are measured and all pass; the other 19 are unreachable (evicted or unloadable), so the 40-model panel is descriptive only (App. S29).
- *Contamination probed, and a screener implicated.* One large anomaly consistent with disclosed training inclusion and two marginal screener cells on PathVQA fired our rebuild rule; v0.3 removes direct train-split selection but does not establish training-data cleanliness. The CIPE ranking is stable but not invariant across split.
- *Corpus label skew, residual de-inversion, coverage.* Text-only hardness is confounded with the answer prior (v0.3 is still mostly yes), so prior-sensitive analysis is limited to the balanced subset, which is mostly PathVQA; two readout-artefact models remain unresolved; English-language 2D imaging only, mostly general VLMs; source datasets inherit demographic and prevalence biases (Zong et al., 2023), which models can propagate (Omiye et al., 2023); nothing here evaluates clinical usability (details: Appendix S7).

6 CONCLUSION

When blacking out an annotated region moves a medical VLM's answer more than blacking out a box elsewhere, the result depends on the second box, not only on the model. Under a control that could land in air every model we ran looked region-sensitive; Proposition 1 shows how such a control can produce that, although beyond one marginal MedGemma-4B CIPE contrast our runs do not identify how much of it did. Matched on tissue, the answer is graded: sensitivity to the annotated union is lower in general models, where detection mostly depends on tie scoring, and the MedGemma lead survives image clustering but not an exposure audit. For the general models the control's effect remains unidentified, as do question-target alignment and specialisation; CIPE and the non-blind follow-up carry the narrower evidence stated above. An occlusion audit should report the tissue coverage of both occluded regions, match the control within a stated tolerance, drop and count items with no admissible placement, report W^* beside W with image-resampled intervals, and state whether the boxes annotate the image or the question target and whether the items sit on images the model's developers disclose training on.

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.

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. doi: 10.1109/CVPR.2018.00522.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

Alexander Amini, Anna Banaszak, Harold Benoit, et al. LFM2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 129696–129742, 2024.

Mohammad Asadi et al. MIRAGE: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. Abusing images and sounds for indirect instruction injection in multi-modal LLMs. *arXiv preprint arXiv:2307.10490*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Stella Biderman et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Léo Butsanets, Charles Corbière, Julien Khlaut, Pierre Manceron, and Corentin Dancette. RadImageNet-VQA: A large-scale CT and MRI dataset for radiologic visual question answering, 2025. URL `https://arxiv.org/abs/2512.17396`.

Mu Cai, Haotian Liu, Dennis Park, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, and Yong Jae Lee. ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0850.

Jan Clusmann, Dyke Ferber, Isabella C. Wiest, Carolin V. Schneider, Titus J. Brinker, Sebastian Foersch, Daniel Truhn, and Jakob Nikolas Kather. Prompt injection attacks on vision language models in oncology. *Nature Communications*, 16:1239, 2025. doi: 10.1038/s41467-024-55631-x.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. doi: 10.1038/s42256-021-00338-7.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024. doi: 10.1145/3664647.3685520.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics (TACL)*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410.

Federico Felizzi, Olivia Riccomi, Michele Ferramola, Francesco Andrea Causio, Manuel Del Medico, Vittorio De Vita, Lorenzo De Mori, Alessandra Piscitelli, Pietro Eric Risuleo, Bianca Destro Castaniti, Antonio Cristiano, Alessia Longo, Luigi De Angelis, Mariapia Vassalli, and Marcello Di Pumpo. Are large vision language models truly grounded in medical images? Evidence from Italian clinical visual question answering. In *MMRL4H Workshop at EurIPS*, 2025.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.371.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. doi: 10.18653/v1/2021.emnlp-main.775.

Jack Gallifant, Majid Afshar, Saleem Ameen, Yindalon Aphinyanaphongs, Shan Chen, Giovanni Cacciamani, Dina Demner-Fushman, Dmitriy Dligach, Roxana Daneshjou, Chrystinne Fernandes, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*, 31:60–69, 2025. doi: 10.1038/s41591-024-03425-5.

Itai Gat, Idan Schwartz, and Alexander Schwing. Perceptual score: What data modalities does your model perceive? In *Advances in Neural Information Processing Systems*, volume 34, pp. 21630–21643, 2021.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

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.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.670.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.62.

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.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMed-VQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. doi: 10.18653/v1/2020.acl-main.386.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. RadGraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Hanqi Jiang, Junhao Chen, Mingyu Kang, Hyeokjae Kwon, Yi Pan, Lifeng Chen, Weihang You, Haozhen Gong, Ruiyu Yan, Jinglei Lv, Lin Zhao, Hui Ren, Quanzheng Li, Tianming Liu, and Xiang Li. MedVIGIL: Evaluating trustworthy medical VLMs under broken visual evidence, 2026. URL https://arxiv.org/abs/2605.07919.

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

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: Insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024a.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, 2024b.

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. In *Transactions on Machine Learning Research (TMLR)*, 2024a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023a.

Dongxu Li, Yudong Liu, Haoning Wu, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023b.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in LLMs, 2025. URL https://arxiv.org/abs/2505.11423.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023c. doi: 10.18653/v1/2023.emnlp-main.20.

Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. In *Transactions on Machine Learning Research (TMLR)*, 2023.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024b. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`. Blog post.

Mahshad Lotfinia, Sebastian Ziegelmayer, Lisa Adams, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Vision-language models for chest radiography do not always need the image. *arXiv preprint arXiv:2606.17710*, 2026.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Andrés Marafioti, Orr Zohar, Miquel Farré, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 7670–7678, 2025. doi: 10.24963/ijcai.2025/853.

Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, et al. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9:429, 2022. doi: 10.1038/s41597-022-01498-w.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6, 2023. doi: 10.1038/s41746-023-00939-z.

OpenAI. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning into quantiles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015. doi: 10.1609/aaai.v29i1.9602.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Conference on Computational Natural Language Learning (CoNLL)*, 2023. doi: 10.18653/v1/2023.conll-1.21.

Letitia Parcalabescu and Anette Frank. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models and tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. doi: 10.18653/v1/2023.acl-long.223.

Jakub Pokrywka, Łukasz Grzybowski, Antoni Lasik, Marek Kubis, Jeremi Ignacy Kaczmarek, and Wojciech Kusa. Polish medical visual question answering: Vision-language models underutilize visual evidence. *arXiv preprint arXiv:2608.12928*, 2026.

Pranav Rajpurkar and Matthew P Lungren. The current and future state of AI interpretation of medical images. *New England Journal of Medicine*, 388(21):1981–1990, 2023. doi: 10.1056/NEJMra2301725.

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*, 2023. doi: 10.18653/v1/2023.findings-emnlp.206.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. doi: 10.18653/v1/D18-1437.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18770–18795, 2022.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Binesh Sadanandan and Vahid Behzadan. Attention without grounding: Causal evaluation of visual explanations in medical VLMs. *arXiv preprint arXiv:2607.18577*, 2026.

Andrew Sellergren et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. MediConfusion: Can you trust your AI radiologist? probing the reliability of multimodal medical foundation models. In *International Conference on Learning Representations (ICLR)*, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting GPT-3 to be reliable. In *International Conference on Learning Representations (ICLR)*, 2023.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1), 2020. doi: 10.23915/distill.00022.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. doi: 10.18653/v1/2023.emnlp-main.330.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical AI. *NEJM AI*, 1(3), 2024. doi: 10.1056/AIoa2300138.

Xingyu Wan, Qiaoying Teng, Jun Chen, Yonghan Lu, Deqi Yuan, and Zhe Liu. Eliminating language bias for medical visual question answering with counterfactual contrastive training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 194–204, 2025. doi: 10.1007/978-3-032-04978-0_19.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Lin, Wenbin Chang, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics (TACL)*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, et al. Chest ImaGenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 140334–140365, 2024.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. MedTrinity-25M: A large-scale multimodal dataset with multigranular annotations for medicine. In *International Conference on Learning Representations (ICLR)*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Bruce Changlong Xu, Lan Wu, and Alexander Ryu. A controlled audit of pretraining contamination in public medical vision-language benchmarks, 2026. URL https://arxiv.org/abs/2606.10066.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? An embarrassingly simple probing evaluation of large multimodal models in medical VQA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19188–19205, 2025. doi: 10.18653/v1/2025.findings-acl.981.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63565–63598, 2024. doi: 10.52202/079017-2030.

Anas Zafar, Leema Krishna Murali, Siddhant Bharadwaj, Ashish Vashist, and Jia Wu. Do medical vision language models actually see? a counterfactual grounding framework and hard-negative contrastive training for visually-reliant medical VLMs, 2026a. URL https://arxiv.org/abs/2607.03647.

Anas Zafar, Leema Krishna Murali, and Ashish Vashist. Beyond accuracy: Evaluating visual grounding in multimodal medical reasoning, 2026b. URL https://arxiv.org/abs/2603.03437.

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

Chenlu Zhan, Peng Peng, Hanrong Zhang, Haiyue Sun, Chunnan Shang, Tao Chen, Hongsen Wang, Gaoang Wang, and Hongwei Wang. Debiasing medical visual question answering via counterfactual training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 382–393, 2023. doi: 10.1007/978-3-031-43895-0_36.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. MultiTrust: A comprehensive benchmark towards trustworthy multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 49279–49383, 2024.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 15004–15019, 2024. doi: 10.18653/v1/2024.emnlp-main.836.

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

Yongshuo Zong, Yongxin Yang, and Timothy Hospedales. MEDFAIR: Benchmarking fairness for medical imaging. In *International Conference on Learning Representations (ICLR)*, 2023.

USE OF AI ASSISTANTS

We used generative AI tools (Anthropic Claude and OpenAI Codex agents) to implement experiment code and check it for correctness, to collect public datasets, to check references, and to polish the writing. The authors take full responsibility for all content.

A PROOF OF PROPOSITION 1

Fix one item and condition on the annotated region A and one control placement C; write $\phi_A = \phi(A)$, $\phi_C = \phi(C)$ and $\pi = \Pr(\Delta_A = \Delta_C)$. All statements about W and W^* follow from the per-comparison statements by linearity of expectation over items and draws; no independence across items or draws is needed.

(a) Exchangeability means $(\Delta_A, \Delta_C) \overset{d}{=} (\Delta_C, \Delta_A)$, so $\Pr(\Delta_A > \Delta_C) = \Pr(\Delta_C > \Delta_A)$. The events $\{\Delta_A > \Delta_C\}$, $\{\Delta_A < \Delta_C\}$ and $\{\Delta_A = \Delta_C\}$ partition the sample space, hence $2\Pr(\Delta_A > \Delta_C) + \pi = 1$. Averaging over the nK comparisons in (2), $\mathbb{E}W = \frac{1}{nK}\sum_{i,k}\frac{1}{2}(1-\pi_{ik}) \le \frac{1}{2}$ and $\mathbb{E}W^* = \frac{1}{nK}\sum_{i,k}\left[\frac{1}{2}(1-\pi_{ik}) + \frac{1}{2}\pi_{ik}\right] = \frac{1}{2}$. This part uses no model for Δ; it is why the grey tick in Figure 2(a) sits at $(1 - \text{tie share})/2$.

(b) Under (4), $\Delta_A > \Delta_C$ if and only if $\xi = \varepsilon_C - \varepsilon_A < \tau + h(\phi_A) - h(\phi_C) =: \tau + b$. Because ξ is independent of the placement and F is continuous, $\Pr(\Delta_A > \Delta_C \mid C) = \Pr(\xi < \tau + b) = F(\tau + b)$. Exchangeability of $(\varepsilon_A, \varepsilon_C)$ gives $\xi \overset{d}{=} -\xi$, so $F(0) = \frac{1}{2}$ by continuity. If $\phi_C \le \phi_A$, monotonicity of h gives $b \ge 0$ and monotonicity of F gives $F(\tau + b) \ge F(\tau)$, with strict inequality when $b > 0$ and F is strictly increasing on $[\tau, \tau + b]$. With $\tau = 0$ this is $F(b) > F(0) = \frac{1}{2}$: the annotated region wins more than half the time although it has no specific effect. Averaging over comparisons (with $\tau = 0$ and F strictly increasing), $\mathbb{E}W > \frac{1}{2}$ whenever every comparison has $b \ge 0$ and at least one has $b > 0$.

(c) If F has a density bounded by $\bar{f}$, then $|F(x) - F(y)| \le \bar{f}|x - y|$, so $|F(\tau + b) - F(\tau)| \le \bar{f}|b| \le \bar{f}L|\phi_A - \phi_C| \le \bar{f}L\delta$ when h is L-Lipschitz. For $\delta = 0$ the win probability equals $F(\tau)$; since $F(0) = \frac{1}{2}$ and F is strictly increasing, $F(\tau) > \frac{1}{2}$ if and only if $\tau > 0$. □

Remarks on scope. (i) W and W^* depend on the shifts only through the order of Δ_A and Δ_C, so the additive form (4) can be read on any strictly increasing transform of the total-variation scale; nonnegativity of TV is not in conflict with it. Because F is continuous the model produces no exact ties; the observed ties (4–9% of comparisons, Fig. 2(c)) lie outside it and are covered by part (a) alone. (ii) The implementation matches coverage per box within 0.15, so for a union of disjoint boxes the union's coverage gap is at most 0.15 as well; overlapping annotated boxes can loosen this. 58 of the 229 admitted items have overlapping annotated boxes, and there the mutually disjoint control boxes also occlude more pixels than the annotated union. (iii) Nothing in the argument

identifies τ with reliance on the question target: τ collects every difference between A and C other than foreground fraction, including position and anatomical coherence. (iv) The proposition says why a background-favouring comparator is invalid; it does not estimate how much of the difference between the historical legacy-control and foreground-matched runs is due to the control, because those runs are different QA samples (Appendix S4).

B Additional related-work context

B.1 Supporting detail

Shortcut exploitation, blind evaluation, and faithfulness metrics. That medical VQA accuracy can coexist with weak image use is supported by *counterfactual splits* that re-partition answer priors (VQA-CP (Agrawal et al., 2018), SLAKE-CP (Zhan et al., 2023), VQA-RAD-CP (Wan et al., 2025)) and by *blind evaluation*, where removing the image costs three of four frontier models only 2–9 points on 60 Italian items (GPT-4o loses 28) (Felizzi et al., 2025) or retains on average 70–80% of image-enabled accuracy (Asadi et al., 2026); on chest radiography specifically, occlusion probes find the image often unnecessary (Lotfinia et al., 2026). MedTrust-IN instead *curates* panel-defined text-only-hard items via a text-only screen. On the metric side the field has moved to intervention-based faithfulness (FPVG (Reich et al., 2023); HEAL-MedVQA (Nguyen et al., 2025)); CIPE inherits the balanced-pairs intuition of Goyal et al. (2017) and the modality-reliance line (Hessel & Lee, 2020; Gat et al., 2021; Parcalabescu & Frank, 2023) but is a per-sample counterfactual-KL score, with "faithfulness" following Jacovi & Goldberg (2020). Masking tests are known to be affected by out-of-distribution occlusion (Fong & Vedaldi, 2017; Hooker et al., 2019). Our legacy control raised a different issue, a comparator that covered less tissue (Appendix S4.2); Proposition 1(b) shows how that can raise W, and we do not measure how much it did (§4.1). Our audit documents an *answer-extraction* artefact in which a model appears *perfectly* image-invariant to same-readout controls but not to a raw-distribution check; evaluation-infrastructure studies motivate checking answer extraction and harness sensitivity (Duan et al., 2024; Biderman et al., 2024). Extended discussion and the full citation set are in Appendix S22.

S1 Training-data exposure of the specialist models

The corrected region readout (§S3) leaves both MedGemma checkpoints ahead of every general model in point estimate. This section reports why we no longer read that ranking as evidence about medical specialisation.

What the specialist family trained on, in its own words. MedGemma's technical report lists SLAKE (450 examples) among its vision-encoder, pretraining and RL training data (its Table 1) and, for its SLAKE evaluation, states verbatim that "[f]or SLAKE, we used the default train/test split" (Sellergren et al., 2025). The same report notes that VQA-RAD's original test set "includes some duplicated images in the train and test sets (with different questions)" and that its authors therefore built their own splits to avoid that contamination — for VQA-RAD. Not for SLAKE.

What our items rest on. We reproduced the region arm's item selection exactly (it is deterministic: closed English yes/no items carrying a `detection.json`, shuffled under a fixed seed) and checked each item's image against the official splits. **213 of 229 items (93.0%) sit on images that appear in the SLAKE train split; only four items are test-only.** Further, SLAKE's splits are *not* image-disjoint: $586 + 174 + 180 = 940$ image-name slots cover only 642 distinct images, and 52 of our items sit on images appearing in both train and test. Receipt: `outputs/region_exposure_audit.json`.

Why no subset rescues it. The obvious repair — restrict to items the specialist did not train on — **cannot be built from this dataset**. Four test-only items are not a cohort, and because images recur across splits, a test-split restriction would not remove image exposure even if it were large enough. The specialist-versus-general contrast on SLAKE is therefore confounded with training exposure *by construction*: the specialist family's own technical report places this split in its training data. We make no complementary claim about the four general models. Our receipt audits which SLAKE

split each image falls in — it does not audit any model's training corpus, and we did not audit theirs. What is documented is an asymmetry in *disclosed* exposure: one family states it trained on this split, the other four make no statement we can check either way. That is enough to make the contrast uninterpretable, and not enough to assert that the general models never saw these images.

What we claim instead. The measurement stands and the numbers are unchanged; what changes is what they are called. On a dataset a model family has trained on, occluding the annotated organ moves that family's answer distribution more than a tissue-matched control does. That is a statement about differential answer-distribution sensitivity under exposure. It is not evidence about where a model looks, not evidence of anatomical understanding, and not evidence of clinical reliance. Occlusion also remains a compound intervention: our control matches size, tissue coverage and non-overlap, but not anatomical coherence.

A disclosure-clean follow-up, run as exploratory. A region control on a corpus absent from the specialist's disclosed training data, with the same foreground-matched design and the corrected readout. We first named this experiment rather than running it; it has since been run under a registration committed before any file was downloaded (`PREREG_exposure_clean_region.md`), and it is reported in full in §S2. We did not search over subsets, readouts or models after the SLAKE measurement, and the consequence above was registered before the paper was edited (`PREREG_region_exposure_consequence.md`).

S2 Exploratory disclosure-clean, target-aligned follow-up

Registration, dataset, and what "clean" may mean. The registration (`PREREG_exposure_clean_region.md`) fixed the dataset, models, item construction, frozen instrument and branch consequences before any file was downloaded. Primary candidate was VinDr-CXR (Nguyen et al., 2022) — absent from every MedGemma disclosure surface we checked (technical report v4 and three model cards), an original hospital collection with no derivation chain into anything MedGemma discloses — with ChestX-Det as the registered fallback if credentialed PhysioNet access could not be arranged in time. The fallback ran. Its registered caveat applies verbatim: ChestX-Det's (Lian et al., 2021) images are a subset of NIH ChestX-ray14 (Wang et al., 2017), which has been openly mirrored since 2017, so a web-crawl exposure risk that VinDr-CXR would not carry cannot be ruled out. In either case the claim is "absent from all *disclosed* training data," never "never seen": MedGemma's card hedges with "including," its Table 1 lists two sources only as "Internal", and the Gemma 3 base crawl is undisclosed. The specialist is MedGemma-4B *only*: the 27B multimodal variant trained on Chest ImaGenome (Wu et al., 2021) "to enable anatomy localization" (report Appendix F) and is excluded from a clean-occlusion comparison by construction.

Items and instrument. 200 yes $+$ 200 no items over ChestX-Det's 13 finding classes (15–16 per class, seed 42), in the question core of MedGemma's own published prompt template ("Is there <finding> in this image?"); the template's chain-of-thought suffix is dropped because the frozen instrument reads yes/no logprobs, not generations (disclosed deviation). Yes-items occlude the queried finding's boxes. No-items ask about a finding the image's reader consensus lacks and occlude the *other* findings' boxes — and, as built, *all* of them: the registration said "a box from a different finding," singular and geometry-preserving, but the committed loader supplies every other finding's boxes, so no-items occlude 2.73 boxes on average against the yes arm's 1.43 (five no-items carry duplicate boxes). The no arm therefore does *not* preserve the yes arm's geometry; it measures sensitivity to occluding *other* findings' regions — cross-finding context — not localization of the queried finding. Only the yes arm's construction matches the registered estimand, and Amendment 2 (non-blind, below) registers arm-stratified reporting with the yes arm primary. The foreground-matched control and corrected-TV readout are byte-identical to the SLAKE arm; the loader is the only code that changed. 315 of 400 items admitted a valid tissue-matched control (168 yes, 147 no); the committed reader refuses below 120.

The registered verdict was an instrument fail, and the amendment is non-blind. The registered branch E3 fired: the black-vs-white positive control — built to catch a harness bug in which the image is never injected — came back below its committed bar on the specialist (0.016 against

IMAGE_SENSITIVITY_THRESH $= 0.02$). Its premise, however, does not survive this question type: on "is there a finding" questions a model that correctly ingests a *blank* image answers *no* at either colour, so colour-insensitivity is correct behaviour, not the bug the control was built to catch. The receipts themselves settle the ambiguity: a text-only forward would give the same $p(\text{yes})$ for one question on every image, and MedGemma-4B's within-question spread across real images is 0.384 (per-item $p(\text{yes})$ spans 0.000–1.000). The two diagnostics disagree in *opposite* directions on two models: MedGemma-4B fails colour but is image-driven, while Qwen2-VL-7B passes colour (0.208) yet is essentially image-insensitive on real content (spread 0.019) — and its combined $W = 0.535 \approx 0.5$ is what a *low-but-noisy* image effect produces, not a mechanical necessity: the readout scores strict wins ($\text{TV}_{\text{organ}} > \text{TV}_{\text{control}}$), so a perfectly image-invariant model would give $W = 0$ (every comparison a tie), while occlusion effects near the noise floor approach the coin-flip 0.5; either way, below the bar W stops being evidence about localization. Amendment 1 therefore replaced the gate for this question type with an ingestion diagnostic (within-question spread ≥ 0.05), **after all five W values had been seen**. We report that non-blindness rather than pretend otherwise; every consequence was bound in the amendment before the amended reader ran, and the reader's original 0.5 bar is documented there as a transcription error against the instrument's committed 0.02.

A second non-blind amendment: arms, seed, and framing. An independent check found three further defects, fixed by Amendment 2 — also written with every value already seen, also binding its consequences before the amended pipeline ran. *(a) Arms.* The no arm's construction does not match the registered estimand (above), so reporting is arm-stratified: the yes arm is primary and the no arm is a labelled secondary measuring cross-finding context sensitivity. *(b) Seed.* The receipts pipeline had run with seed 0 where the registration promised the frozen analysis's seed 42; it now runs at 42. W is deterministic given the receipts and did not move; the bootstrap q shifted within Monte-Carlo error (MedGemma-4B combined $1.7 \times 10^{-4} \to 2.5 \times 10^{-4}$). The registration's phrase "permutation null" misnamed the frozen implementation, an *item bootstrap*; the term is corrected in the registration, the implementation untouched. *(c) Framing.* Amendment 1's spread bar (0.05) was chosen after the values were seen, so it is demoted from an exclusion gate to a diagnostic label: the ingestion SD is reported continuously for all five models (Table 2), Qwen2-VL-7B is described as image-insensitive *under this yes/no readout* with its W in the table rather than excluded, and no exclusion enters the ordering — MedGemma-4B exceeds it regardless. The bar still gates the specialist itself: an image-insensitive specialist has no interpretable W.

Exploratory result (E1″). Table 2 reports all three strata for all five models. Both amendments admitted data already seen, so no branch below is a confirmatory replication. On the primary yes arm — the estimand the registration fixed — MedGemma-4B's corrected-TV $W = 0.742$ ($q = 5 \times 10^{-4}$, $n = 168$) exceeds every general VLM: InternVL3-8B 0.626, Qwen2.5-VL-7B 0.607, Qwen2-VL-7B 0.570, Gemma3-4B 0.526. Combined over both arms, $W = 0.681$ ($q = 3 \times 10^{-4}$, $n = 315$); on the no arm — the labelled cross-finding secondary — MedGemma-4B is again first (0.612, $q = 5 \times 10^{-3}$, $n = 147$). Three scope sentences travel with these numbers wherever they appear: one checkpoint on one corpus; "clean" means absent from *disclosed* training data; and both admitting amendments were non-blind. The SLAKE withdrawal stands regardless — nothing here rehabilitates a measurement made on the specialist's own training images.

S3 Normalisation correction of the region readout

Our first region readout was not total variation. The helper that persists answer distributions returns max-shifted likelihood weights, $[\exp(\ell_{\text{yes}} - m), \exp(\ell_{\text{no}} - m)]$ with $m = \max \ell$, so the larger entry is exactly 1 and each vector sums to $(1, 2]$. Over the committed receipts, 2378 **of** 2748 **stored vectors do not sum to** 1 (range 1.000–2.000, mean 1.104). CIPE normalises internally and is unaffected; total variation and $|\Delta p(\text{yes})|$ read the raw weights, so every number they produced inherited the defect. We found this while answering a reviewer-style audit of our own pipeline, and we report it here rather than silently substituting corrected values.

The correction is receipt-only — no inference was re-run — and was pre-registered with its branches before it was applied (PREREG_region_tv_correction.md; that registration also records that it was *not* blind, since a preliminary read-only diagnostic had already indicated the likely outcome). Two estimands are now reported rather than one: the item-level difference

Table 2: Exploratory disclosure-clean follow-up, arm-stratified per Amendment 2 (non-blind, disclosed; not confirmatory). The yes arm is primary: it occludes the queried finding, matching the registered estimand. The no arm is a labelled secondary measuring cross-finding context sensitivity: it occludes *all* other findings' boxes, 2.73 per item against the yes arm's 1.43. Ingestion SD is the within-question spread of $p(\text{yes})$ across real images, reported for all five models; Qwen2-VL-7B (0.019) is image-insensitive under this yes/no readout, so its W is uninformative rather than evidence — reported, not excluded, and MedGemma-4B exceeds it regardless. q: BH-FDR across the five models within each stratum; item bootstrap, $B = 10{,}000$, seed 42; printed to one significant figure (the combined MedGemma-4B q is 2.5×10^{-4}).

Model	Ingestion SD	Yes arm (primary) W (q), n=168	No arm (secondary) W (q), n=147	Combined W (q), n=315
MedGemma-4B	0.384	**0.742** (5×10^{-4})	0.612 (5×10^{-3})	0.681 (3×10^{-4})
InternVL3-8B	0.112	0.626 (5×10^{-4})	0.578 (0.026)	0.604 (3×10^{-4})
Qwen2.5-VL-7B	0.135	0.607 (1×10^{-3})	0.581 (0.023)	0.595 (3×10^{-4})
Gemma3-4B	0.129	0.526 (0.358)	0.569 (0.052)	0.546 (0.038)
Qwen2-VL-7B	0.019	0.570 (0.016)	0.495 (0.879)	0.535 (0.087)

$D_i = \text{TV}_{\text{organ},i} - \overline{\text{TV}}_{\text{control},i}$, and the win rate W_i that averages the indicator *over* the five control draws instead of comparing against their mean. The quantity our first analysis reported, $\mathbb{1}[\text{TV}_{\text{organ}} > \overline{\text{TV}}_{\text{control}}]$, is deprecated: averaging five draws is legitimate variance reduction for D, but it does not carry the chance-0.5 interpretation we first attached to it. All five single-draw win rates are reported so that no draw is selected.

Historical registered reading under strict wins: R1. This is not a tie-robust classification: Appendix S5 finds that the general-model verdict changes under half-tie scoring. Under the original convention, both specialist checkpoints remain significant ($W = 0.639$ for MedGemma-27B and 0.612 for MedGemma-4B, $q = 3 \times 10^{-4}$, bootstrap CIs excluding 0.5; $D = +0.059$ and $+0.051$ with CIs excluding zero) and three of the four general models do not ($q = 0.099$–0.316). Two things did change and both are reported: the magnitudes moved (the deprecated quantity gave MedGemma-4B 0.646 and MedGemma-27B 0.594), and the *within-family ordering flipped* — MedGemma-27B is now the stronger checkpoint, where the defective readout had put MedGemma-4B first. Any statement in this paper about which specialist checkpoint is strongest is therefore a corrected statement, and the uncorrected ordering is superseded. Receipt: `outputs/region_tv_corrected.json`, emitted by `code/scripts/region_tv_corrected.py`.

S4 Historical sample identity and the shared subset

Four general-model minimum-overlap receipts disagree with the foreground candidate sequence on box counts at 310 of 400 identical `slake_i` labels. Replaying the original loader against the pinned official SLAKE `train.json` revision reproduces their box-count sequence 400/400; the new foreground sequence instead comes from the later per-image-QA fallback. These are different QA samples, not a controlled switch on 229 identical items. The old MedGemma-4B receipt uses the fallback sequence and its annotated-mask CIPE agrees with the new run on all 229 admitted QA, so it must not be treated as sharing the old general roster. Because it shares the matched run's sequence, we formed the paired legacy-versus-matched contrast for MedGemma-4B on those 229 QA post hoc (Appendix S4.1).

Matching unique (image path, question, answer) keys yields 62 shared candidate QA, with identical ordered annotation boxes, of which 34 were admitted to the foreground analysis. On those 34, each general model's old/new annotated-mask CIPE agrees exactly. This is strong conditional reconstruction evidence; historical receipts omitted QA text and pixel hashes, so it is not independent verification of every old pixel. One duplicated source key is excluded from one-to-one matching. The source revision, source file hash, mappings and execution checks are in the supplementary material under `outputs/legacy_mapping_20260923/`.

Paired description on the actual intersection. Before inspecting control outcomes we fixed $D_L = \mathrm{CIPE}_{\text{legacy control}} - \mathrm{CIPE}_{\text{annotation}}$ and $D_F = \overline{\mathrm{CIPE}}_{\text{foreground controls}} - \mathrm{CIPE}_{\text{annotation}}$. Positive $D_L - D_F$ means the old CIPE contrast is larger. Both are CIPE contrasts oriented control minus annotation, not the TV-based D of Eq. (3). We resample the 33 images with all QA rows (20,000 draws, seed 20260923), retaining QA weighting. Only Qwen2-VL's Bonferroni 98.75% interval over four models excludes zero (Table 3). Ordinary 95% intervals and all per-item contributions also ship; no model or QA was selected by the sign. This post-hoc 34-QA intersection, conditional on one stored legacy mask versus five matched masks per QA, cannot establish a general control effect over 229 items or under the TV readout.

Model	Paired $D_L - D_F$	Image-cluster 98.75% CI
Gemma3-4B	$+0.00457$	$[-0.03850, +0.04518]$
InternVL3-8B	-0.00161	$[-0.01647, +0.00632]$
Qwen2.5-VL-7B	$+0.02819$	$[-0.00379, +0.08316]$
Qwen2-VL-7B	$+0.00620$	$[+0.00196, +0.01265]$

Table 3: Same-QA, same-CIPE comparison on 34 recovered shared QA (33 images), with four-model-adjusted intervals. Source: `paired34_stored_mask_descriptive.json`; generator: `paired_legacy_foreground_34_20260923.py`.

S4.1 MedGemma-4B: legacy versus matched control on the same 229 QA

Why only this model, and why CIPE. MedGemma-4B is the one model whose legacy run shares the matched run's sample sequence. Before any control value was read, an identity gate aligned the two receipts by item ID (not list position) and found all 229 admitted IDs in the legacy receipt with equal box counts, ordered boxes equal to the source mapping, and float-identical annotated-mask CIPE (49 of these sit at the clip boundary and carry no fingerprint). The legacy receipt stores only per-item CIPE, not answer distributions or control boxes, so total variation cannot be computed for it; its control boxes were not stored, but they can be replayed from the seed on CPU (Appendix S4.2). The contrast is therefore a CIPE analogue of Eq. (2), not the TV-based W of Table 1: an annotation win is $\mathrm{CIPE}_A < \mathrm{CIPE}_C$, $w_i^L = \mathbf{1}[\mathrm{CIPE}_{A,i} < \mathrm{CIPE}_{L,i}]$ for the one legacy draw, $w_i^F = \frac{1}{5}\sum_k \mathbf{1}[\mathrm{CIPE}_{A,i} < \mathrm{CIPE}_{F,ik}]$ averaged over the five matched draws, ties scored zero, and the primary estimand is $\Delta W = \frac{1}{229}\sum_i (w_i^L - w_i^F)$. Estimand, image-cluster bootstrap (175 images, 20,000 draws, seed 20260924), sign-flip test, decision rule and the sentence to print for each outcome were committed before the analysis (`notes/PREREG_medgemma4b_legacy_vs_matched_20260924.md`); the registration records that the aggregate legacy and matched rates on other denominators were already known.

Result. The registered verdict is that the win rate is higher under the legacy draw (Table 4): $\Delta W = +0.054$, 95% CI $[+0.005, +0.105]$, sign-flip $p = 0.040$ at the registered seed 20260924 (the analysis script used 20260925 and gave 0.041; this deviation is logged beside the registration), and half-tie scoring agrees. It is borderline. The interval nearly reaches zero, four of the five single-matched-draw contrasts have intervals that include zero, and, in a post-hoc description that was not registered, 64 items on which all seven CIPE values agree to within 10^{-6} contribute $+0.007$ of the difference (142 items within 10^{-4} contribute $+0.021$). The comparison changes placement rule and coverage together (the replayed legacy boxes cover less tissue, Appendix S4.2), so it does not isolate foreground coverage as the cause (Proposition 1). It is one model, one legacy draw against five matched draws, on the CIPE readout; it says nothing about the four general models, whose legacy runs used different QA. W_F is not the 0.585 of Table 19, which compares against the mean control CIPE. Nor does it account for the historical number: on all 400 legacy items the same strict rate is 0.700 against 0.642 on these 229, so the item pool and the control move it by similar amounts (-0.058 and -0.054 on CIPE), and the historical difference is not attributed to the control.

S4.2 Coverage replay of the legacy placement

Why a replay is possible, and how it was checked. The legacy receipt stores no control boxes, but the placement is deterministic. The legacy script creates `random.Random(42)` after the positive

Quantity (MedGemma-4B, n=229 QA, CIPE)	Estimate	Image-cluster 95% CI
W_L, one legacy draw (strict)	0.642	$[0.581, 0.703]$
W_F, five matched draws (strict)	0.588	$[0.536, 0.640]$
$\Delta W = W_L - W_F$ (primary; sign-flip $p = 0.040$)	$+0.054$	$[+0.005, +0.105]$
W_L^* / W_F^* (half-tie)	0.710 / 0.647	—
ΔW^* (half-tie)	$+0.062$	$[+0.019, +0.108]$
$D_L^{\mathrm{CIPE}} - D_F^{\mathrm{CIPE}}$ (as Table 3)	$+0.023$	$[+0.002, +0.047]$
ΔW with matched draw 1 only	$+0.039$	$[-0.018, +0.102]$
ΔW with matched draw 2 only	$+0.035$	$[-0.022, +0.092]$
ΔW with matched draw 3 only	$+0.066$	$[-0.005, +0.139]$
ΔW with matched draw 4 only	$+0.074$	$[+0.009, +0.142]$
ΔW with matched draw 5 only	$+0.057$	$[-0.004, +0.118]$

Table 4: Paired legacy-versus-matched CIPE win rates for MedGemma-4B on the same 229 QA (175 images). Ties: 13.5% of legacy and 11.9% of matched comparisons. Receipts: `outputs/medgemma4b_paired_20260924/` (`identity_gate.json`, `paired_contrast.json`, `saturation_descriptive.json`, `signflip_registered_seed.json`); generators in `code/scripts/`: `medgemma4b_legacy_matched_identity_20260924.py`, `medgemma4b_legacy_vs_matched_20260924.py`, `medgemma4b_paired_saturation_descriptive_20260924.py`, `medgemma4b_paired_registered_seed_20260924.py`.

control, and its only consumer is the placement routine, which draws up to 30 positions per annotated box and keeps the first with zero overlap (else the least); how much of the stream an item consumes depends only on image size and box geometry. A row is written only on success and all 400 rows are present, so no item skipped the stream, and the placement routine is byte-identical between the run commit and the replay. Before any legacy box or coverage was computed we registered the identity checks, estimand, decision rule and the sentence to print for each outcome (`notes/PREREG_legacy_coverage_replay_20260925.md`). No model is loaded. The replay uses the seed-42 item order, sizes and boxes of the image mapping (Appendix S5) and the source images, each checked against its stored SHA-256. Three checks pass. (I1) Recomputed annotated coverage for all 400 items and matched coverage for the 229 admitted items equal the stored values exactly. (I2) Replaying the matched placement from the same seed reproduces all 229×5 stored matched box sets and the $229/171$ partition. (I3) On 21 items the replayed legacy boxes fall on pixels that are already black, so the masked image equals the original and the stored legacy CIPE must be exactly 1; all 21 are. The replayed legacy boxes themselves cannot be compared with stored ones, since none were stored, and the model-forward check (rerunning MedGemma-4B on them) was not run. Post hoc, we measured what I3 can detect: none of 200 wrong seeds passes it, but 13 of 20 one-item offsets of the stream (one at every 20th item) do; such an offset needs an item that failed before placement, and none did.

Items	annotated	matched (5 draws)	legacy (replayed)
229 admitted, all	0.961	0.943	0.659
CT (82)	—	0.873	0.460
MRI (47)	—	0.959	0.343
X-ray (100)	—	0.993	0.971
boxes with $\phi < 0.5$	2.9%	3.1%	48.4%
400 legacy items	0.957	—	0.517

Table 5: Mean foreground coverage ϕ (means of per-item means) of the annotated boxes, the foreground-matched draws and the replayed legacy boxes of the MedGemma-4B runs. On all 400 legacy items the annotated-minus-legacy gap is 0.440 $[0.399, 0.481]$ (296 images; $[0.399, 0.480]$ without the one item whose image identity is ambiguous). Receipts: `outputs/legacy_coverage_replay_20260925/` (`identity.json`, `coverage.json`, `gate_power_posthoc.json`); scripts `legacy_coverage_replay_20260925.py` and `legacy_replay_gate_power_20260925.py` (in the review supplement). Modality rows are descriptive and were not registered.

Result. The registered verdict is that the legacy placement covered much less tissue (Table 5): on the 229 items, mean coverage 0.659 against 0.943 for the matched draws, a gap of 0.284 $[0.232, 0.338]$ (image-cluster 95% interval, 175 images, 20,000 draws, seed 20260925). The gap sits in CT and MRI; on X-ray the two are close (a descriptive split, not registered). Across 200 other seeds the 229-item legacy mean ranges from 0.612 to 0.683 (post hoc), so the conclusion does not rest on the exact draws. This measures coverage, not its effect. It does not say how much of MedGemma-4B's paired $\Delta W = +0.054$ came from coverage, since the placement rule changed too, and it says nothing about the general-model legacy runs, which used different QA and are not replayed. Coverage is ϕ of §3 with the matched script's box slicing; the legacy fill also blackens each box's closing row and column, which ϕ does not count. Three presentation choices were made after the result and are not registered: the abstract quotes the 229-item means and the introduction summarises them in words; the main-text sentence sits after Proposition 1(b), with its wording adapted to that position, instead of in the "Two controls" paragraph of §3 that the registration named; and the per-modality rows of Table 5 were added and are cited in the main text.

S5 Post-hoc image-cluster and exact-tie sensitivity

Recovering the replication unit. Historical `slake_i` receipts omit source image names. Replaying the original seed-42 loader on source per-image QA metadata matches all 400 ordered annotation-box lists and image dimensions in each of six receipts ($2{,}400/2{,}400$). Every included item's box/dimension fingerprint identifies one QA-referenced source image; one excluded item has an unresolved collision and enters no estimate. The 229 retained QA rows span 175 distinct source-image byte hashes: 129 images contribute one QA, 39 two, 6 three and one four. This validates current image grouping conditionally on source metadata; absent historical QA text/pixel hashes cannot be independently reconstructed from the model receipts.

Plan and estimands. Before computing outcomes we fixed $20{,}000$ image-cluster bootstrap draws (seed 20260922), preserving all questions and five control draws within each sampled image. The statistic is the QA-weighted ratio of summed scores to summed QA counts; image-equal weighting is reported separately in the receipt. The normalised-TV strict-win point estimates reproduce the original six W values exactly. Their image-clustered 95% intervals are $[0.589, 0.688]$ and $[0.560, 0.664]$ for MedGemma-27B and -4B. Both also retain positive mean TV differences: 0.059 $[0.031, 0.089]$ and 0.051 $[0.021, 0.085]$.

Exact ties change the general-model reading. The original W scores strict wins as one and ties as zero. Exact ties occur in 4.3–9.0% of the stored comparisons. Under exchangeability of the two occlusions, its expectation is $(1 - P(\text{tie}))/2$, not 0.5. We therefore also report $W^* = P(\text{win}) + \frac{1}{2}P(\text{tie})$, without modifying the original estimator or receipts. All four general models show weaker positive W^* effects (Table 6); all six conditional image-cluster sign-flip tests on W^* survive Holm adjustment. Sign flips assume independent image clusters and symmetric cluster score sums under the 0.5 null; these are post-hoc sensitivity tests, not a new confirmatory registration. With strict wins, both MedGemma effects survive Holm but InternVL3's adjusted $p = 0.0528$ does not; BH retains it ($q = 0.0264$). Categorical specialist/general separation is therefore not robust.

Model	Strict W	Exact ties	Half-tie W^* [cluster CI]	Holm p^*
Gemma3-4B	0.545	4.5%	0.567 [0.520, 0.615]	0.0207
InternVL3-8B	0.563	6.4%	0.595 [0.547, 0.642]	0.0010
MedGemma-27B	0.639	4.3%	0.661 [0.611, 0.708]	0.0003
MedGemma-4B	0.612	9.0%	0.657 [0.610, 0.704]	0.0003
Qwen2.5-VL-7B	0.536	4.5%	0.559 [0.509, 0.610]	0.0373
Qwen2-VL-7B	0.525	7.1%	0.560 [0.511, 0.609]	0.0373

Table 6: Post-hoc image-cluster sensitivity, 229 QA items on 175 images. CI resamples images; Holm-adjusted sign-flip tests are conditional on the assumptions stated above. The source is `outputs/image_cluster_20260922/image_cluster_sensitivity.json`, generated by `code/scripts/image_cluster_audit_20260922.py`.

S5.1 RECEIPT-ONLY INSTRUMENT CHECKS

Registration. `notes/PREREG_region_instrument_20260926.md` was committed before these receipts were computed. It lists two receipt-only additions as descriptive, with no decision attached: a placebo and a tolerance tightening. Both read the six stored foreground-matched receipts; no model is run. The same registration fixes a marker positive control, a fill, tolerance and threshold grid, and a coverage-isolation arm that need new inference; none of those has run, and nothing in this paper reports them.

Negative control (placebo). For each model and each $k \in \{1, \ldots, 5\}$, control draw k plays the annotated region and is scored against the other four draws with W^*. There is no annotation effect by construction, so W^* should sit at one half. The registration flags the instrument if at least 6 of the 30 image-cluster intervals exclude 0.5; none does (W^* between 0.465 and 0.545; Table 7). This checks the null calibration of W^* on matched draws; it is not a positive control, and the instrument's power for a small annotation effect remains unmeasured (§5).

Tolerance tightening. Keeping only the stored draws whose largest per-box coverage gap is at most 0.05 (or 0.10), and the items with at least one such draw, gives the W^* in Table 7. The stored draws allow only tightening; looser tolerances, other foreground thresholds and other fills need new placements and inference and are untested.

Model	Placebo W^*, 5 draws (CIs $\not\ni 0.5$)	W^*, gap ≤ 0.05 (194 items, 149 images)	W^*, gap ≤ 0.10 (221 items, 171 images)
MedGemma-27B	0.472–0.538 (0/5)	0.660 [0.602, 0.715]	0.654 [0.604, 0.704]
MedGemma-4B	0.479–0.517 (0/5)	0.673 [0.617, 0.727]	0.659 [0.609, 0.711]
InternVL3-8B	0.469–0.545 (0/5)	0.583 [0.528, 0.637]	0.603 [0.553, 0.653]
Gemma3-4B	0.465–0.519 (0/5)	0.546 [0.490, 0.602]	0.576 [0.525, 0.626]
Qwen2.5-VL-7B	0.481–0.537 (0/5)	0.558 [0.498, 0.618]	0.559 [0.505, 0.612]
Qwen2-VL-7B	0.477–0.529 (0/5)	0.570 [0.514, 0.626]	0.560 [0.508, 0.611]

Table 7: Receipt-only instrument checks (descriptive, registered before computation). Brackets are image-cluster 95% intervals (20,000 draws, seed 20260926); full-set values are in Table 6. Receipt: `results/region_instrument_20260926/receipt_only.json`; script `region_instrument_analysis_20260926.py`.

S6 ARTEFACT TAXONOMY AND PRESCRIBED CONTROLS

What the audit teaches. A counterfactual faithfulness metric is only as trustworthy as the decoding that reads the answer *and* the composition of the control it is compared against; neither may be shared with the thing being validated. Three clean-looking results failed through a shared readout, a shared counterfactual term and pseudo-replication. The region comparison additionally exposed unstable sample identity, preventing the original causal attribution to control placement. These failures belong to the pipeline around the metric; CIPE's own head-to-head loss is reported separately (§4.3).

An artefact taxonomy and prescription. We report all reversals in full —including our own first misdiagnosis—because the failure modes we hit plausibly recur in any counterfactual VLM-faithfulness metric (a prescription grounded in one end-to-end case, not a proof across metrics). *(1) Answer-extraction:* a fixed decoding can saturate the metric to a constant while the image is fully causal, and controls that share the readout (raw-KL saturation, black-vs-white sensitivity) are blind to it—only a raw-distribution control catches it. *(2) Image-injection:* a harness can genuinely drop the image for some architectures; the positive control must be readout-agnostic or it inherits (1). *(3) Metric–stratifier circularity:* a stratified validation sharing a term with the metric manufactures separation with no faithfulness content. *(4) Control-region composition:* a control region drawn without matching the content of the intervened region measures where in the frame an occlusion falls rather than what it covers—this is the one our own prescribed controls did *not* catch, because a positive control certifies that the image is read, not that the right region is (§4.1).

We propose a *raw-distribution positive control*, *per-model answer-extraction validation*, a *circularity stress-test*, and a *content-matched control region* as required reporting, and implement all four as reusable, self-checking components (to be released upon publication). Two secondary observations: the panel scale trend is directional but specification-dependent on the 21 independently certified models (mixed-model coefficient -0.079, -0.052 over all 40; a family random intercept gives p=0.033, OLS with family fixed effects gives p=0.237; not claimed; all-40 p=0.064 for comparison only, §4.3); and `text_only`'s reactivity suggests an *image-presence cue* (the model treats "an image was provided" as a confidence signal independent of content)—a hypothesis confounded by the `text_only` "no image provided" sentinel (§3; Appendix S7), and a shortcut accuracy-gap benchmarks miss by holding image presence constant.

S7 Extended limitations

Partially-recovered cluster: of the five readout-artefact models, three (Idefics2-8B, LLaVA-1.5-7B, Fuyu-8B) recover and are re-scored, but two (Aria-MoE-25B-3A, VipLLaVA-7B) remain unresolved (Aria was re-audited but shows near-zero yes/no mass and no black-vs-white shift; VipLLaVA hits a state-dict incompatibility; §4.2). *Weak / dataset-dependent de-inversion*: after decircularization it survives only on PathVQA (H1 0.431, 95% CI $[0.413, 0.446]$) and not on VQA-RAD or SLAKE (§4.4); pooled item-level construct validity is modest (AUROC $\approx$0.53). *Region estimand and selection*: the SLAKE arm masks the union of all image-level boxes, not a question-aligned target. Only $229/400$ items admit a foreground-matched placement; accepted items carry fewer boxes than exclusions (1.66 vs. 4.46 on average). The historical receipts omit image names. A post-hoc source-metadata reconstruction identifies all included image groups and supports image-clustered sensitivity (Appendix S5); it does not verify the historical QA text or pixel bytes. Original QA-item intervals remain in Appendix S39. *Metric design*: the `text_only` condition inserts a "no image provided" sentinel, so its KL conflates image removal with a template change; in panel runs `anatomy_mask` is a centred-square control. CIPE loses its head-to-head against simpler readouts, and total variation is retained (Appendix S10). *Provenance and contamination*: v0.3 is rebuilt on source *test* splits after two marginal PathVQA screener gaps fired the registered rule (§4.4); test membership removes the original split confound but does not prove absence from model training. The CIPE ranking is stable but not invariant across split. *Corpus semantics*: MedTrust-IN is panel-defined text-only-hard, not a human-verified necessity label; $59/1{,}122$ items are unsolved by every roster model even with the image, while $871/1{,}122$ are solved by a roster majority. *Coverage and equity*: we evaluate English-language 2D imaging; the panel is mostly general VLMs plus two exposed MedGemma checkpoints; source datasets inherit demographic and prevalence biases (Zong et al., 2023), which models can propagate (Omiye et al., 2023).

S8 Broader impact, ethics and reproducibility statements

Broader Impact

MedTrust-IN and the foreground-matched protocol support routine auditing of whether medical VLM answers respond to image content. They offer no clinical-workflow, patient-outcome or site-shift evidence and are not a substitute for prospective clinical evaluation, regulatory review or site-specific validation (Rajpurkar & Lungren, 2023). A panel-defined text-only-hard corpus inherits its panel's answer prior (85.6% yes in v0.3), so accuracy comparisons should use the balanced subset, and a favourable region audit is a necessary, not sufficient, screen.

Ethics Statement

All data are publicly available, de-identified medical imaging datasets (SLAKE, VQA-RAD, PathVQA, and ChestX-Det (Lian et al., 2021) for the disclosure-clean region arm — a re-annotated subset of NIH ChestX-ray14 (Wang et al., 2017), used under its public release with the expert box annotations taken as shipped); no new patient data were collected and no human subjects were involved. The v0 results rely solely on programmatic counterfactuals; we recruited no new human annotators (the annotator study is deferred v1 scope, Appendix S36), though the external region val-

idation (§4.1) uses SLAKE's existing human region annotations. We follow the datasheet protocol of Gebru et al. (2021) (Appendix S32). Generative-AI use is disclosed in the Use of AI Assistants statement; all reported numbers derive from the per-record dumps via the committed readers.

REPRODUCIBILITY STATEMENT

Code, data-processing scripts, model configurations, prompt templates, MedTrust-IN v0.3 with its builder and datasheet, and the per-record result dumps will be released upon publication. An anonymized subset for review is provided in the supplementary material: the per-item receipts of the six foreground-matched, four legacy general-model and legacy MedGemma-4B SLAKE region runs, the recovered image mapping and legacy-identity mapping, the corrected-TV receipt, and the analysis scripts with an offline verifier (Python and NumPy only) that recomputes the image-cluster, legacy-identity, paired-subset and MedGemma-4B paired-contrast analyses from those receipts. The supplement also holds, hash-checked but not re-run, every registration file the paper cites by name, with the first and last commit of each in our project repository (not an external registry), the legacy-coverage replay, ChestX-Det and receipt-only instrument-check receipts with their scripts, and the MedTrust-IN v0.3 items (question text for SLAKE and VQA-RAD, PathVQA items by reference with a rehydration script, no pixels) with their statistics and datasheet; the supplement README states each source's license and attribution. Model inference cannot be re-run from the supplement: source images are referenced rather than redistributed. Open-source models are identified by their HuggingFace `model_path` (exact commit revisions were not pinned in v0). Two points correct our own first analysis, both verified against the repository. *(i) Nothing is cached.* `code/medtrust/models/base.py` computes a SHA-256 request key per output but implements no cache around it; that module is part of the deferred v1 hosted-API path and is never imported by the pipeline that produced any reported number (`cipe_v0_pilot.py` performs a single linear pass and one terminal write, with no cache read, resume, or incremental output). Reruns are full recomputation. *(ii) There is no held-out split,* and none is offered for review. All items in both v0 and v0.2 are drawn from source *train* splits. Reproducibility of the corpus is scoped as follows: the recommended artefact is **MedTrust-IN v0.3**, built on source *test* splits (v0.2 and v0 retained as superseded); every item is deterministically reconstructable from its source dataset at seed 42, and its `image_ref` names the same split as its `source_split`, gated by the builder, but re-deriving the *screen* that selected it requires the panel's per-item dumps at the full source-pool sizes; 516 of the $1{,}197$ v0 items (43%) were screened at pool sizes (SLAKE n=$1{,}681$, PathVQA n=$2{,}400$) whose dumps exceeded the committed $n \leq 1{,}200$ release. Those dumps will be released with the corpus, which makes the v0.2 screen re-derivable end to end (§4.4).

S9 SUMMARY OF POSITIVE FINDINGS

What survives. (1) Under the foreground-matched control and the corrected TV readout, both MedGemma checkpoints show image-level annotated-region sensitivity that survives image clustering, and general models show weaker effects whose detection depends on tie scoring (Table 1); the family contrast is exposure-confounded (§S1). (2) The previously "saturated" cluster is a readout artefact in the three families that recover under a robust readout with CIPE 0.67/0.92/0.87; Aria and VipLLaVA remain unresolved (Appendix S29). (3) Medical-specialist image benefit appears at two scales on VQA-RAD (MedGemma-4B/27B: with-image accuracy 0.74/0.755 vs. 0.55/0.595 text-only); VQA-RAD is also in MedGemma's disclosed training data, so this benefit is exposure-confounded in the same way. (4) MedTrust-IN v0.3 is a built, datasheeted, positive-controlled, panel-defined text-only-hard corpus, re-issued after its own screen failed an audit. CIPE is reported as a negative and is not recommended for ranking models.

S10 CIPE AGAINST ALTERNATIVE READOUTS

The same runs under CIPE and under corrected TV. Under the paired within-item test—the statistic the original MedGemma result used—the annotated-region effect appears in both specialist checkpoints. Read *under CIPE* — the readout this paper does *not* recommend — the pair is MedGemma-4B 0.592 (p=6.9×10^{-5}; ties count half here, whereas the 0.585 of Table 19 counts strict wins on the same Wilcoxon test) and MedGemma-27B 0.587 (p=5.2×10^{-4}), with the four

general VLMs at 0.445–0.515 (no $p < 0.05$). Read under the corrected total variation (§S3) the pair is 0.639 (MedGemma-27B) and 0.612 (MedGemma-4B), the within-family ordering is the reverse of CIPE's, and InternVL3-8B also reaches significance (0.563, q=0.030), so the separation is graded rather than categorical. A tissue-matched region control therefore ranks the two specialist checkpoints above every general model on either readout, and does so in two checkpoints rather than one — a ranking §S1 then shows cannot be read as specialisation, because 93.0% of these items sit on images the specialist family trained on. On this secondary paired statistic, computed on the unnormalised vectors of Appendix S3, total variation is not below CIPE in either specialist; that does not show it is the better readout. **The matched-control ranking is exposure-confounded, and the metric is not validated.**

Addresses the standing objection that CIPE's ranking may track answer-distribution sharpness rather than visual reliance, and that the paper never compares it against the cheaper readouts a rate-based audit would use. Both probability vectors are in the committed records, so this needs no new inference (`code/scripts/cipe_metric_comparison.py`). Restricted to the 17 n=500 evaluations that pass the positive control. Every score is oriented so that higher = less image-reactive, matching CIPE's sign, then scored by AUROC against per-item correctness.

Table 8: **CIPE buys no item-level signal over a flip rate.** Macro-mean AUROC 0.529 for CIPE against 0.539 for the argmax flip rate; the paired difference favours the flip rate by 0.010, which wins 8 of 17 evaluations at Wilcoxon p=0.55—statistically indistinguishable, and the flip rate is far cheaper. CIPE also ranks items almost identically to total variation and to $|\Delta p(\text{yes})|$ ($\rho = 0.924$ for both), so the clipping and the KL asymmetry contribute little. All readouts sit near chance, which is consistent with the weak construct validity reported in §4.3.

evaluation	CIPE	raw KL	JSD	TV	$\lvert\Delta p\rvert$	flip
pathvqa_gemma3_4b	0.474	0.454	0.449	0.454	0.454	0.461
pathvqa_internvl3_8b	0.471	0.452	0.457	0.464	0.464	0.544
pathvqa_qwen25vl7b	0.461	0.464	0.459	0.469	0.469	0.472
pathvqa_qwen2vl7b	0.373	0.370	0.360	0.404	0.404	0.463
pathvqa_qwen3vl8b	0.574	0.559	0.537	0.553	0.553	0.530
slake_gemma3_27b	0.629	0.559	0.552	0.587	0.587	0.574
slake_gemma3_4b	0.544	0.497	0.520	0.533	0.533	0.534
slake_internvl3_8b	0.564	0.508	0.451	0.526	0.526	0.558
slake_qwen25vl7b	0.540	0.492	0.448	0.485	0.485	0.561
slake_qwen2vl2b	0.409	0.407	0.353	0.406	0.406	0.458
slake_qwen2vl7b	0.327	0.323	0.261	0.356	0.356	0.397
slake_qwen3vl8b	0.626	0.515	0.516	0.566	0.566	0.549
vqa-rad_gemma3_27b	0.668	0.633	0.628	0.649	0.649	0.633
vqa-rad_gemma3_4b	0.605	0.583	0.584	0.599	0.599	0.581
vqa-rad_internvl3_8b	0.598	0.536	0.483	0.570	0.570	0.659
vqa-rad_qwen25vl7b	0.547	0.528	0.519	0.567	0.567	0.631
vqa-rad_qwen3vl8b	0.590	0.572	0.518	0.567	0.567	0.560
macro-mean AUROC	**0.529**	**0.497**	**0.476**	**0.515**	**0.515**	**0.539**
Spearman ρ with CIPE	—	+0.861	+0.862	+0.924	+0.924	+0.781

S11 MEDTRUST-IN TEST-SPLIT REBUILD (V0.3)

Triggered by the screener rule of §4.4. The roster was re-screened on the complete closed yes/no test split of each source dataset through the identical harness, and the corpus rebuilt under the same fail-closed positive control and constant integer criterion; the builder now refuses a build whose dumps mix splits. Receipts: `results/screening_test/`; artefact and datasheet: `data/medtrust_in_test/` (composition: Figure 10). Qwen2-VL-7B is disqualified on VQA-RAD test (acc_delta −0.044), so that roster is five rather than six.

Table 9: **Three corpus builds, and what each control actually did.** v0's guard was a no-op when the flag was absent—`.get(...) is False` passes when the key is absent—and where the flag was false, de-duplication re-admitted the same model's older run (Appendix S27), so it disqualified nothing. v0.2 made the control fail closed and it fired 12 times, but 41 of its 64 dumps still carried no recorded flag and were certified by the derived control alone. v0.3 is the first build where every dump carries the recorded black-vs-white flag. The shrink from v0.2 is the price of that fix, not a regression: the corpus is 29% smaller and every other property improves.

	v0 (train)	v0.2 (train)	**v0.3 (test)**
items	1,197	1,581	**1,122**
yes-skew	86.1%	93.9%	**85.6%**
label-balanced subset	332	194	**324**
screening dumps	64	64	18
missing the control flag	41	41	**0**
model×dataset pairs disqualified	0 (re-admitted)	12	1
`derived`-only items	—	179	**0**

Table 10: **MedTrust-IN v0.3, built from source test splits.** Smaller than the train-split v0.2 (1,122 vs. 1,581), almost entirely because SLAKE's test split holds 355 closed items against 1,681 in train. Every other property improves: the yes-skew falls 93.9% → 85.6%, "no" items rise from 97 to 162, the label-balanced subset grows 194 → 324, and all 1,122 items are certified `recorded+derived` rather than the 179 derived-only items v0.2 carried.

dataset	test pool	roster	curated	yes	no	balanced
vqa-rad	251	5	85	77	8	16
slake	355	6	121	121	0	0
pathvqa	3362	6	916	762	154	308
total			**1,122**	**960**	**162**	**324**

S12 TRAIN-VERSUS-TEST CONTAMINATION PROBE

Receipts: `results/contamination/contam_*.json`; analysis `code/scripts/cipe_contamination_analysis.py`; summary `outputs/cipe_v0/contamination_probe.{txt,json}`. Submitted as 48 runs (8 models × 3 datasets × 2 splits); all 24 cells are complete. The probe reuses the paper's own harness through `cipe_v0_pilot.py -slake_split`, so train and test differ only in split membership. n is matched between splits and capped by the smaller (test) closed yes/no pool: VQA-RAD 250 (of 940 train / 251 test closed), SLAKE 350 (of 1,681 / 355), PathVQA 500 (of 9,751 / 3,362). Items are disjoint across splits, so the gap is bootstrapped as a difference of independent means (10,000 resamples, seed 42). * marks the three cells whose 95% CI excludes zero (one survives BH).

Image benefit across splits. An accuracy gap cannot separate contamination from difficulty drift. A cell's *image benefit* $\mathrm{IB} = \mathrm{acc} - \mathrm{acc}_{\text{text-only}}$ measures the accuracy gained from the image; its cross-split change alone does not distinguish memorisation from difficulty or composition shifts. Over 24 cells the mean train-to-test drop is $+0.020$ (sd 0.066) and **exactly one collapses**: SmolVLM-256M on VQA-RAD, $+0.364 \to +0.080$, a drop of $+0.284$ — 4.0 sd above the mean and $3.6\times$ the largest of the other 23 ($+0.080$). Qwen2.5-VL-32B, which demonstrably reads the image, holds $+0.276$/$+0.308$. This pattern supports a contamination hypothesis for the cell with documented training inclusion, without separating memorisation from other causes of split sensitivity. The $+0.020$ is the observed mean, not a bound on contamination elsewhere. Text-only accuracy ($\mathrm{acc}_{\text{text-only}} = 0.568$) alone does not identify the cause of the image-benefit drop. The other 21 cells are null ($|\mathrm{gap}| \le 0.056$, widest bound 0.132). Completing the probe also retires two claims of our first analysis—split-invariance, and that no screener was implicated.

Is the probe specific? A detector firing on an unexposed model is not detecting exposure. Xu et al. (2026) audit these datasets by *provenance* and find image-side source overlap (19.8% of SLAKE-

Table 11: **One cell carries a large split anomaly.** SmolVLM-256M loses 34 accuracy points moving from VQA-RAD train to test, landing near that split's majority baseline, consistent with documented training inclusion but not by itself identifying memorisation; it is the only cell surviving Benjamini-Hochberg across all 24. Two further cells exclude zero without surviving multiplicity correction—InternVL3-8B ($+0.066$, q=0.251) and Gemma-3-27B ($+0.062$, q=0.289), both on PathVQA—and because both are corpus screeners they fire our rebuild rule (§4.4). The remaining 21 cells are null. Qwen2.5-VL-32B is null on all three datasets, twice with a *negative* gap. CIPE means barely move with split and the ranking is stable, though not identical (Spearman $\rho = 0.952$ VQA-RAD, 0.952 SLAKE, 0.929 PathVQA).

Model	dataset	n/split	$\mathrm{acc}_{\mathrm{tr}}$	$\mathrm{acc}_{\mathrm{te}}$	gap	95% CI	$\mathrm{CIPE}_{\mathrm{tr}}$	$\mathrm{CIPE}_{\mathrm{te}}$
smolvlm_256m*	vqa-rad	250	0.9320	0.5880	+0.3440	[+0.2760, +0.4120]	0.553	0.580
qwen25vl7b	vqa-rad	250	0.7840	0.7280	+0.0560	[−0.0200, +0.1320]	0.627	0.659
qwen25vl32b	vqa-rad	250	0.8240	0.8600	−0.0360	[−0.1000, +0.0280]	0.594	0.584
internvl3_8b	vqa-rad	250	0.6560	0.6240	+0.0320	[−0.0520, +0.1160]	0.620	0.654
qwen3vl8b	vqa-rad	250	0.7520	0.7720	−0.0200	[−0.0960, +0.0560]	0.620	0.659
qwen2vl7b	vqa-rad	250	0.5240	0.5080	+0.0160	[−0.0680, +0.1040]	0.716	0.730
gemma3_4b	vqa-rad	250	0.5440	0.5320	+0.0120	[−0.0760, +0.1000]	0.575	0.589
gemma3_27b	vqa-rad	250	0.6160	0.6240	−0.0080	[−0.0920, +0.0760]	0.512	0.543
qwen25vl7b	slake	350	0.8029	0.7514	+0.0514	[−0.0086, +0.1114]	0.537	0.520
qwen3vl8b	slake	350	0.8314	0.7800	+0.0514	[−0.0057, +0.1086]	0.528	0.505
gemma3_27b	slake	350	0.8143	0.7686	+0.0457	[−0.0143, +0.1057]	0.566	0.536
smolvlm_256m	slake	350	0.5800	0.5343	+0.0457	[−0.0286, +0.1200]	0.558	0.564
gemma3_4b	slake	350	0.7114	0.6857	+0.0257	[−0.0429, +0.0943]	0.534	0.530
qwen25vl32b	slake	350	0.8486	0.8600	−0.0114	[−0.0629, +0.0401]	0.490	0.497
qwen2vl7b	slake	350	0.5371	0.5457	−0.0086	[−0.0829, +0.0629]	0.737	0.737
internvl3_8b	slake	350	0.7171	0.7200	−0.0029	[−0.0686, +0.0657]	0.529	0.519
internvl3_8b*	pathvqa	500	0.7000	0.6340	+0.0660	[+0.0100, +0.1220]	0.652	0.654
gemma3_27b*	pathvqa	500	0.6940	0.6320	+0.0620	[+0.0040, +0.1200]	0.603	0.573
qwen25vl32b	pathvqa	500	0.7320	0.6940	+0.0380	[−0.0180, +0.0960]	0.646	0.628
qwen25vl7b	pathvqa	500	0.6740	0.6460	+0.0280	[−0.0320, +0.0880]	0.694	0.677
smolvlm_256m	pathvqa	500	0.5660	0.5420	+0.0240	[−0.0380, +0.0860]	0.650	0.639
qwen2vl7b	pathvqa	500	0.5900	0.5740	+0.0160	[−0.0460, +0.0760]	0.820	0.809
gemma3_4b	pathvqa	500	0.6160	0.6080	+0.0080	[−0.0520, +0.0680]	0.668	0.664
qwen3vl8b	pathvqa	500	0.6900	0.6940	−0.0040	[−0.0620, +0.0540]	0.660	0.636

En images flagged under SigLIP-B-16), which they read as source or distributional overlap rather than confirmed per-image memorisation, and report that failure for their own detectors, whose signals BLIP-2 (Li et al., 2023b) reproduces despite lacking plausible medical-VQA exposure. Their methods and ours measure different things. Ours admits a known answer: SmolVLM-256M's documented corpus contains VQA-RAD's *train* split (100% of its 1,793 turns, 0 test-exclusive) and neither SLAKE nor PathVQA. **The one cell we flag is the one with documented training-set inclusion**; that model's two other datasets, without disclosed inclusion, are null. The dataset pattern agrees with the disclosures, but model-axis specificity remains untested and two pre-registered controls still cut against us (below).

Bounding the effect. Three things bound it. *(i) It does not generalize*: the other 21 cells have $|\text{gap}| \leq 0.056$ with every CI covering zero, and the widest CI bound over them, 0.132, is the largest single-cell bound observed, not a general upper bound (Qwen2.5-VL-32B is null on all three datasets, twice *negative*; SmolVLM-256M is null on PathVQA). *(ii) Two screener cells are implicated*: SmolVLM-256M is a panel row and never screened a corpus item, but InternVL3-8B and Gemma-3-27B on PathVQA are screeners and exclude zero (marginally, q=0.251 and q=0.289), which fires the rebuild rule. *(iii) The panel ordering is stable but not invariant*: $\rho = 0.952$ on VQA-RAD, 0.952 on SLAKE, 0.929 on PathVQA, and SmolVLM-256M holds the same VQA-RAD rank (second most image-reactive, after Gemma-3-27B) on both splits, so its accuracy anomaly does not move its CIPE rank. We therefore retract our first analysis's blanket claim that every panel-level faithfulness ranking is contamination-confounded—the ranking largely survives the split—and equally withdraw our own previous over-claim that it survives it exactly.

The specificity control, pre-registered. Pre-registration: `PREREG_contamination_specificity_control.md`, committed before the runs; receipts `outputs/cauldron_vqarad_provenance.json` and `outputs/cipe_v0/contamination_control.json`; analysis `code/analysis/contamination_summary.py`, which holds control models out of the panel's reference distribution so a control cell cannot shift the mean and sd it is judged against. All four outcome branches were fixed in advance, including "our detector is not specific either".

Ground truth. Xu et al. (2026) could only compare cohorts of models against each other; we can anchor to a documented corpus. The Cauldron (Laurençon et al., 2024b)—the training mixture named on SmolVLM-256M's model card—ships 50 configs, of which `vqarad` is one and neither SLAKE nor PathVQA appears. Its `vqarad` config has a `train` split only. Verified against the source rather than the listing: its 313 image rows carry 1,793 question turns, exactly the row count of VQA-RAD *train*; 1,635 questions occur only in train, 158 occur in both splits (generic phrasings such as "is there a fracture" recur), 0 are test-exclusive and 0 fail to match. The Cauldron stores questions with a sampled brevity instruction appended ("…\n Answer briefly."), so a whole-string match returns 0% and reads as a refutation; we match per line.

Arm A (external model) — Branch 4, invalid. The failure Xu et al. (2026) report of their own detectors is that "BLIP-2 reproduces the apparent positive signals despite lacking plausible medical-VQA exposure", leaving them "unreliable as standalone membership-inference signals on small medical-VLM cohorts". BLIP-2, that same baseline, cannot run our statistic at all: `Blip2ForConditionalGeneration.forward` takes `pixel_values` as a required positional argument, so the text-only forward that $\text{IB} = \text{acc} - \text{acc}_{\text{text-only}}$ requires is undefined for it, whereas `LlavaForConditionalGeneration` defaults it to `None`. Their statistics are pure-likelihood and run on any model; ours needs a model conditionable on text alone. LLaVA-1.5-7B—documented instruction mixture with no medical VQA corpus and no medical fine-tuning—was the stand-in, and failed the harness's own black-vs-white injection gate at n=8 (image sensitivity 0.0000; $\text{acc} = \text{acc}_{\text{text-only}} = 0.625$; 8/8 predictions "yes" with and without the image). The cause is known in this project: LLaVA-family models place first-token mass on "Yes"/"No" rather than the space-prefixed lowercase ids this readout scores, a readout degeneracy, not a dead image. The consequence for a *control* is the point: its image benefit is identically 0 on both splits, so its transfer drop is exactly 0 and would have read as the cleanest null in the study while carrying no information at all. A specificity control can be silently vacuous, and only a positive control on the control detects it. Idefics3-8B, a non-panel model run through the same path, passes the gate at 0.207, so the failure is model-specific rather than a broken harness.

Arm B (corpus replication) — B-null. SmolVLM-2.2B shares the documented Cauldron mixture at $\sim 9\times$ capacity and passed the injection gate on all six cells (0.023–0.154). Its drops are $+0.124$ (VQA-RAD), $+0.058$ (PathVQA), $+0.023$ (SLAKE), and its VQA-RAD accuracy gap is $+0.120$ ($0.764 \rightarrow 0.644$). The VQA-RAD cell is 1.6 sd above the panel mean and is the largest drop anywhere outside the flagged cell, but it clears neither the committed 0.20 flag nor the pre-registered within-model contrast, so we record B-null: the collapse does not replicate at this capacity. We also record a defect in our own pre-registration rather than resolve it conveniently—the Arm-B criterion was written as "clearly above its own SLAKE/PathVQA drops" with no number attached, while the threshold committed in code is 0.20; the two can disagree, and here they would have if the margin had fallen differently. Both readings are reported.

Class imbalance is not the mechanism. For the flagged SmolVLM-256M/VQA-RAD cell, gold is near-balanced (130/120 train, 133/117 test), so a majority answerer scores ≈ 0.52, not 0.932, and the model's predictions have not collapsed to one class either (145/105 train, 170/80 test).

Decision status. All six MedTrust-IN screeners are now measured (18 screener cells). Two show a gap whose 95% CI excludes zero—InternVL3-8B ($+0.066$, q=0.251) and Gemma-3-27B ($+0.062$, q=0.289)—and both are on PathVQA. Neither survives multiplicity correction on its own, but our pre-registered rule keys on the interval, not on the corrected q, and we do not get to reinterpret the rule after seeing which way it points: **the rule fired and the corpus is rebuilt from source test splits** (§4.4). The concentration on one dataset, which supplies 654 of the 1,581 items, is what makes bounding these cells the weaker option. The remaining four screeners are null (max $|\text{gap}| = 0.056$). SmolVLM-256M, the cell with the largest split anomaly and documented training inclusion, is a panel row and never screened a corpus item.

S13 Panel bookkeeping and protocol detail

Counterfactual detail. C3 modality-swap is deferred to v1 pending a curated cross-modal pair index. In the CIPE *panel* runs (Tables 12, 13, on VQA-RAD/SLAKE/PathVQA) the builder passes no bounding box, so `anatomy_mask` is the centred-square fallback for 100% of items—it is a *centre-occlusion* control, not a diagnostic-region mask, and carries no region-specific claim. Region-specific evidence comes *only* from the separate SLAKE annotated-region control of §4.1 (Table 18), which uses `detection.json` boxes, and whose general-model effect is graded rather than absent under the foreground-matched control (Table 1).

Metric detail. The clip is a bounding convenience so scores share a common $[0, 1]$ range; it *binds* (raw KL>1 nat, discarding tail magnitude) on 13.9% of Qwen2-VL-7B records (15.4% `image_swap`, 25.4% `text_only`, 0.8% `anatomy_mask`), so the clip mostly affects the strong-shift tail; the 0.7 `frac_high` cut used in the tables is a *descriptive* reporting threshold, not a calibrated decision boundary. The metric and variant generators are unit-tested (68/68 tests); the code will be released upon publication.

Models evaluated and bookkeeping. The v0 work evaluates 50 unique open-source VLMs (63 evaluations after retry replicates) spanning eleven architecture families, plus the two medical specialists MedGemma-4B/27B-it (§4.1)—the complete set of models evaluated for any number in this paper (a broader proprietary / medical-specialist panel and the six-dimension protocol are deferred v1 scope, run for no result here; Appendices S18, S17). Fixing the recurring counts once, and mechanically: every number in this paragraph is regenerated from the per-record dumps by `code/scripts/regen_model_counts.py`, which fails loudly rather than let the paper and the artefact drift apart again. Of 80 attempted runs, 63 produced valid parses over 50 distinct models. Identity here is the HuggingFace `model_path`, not the filename tag: three tags name one Qwen3-VL-32B and two name one Qwen3-VL-30B-A3B, so a tag-based count reads 52 and double-counts two models. Table 13 names 40 of the 50 and Table 20 the 8 readout-artefact models; the remaining two (Qwen2.5-VL-32B, Llama-3.2-11B-Vision) are not tabulated but appear in the per-evaluation summary `outputs/cipe_v0_summary_latest.csv` (to be released), so no evaluated model is unreported. Further, 17 runs produced zero parsed records (BLIP-2, InstructBLIP$\times$2, Kosmos-2/2.5, Idefics2-base, InternVL2-2B, GLM-4V-9B, GLM-Edge-V-5B, Phi-3.5-Vision$\times$2, PaliGemma$\times$6) and are excluded as *evaluation failures*, not zero-CIPE: each was forced past our

family-detection guard with an explicit generic adapter, so those zeros are a property of our harness and not of the models. The scaling fit (§4.3) is headlined on the 21 independently certified models, with all 40 non-saturated models with disclosed parameter counts for comparison only, one run per model — checked, not assumed: those 40 rows carry 40 distinct `model_path` values, so the n contains no pseudo-replication. The per-item construct validation (Table 14) uses the n=500 re-runs. Our first analysis called these 21 evaluations a "clean panel" although they had never been run through the positive control; 4 of the 21 in fact fail it—including the confirmed image-blind SLAKE LLaVA-OneVision run—so they are excluded, leaving 17. The headline is unchanged by the correction (macro-mean AUROC 0.5291 over 21 vs. 0.5295 over the 17 certified; decircularized $0.537 \rightarrow 0.533$), so we report the certified 17 and note that the number was not resting on the uncontrolled rows. CIPE results are point estimates over n samples $\times$ three counterfactuals; the three Qwen2/2.5-VL backbones marked † in Table 13 use n=500, all other rows n=200. (Qwen2.5-VL-3B's displayed row is n=200; its separate n=500 robustness run is reported in App. S39.) All inference is HuggingFace Transformers at temperature 0 over yes/no logprobs (zero-cost local). The per-family CIPE comparison is reported *descriptively*, not inferentially: several rows are byte-identical retry replicates, so treating evaluations as independent would be pseudo-replication (we deduplicate before the scaling fit, §4.3); sub-2pp `frac_high` differences are treated as ties. A sample-size robustness check at n=500 (1,500 records each) on four backbones preserves the within-family ordering, the larger model having the lower mean CIPE (Qwen2-VL-2B mean CIPE 0.855 vs. 7B 0.708; Qwen2.5-VL-3B 0.794 vs. 7B 0.633).

S14 Scope of the submission

To be explicit about what is load-bearing: this is an **audit-only** submission with a corpus artefact, not a metric paper and not a completed multi-dimensional benchmark. We do not ask the reader to adopt CIPE—we tested it against cheaper readouts and report that it has no demonstrated advantage (§4.3). What we ask them to adopt is the annotated-region sensitivity protocol, the four controls, and the corpus. What the v0 work establishes is (i) the readout, image-injection and circularity controls, with CIPE reported as a negative; (ii) the documented readout artefact and its recovery; (iii) a single-dataset annotated-region sensitivity audit under a foreground-matched control, read through total variation; (iv) an exposure-confounded two-scale MedGemma result; and (v) the MedTrust-IN corpus of 1,122 panel-screened *text-only-hard* items built on source *test* splits (MedTrust-IN v0.3; not an image-sufficiency or human-adjudicated necessity label). Reported v0 measurements draw from three established medical VQA datasets (VQA-RAD, SLAKE, PathVQA); the visual-faithfulness axis is measured via CIPE across 50 open-source models on VQA-RAD. The four-dataset scope that adds OmniMedVQA is deferred design, not measured here. Everything else—the original six-dimension design, composite trustworthiness score, pre-registered composition targets, annotator study, and leaderboard (Appendix S17)—is *deferred future scope*, retained only as CIPE stratification axes, and should not be read as completed work. Readers should weigh only (i)–(v).

S15 First-model CIPE detail (Qwen2-VL-7B)

First model. On a single representative open VLM, CIPE first appeared to expose behaviour accuracy misses: 66.9% of samples score above the 0.7 threshold (median 0.924), and counter-intuitively `text_only` is the *most* reactive counterfactual (mean CIPE 0.463), not the least—removing the image shifts the answer more than swapping or centre-occluding it, suggesting the model treats image *presence*, not content, as a confidence cue (the "image-presence cue," Appendix S6). Per-counterfactual numbers and the distribution figure follow (Table 12, Figure 4).

Qwen2-VL-7B-Instruct on 500 closed-form yes/no VQA-RAD items under three counterfactuals; CIPE values where 1.0 = "model ignores the intervention" and 0.0 = "model fully reacts".

S16 Full clean-panel CIPE table

Per-family descriptive patterns. The panel spans eleven architecture families (Qwen-VL (Wang et al., 2024; Bai et al., 2025b;a), InternVL3 (Zhu et al., 2025), Gemma3 (Gemma Team, 2025), Idefics3/SmolVLM (Laurençon et al., 2024a; Marafioti et al., 2025), LLaVA-arch (Liu et al.,

Table 12: CIPE v0 on Qwen2-VL-7B-Instruct (VQA-RAD, n=500 samples per counterfactual). frac_high = fraction with CIPE > 0.7 (a descriptive cut). All three counterfactuals yield CIPE distributions strongly skewed towards 1.0; `anatomy_mask` (a centred 25%-area square here, not the diagnostic region, §3) is the most striking, with 97% of samples scoring CIPE > 0.7.

Counterfactual	**mean CIPE**	**median**	**frac_high** (> 0.7)	n
image_swap (C1)	0.712	0.944	68.2%	500
anatomy_mask (C2)	**0.950**	**0.989**	**97.0%**	500
text_only (C4)	0.463	0.495	35.6%	500
aggregate	0.708	0.924	66.9%	1500

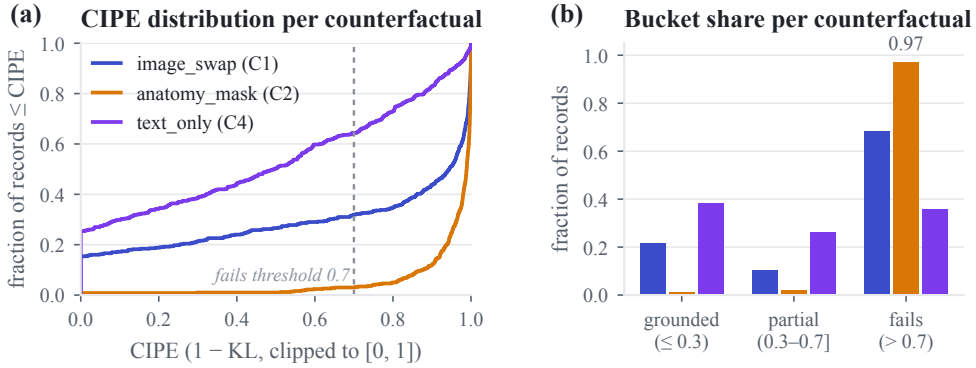

Figure 4: **Centre-square occlusion barely moves the first model's answers.** CIPE v0 on Qwen2-VL-7B-Instruct (VQA-RAD, n=500 records per counterfactual), drawn from `outputs/cipe_v0/results_qwen2vl_7b.json`. (a) Empirical CDF of per-record CIPE for each counterfactual; the dashed line marks the descriptive 0.7 cut. (b) Share of records per CIPE bucket (grounded ≤ 0.3 / partial 0.3–0.7 / fails > 0.7). Centre-occlusion masking leaves 97% of records in the "fails" bucket; on VQA-RAD this mask is a centred square, so region-precise claims rest on the annotated-region control of §4.1.

2024a;b; Li et al., 2024a), Pixtral (Agrawal et al., 2024), Mistral-Small, LFM2-VL (Amini et al., 2025), plus Fuyu, Aria-MoE (Li et al., 2024b) and VipLLaVA (Cai et al., 2024) in the readout-artefact cluster of §4.2, with Idefics2 (Laurençon et al., 2024b)). Two descriptive patterns hold: within families, scale generally lowers `frac_high` (Qwen2/2.5-VL ∼13pp from 2B to 7B, InternVL3 77.3%→59.5%, Gemma3 56.7%→50.7%; Qwen3-VL is flat), and LLaVA-arch/Pixtral sit highest (87.7–98.7%) while the non-reasoning Qwen-VL/InternVL3/Gemma3/Idefics3/SmolVLM rows occupy 47.0–80.3% (Qwen3-VL-2B-Thinking, at 95.2%, is an exception; `text_only` separates cleanest). These are *descriptive*, not an "architecture dominates scale" claim—which in our first analysis rested on the readout-artefact cluster. On the scaling null: a linear term may be the wrong instrument, because InternVL3, the one family we follow from 1B to 78B, is U-shaped rather than monotone (Table 13), with both legs surviving an item-clustered bootstrap. A single slope averaged over families whose CIPE curve turns inside their own range is expected to be small and noisy, so the p=0.064 null is weak evidence about scale, not evidence that scale does not matter. A larger, retry-deduplicated panel is needed to resolve the question (§4.3).

Full per-model CIPE across the open-source panel follows.

The per-model CIPE→correctness construct-validation table follows; the honest summary is the pooled row (mean AUROC 0.529).

The naive (circular) stratified scatter, retained as the "before" panel of the circularity audit:

De-inversion collapse table:

Table 13: CIPE across the open-source vision-language model panel on VQA-RAD (lower CIPE = larger answer shift under the counterfactuals). `frac_high` is the fraction of records with CIPE>0.7. Every cell is generated from the per-record dumps by `code/scripts/regen_table2_panel.py`; no cell is hand-entered. The panel spans Qwen-VL (three generations), InternVL3, Gemma3, Idefics3/SmolVLM, and LLaVA-arch, at scales 256M–78B. Two patterns are worth noting. (i) *Within a family, `frac_high` mostly falls with scale, but not monotonically*: Qwen2/2.5-VL fall 13.4pp from 2B to 7B and Gemma3 falls $56.7\% \to 50.7\%$ from 4B to 27B, but InternVL3 — the one family we can follow across nearly two orders of magnitude — is *U-shaped*: it falls $77.3\% \to 59.5\%$ from 1B to 8B and then climbs back to 68.5% at 78B. Both legs survive an item-clustered bootstrap (1B→8B -17.8pp, 95% CI $[-22.3, -13.5]$; 8B→78B $+9.0$pp, CI $[+4.7, +13.2]$; 10,000 resamples), so the reversal is not item-sampling noise within these single runs; it is a U-shape in CIPE `frac_high`, not in faithfulness: CIPE is not validated (§4.3) and InternVL3-78B is outside the certified subset (Appendix S29). The scaling trend is reported on the 21 independently certified models (§4.3): directional, two specifications disagreeing (MixedLM p=0.033, family-FE OLS p=0.237), so no significance is claimed — and the InternVL3 curve is one reason a *linear* scaling term should not be expected to reach significance here. Qwen3-VL is flat across 2B–32B (55.3–58.3%). (ii) *LLaVA-arch models sit high* (LLaVA-OV-7B 98.7%, LLaVA-NeXT-Vicuna-7B 88.7%), while the non-reasoning Qwen-VL/InternVL3/Gemma3/Idefics3 rows occupy 50.2–80.3%. That separation is *not* clean, and we flag the exception rather than scope it away: Qwen3-VL-2B-Thinking reaches 95.2%, inside the LLaVA-arch band, so architecture does not partition this panel. `text_only` is the cleanest separator. We report these as *descriptive* within- and cross-family patterns, not as an "architecture dominates scale" claim—our first analysis's version of that claim rested on a five-model cluster whose saturation is a readout artefact in the three models we could re-audit (re-scored with a robust readout; Aria and VipLLaVA unresolved, §4.2). †Evaluated at n=500 samples per counterfactual (sample-size robustness check, Appendix S13); all other rows use n=200.

Model	mean CIPE	median	frac_high$_{>0.7}$	img_swap	anatomy_mask	text_only
Qwen-VL family (in-family scale + generation axis)						
Qwen2-VL-2B†	0.855	0.986	80.3%	0.952	0.987	0.626
Qwen2.5-VL-3B	0.793	0.946	73.7%	0.815	0.965	0.598
Qwen2-VL-7B†	0.708	0.924	66.9%	0.712	0.950	0.463
Qwen2.5-VL-7B†	0.633	0.861	60.7%	0.638	0.876	0.385
Qwen3-VL-2B	0.606	0.809	55.3%	0.589	0.851	0.378
Qwen3-VL-4B	0.601	0.862	58.3%	0.576	0.765	0.463
Qwen3-VL-8B	0.619	0.876	58.0%	0.579	0.759	0.520
Qwen3-VL-32B	0.595	0.911	57.0%	0.552	0.831	0.401
Qwen3-VL-30B-A3B-Inst.	0.620	0.884	60.3%	0.594	0.873	0.392
Qwen3-VL-2B-Thinking	0.939	0.990	95.2%	0.913	0.979	0.925
Qwen3-VL-4B-Thinking	0.674	0.895	62.2%	0.798	0.873	0.351
Qwen3-VL-8B-Thinking	0.732	0.931	71.3%	0.867	0.940	0.389
Qwen3-VL-30B-A3B-Thnk	0.610	0.772	56.3%	0.694	0.833	0.304
InternVL3 family (scale axis, 1B–78B)						
InternVL3-1B-hf	0.811	0.941	77.3%	0.818	0.919	0.696
InternVL3-2B-hf	0.787	0.933	75.2%	0.720	0.944	0.698
InternVL3-8B-hf	0.630	0.900	59.5%	0.636	0.934	0.320
InternVL3-14B-hf	0.690	0.910	65.8%	0.701	0.936	0.434
InternVL3-38B	0.668	0.929	64.5%	0.788	0.967	0.247
InternVL3-78B	0.710	0.939	68.5%	0.852	0.963	0.314
Gemma3 family						
Gemma3-4B-it	0.578	0.924	56.7%	0.599	0.807	0.327
Gemma3-12B-it	0.527	0.808	52.0%	0.573	0.693	0.316
Gemma3-27B-it	0.507	0.735	50.7%	0.517	0.831	0.174
Gemma3n-E2B-it	0.580	0.762	52.5%	0.522	0.732	0.487
Gemma3n-E4B-it	0.563	0.722	50.2%	0.541	0.717	0.431
Idefics3 / SmolVLM / other						
Idefics3-8B	0.695	0.876	63.7%	0.669	0.792	0.623
SmolVLM-Inst.-1.7B	0.690	0.864	65.3%	0.747	0.776	0.547
Mistral-Small-3.1-24B	0.757	0.984	73.2%	0.795	0.963	0.513
SmolVLM2-2.2B	0.719	0.921	70.0%	0.733	0.852	0.573
SmolVLM-Synthetic	0.805	0.936	77.8%	0.787	0.866	0.763
SmolVLM2-500M-Video	0.782	0.928	71.8%	0.822	0.900	0.625
SmolVLM-256M	0.551	0.622	47.0%	0.625	0.651	0.378
LLaVA-arch						
Pixtral-12B	0.891	0.994	87.7%	0.958	0.987	0.727
LLaVA-NeXT-Vic.-7B	0.883	0.961	88.7%	0.887	0.946	0.815
LLaVA-NeXT-Video-7B	0.911	0.975	92.0%	0.926	0.951	0.857
LLaVA-OV-0.5B	0.953	0.992	96.3%	0.979	0.986	0.893
LLaVA-OV-0.5B-SI	0.971	0.994	98.7%	0.982	0.990	0.942
LLaVA-OV-7B-SI	0.942	0.983	95.5%	0.915	0.968	0.943
LLaVA-OV-7B	0.969	0.992	98.7%	0.962	0.988	0.957
LFM2-VL-1.6B	0.852	0.962	83.2%	0.811	0.951	0.794
LFM2-VL-450M	0.890	0.977	87.8%	0.904	0.969	0.797

Table 14: **CIPE construct validation: a weak item-level correctness signal (macro-mean AUROC ≈0.53).** `acc` is overall closed-form yes/no accuracy on the n=200 VQA-RAD pilot subset; AUROC(CIPE → correct) is the rank-AUROC of per-item CIPE predicting a correct answer; `acc|lo`/`hi` are accuracies on the below-/above-median-CIPE halves. Recall CIPE=1−clip(KL), so *high* CIPE = image-*invariant* (language-prior) items and *low* CIPE = image-*reactive* items. The top rows are the *tail* of the distribution (most capable models), where AUROC reaches 0.66–0.70; they are *not* typical. The honest summary is the pooled row: over the 17 *certified* n=500 model×dataset evaluations (VQA-RAD/SLAKE/PathVQA; 4 of the original 21 fail the positive control and are excluded) the macro-mean AUROC is 0.529 (0.533 decircularized, i.e. CIPE without `text_only`), i.e. CIPE is only weakly predictive of correctness at the item level. Under the naive readout the readout-artefact cluster (CIPE≡1.0, no variance) sits at AUROC=0.5 by construction, so it is not used in this table.

Model	**acc**	**AUROC**(CIPE → corr)	**acc**|lo	**acc**|hi
Qwen2.5-VL-32B	0.820	**0.700**	0.730	0.910
Qwen3-VL-30B-A3B	0.770	0.675	0.670	0.870
Gemma3-27B-it	0.630	0.666	0.530	0.730
Qwen3-VL-32B	0.710	0.665	0.580	0.840
Qwen3-VL-8B	0.750	0.627	0.660	0.840
InternVL3-8B-hf	0.675	0.623	0.600	0.750
Gemma3-12B-it	0.615	0.615	0.540	0.690
Qwen3-VL-2B	0.635	0.600	0.550	0.720
Excluded readout-artefact cluster (CIPE≡1.0)	0.485	0.500	0.485	—
Pooled (certified n=17)	—	mean **0.529** (decirc. 0.533), median 0.547, 65% (11/17) >0.5		

Table 15: **The strong "de-inversion" is largely a circularity artefact.** Per-model AUROC(CIPE → correct) within each text-only-answerability stratum, n=500 per dataset. *Naive* columns use the same-model text-only stratifier and the full three-variant CIPE (which *includes* `text_only`)—metric and stratifier share a term. Here H1 (text-only-hard) sits far below 0.5 and H3 (language-answerable) far above, appearing to confirm a decisive de-inversion. *Decircularized* columns remove the shared term (CIPE without `text_only`) and use a leave-one-out panel-consensus stratifier. The effect largely collapses: H1 rises to 0.43–0.51 (**de-inversion gone on VQA-RAD and SLAKE**, surviving only on PathVQA) and H3 falls to 0.51–0.61. The honest reading is a *weak, dataset-dependent* faithfulness signal that survives only on PathVQA (Table 24), not the decisive same-metric-opposite-direction result we first reported.

Dataset (n=500)	**Aggregate** (mixed strata)	**H1 text-only-hard** naive	decirc.	**H3 language-answerable** naive	decirc.
VQA-RAD	0.586	0.216	**0.512**	0.832	0.611
SLAKE	0.524	0.215	**0.469**	0.766	0.506
PathVQA	0.470	0.119	**0.431**	0.844	0.512
Pooled	0.527	0.186	**0.47**	0.813	0.54

Naive = self text-only stratifier × full CIPE; decirc. = leave-one-out consensus stratifier × CIPE without `text_only`. Removing `text_only` from the *metric* accounts for most of the change (e.g. VQA-RAD H1 0.216→0.549 with the self stratifier held fixed).

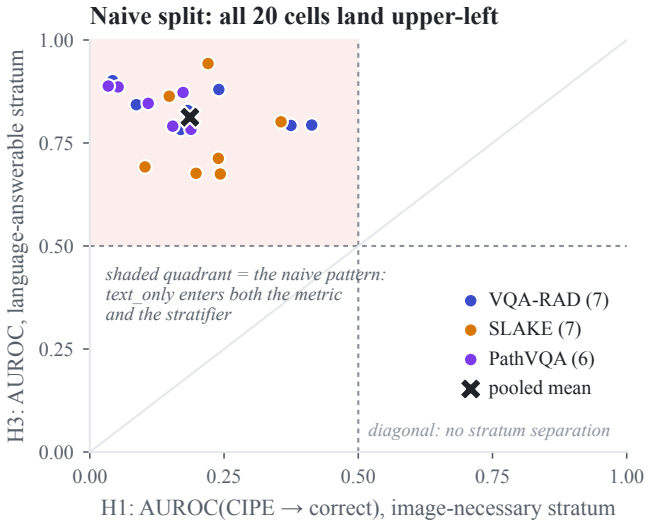

Figure 5: **The naive (circular) stratified analysis**, shown for transparency. Per-model AUROC(CIPE → correct) on the text-only-hard stratum (x, H1) versus the language-answerable stratum (y, H3), one point per (dataset, model) coloured by dataset (n=500 each; 20 cells, recomputed from the committed dumps by `code/scripts/fig_h1h3_naive_scatter.py`). With the same-model text-only stratifier and full CIPE, every cell lands in the upper-left quadrant, which we first read as decisive. Because `text_only` enters both the metric and the stratifier, this separation is largely manufactured: decircularizing (Table 15) moves the H1 cloud toward 0.5 and the effect survives only on PathVQA. We retain the figure as the "before" panel of the circularity audit.

S17 Benchmark design: six trustworthiness dimensions, composite score, and pre-registered composition (deferred to v1)

This appendix records the legacy six-dimension benchmark design. **None of it was evaluated for this paper**: no dimension below was run, no number in the paper depends on it, and it is retained only as a record of scope. Its per-dimension protocols, the composite trustworthiness score and the pre-registered composition targets are deferred to v1.

S17.1 Overview and Design Principles

MedTrust is organized around six trustworthiness dimensions, each motivated by a concrete clinical requirement and operationalized through a dedicated evaluation protocol. Our design principles are:

1. **Clinical grounding.** Each dimension addresses a failure mode that would be consequential in real clinical workflows (e.g., an overconfident wrong diagnosis that pre-empts further workup).
2. **Complementarity.** The six dimensions are chosen to be conceptually distinct and to cover different aspects of model behavior—input-side robustness, output-side calibration, behavioral consistency, evidential faithfulness, and knowledge correctness.
3. **Source diversity.** We draw on four established medical VQA datasets spanning multiple imaging modalities and clinical specialties to ensure that our findings are not artifacts of a single data distribution.
4. **Reproducibility.** All evaluation protocols would be fully automated, requiring no human judgment at inference time, with data and code released alongside the benchmark.

Figure 7 separates source screening from the measured readout and deferred design. The following subsections specify the deferred six-dimension design; they are not additional measured results.

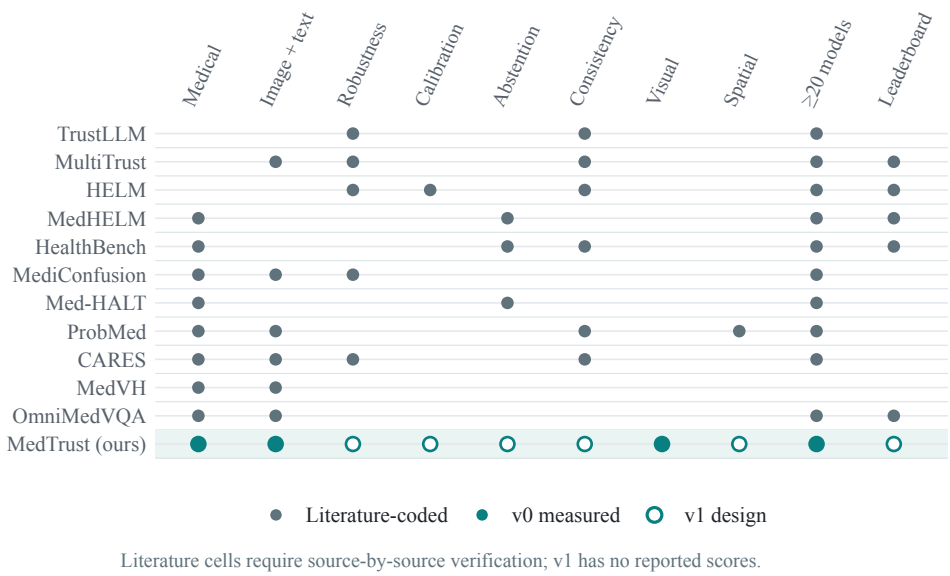

Figure 6: **Scope of benchmark evaluation.** Filled gray marks are literature-coded from each benchmark's own paper and have not been re-verified source by source. Filled teal denotes measured v0 capabilities; open teal denotes registered v1 designs without v0 scores. The matrix is descriptive, not a result table or claim of empirical superiority. Source: the 04_make_figures.py comparison inventory.

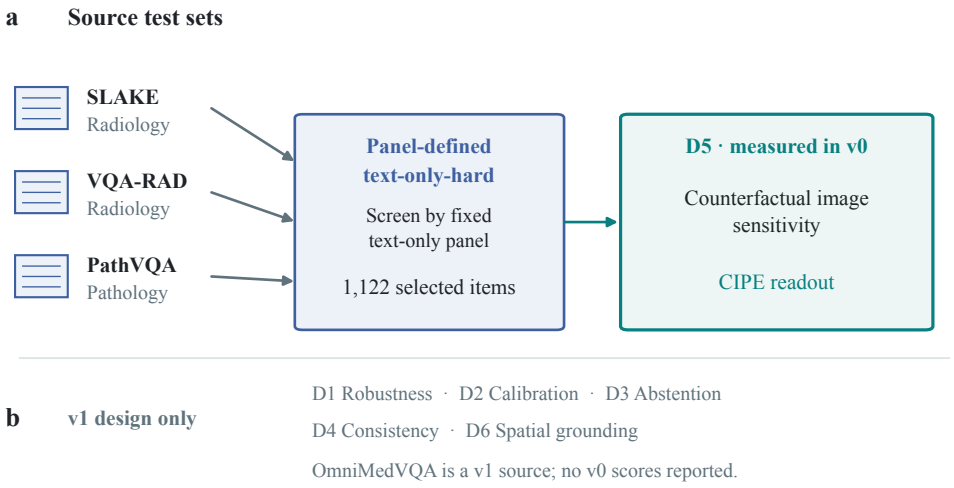

Figure 7: **Source screening and the measured v0 readout.** Three released source test sets are screened with a fixed text-only panel, producing 1,122 panel-defined text-only-hard items. This selection does not establish image necessity. The v0 study measures D5 counterfactual image sensitivity with CIPE. D1–D4/D6 and the OmniMedVQA source are v1 designs, with no v0 scores reported.

S17.2 Dimension 1: Adversarial Robustness

Adversarial robustness measures a model's ability to maintain correct predictions when inputs are perturbed in ways that should not change the ground-truth answer. In the medical setting, this is especially critical: slight variations in image acquisition (different scanners, contrast levels) or question

phrasing should not alter a diagnosis, and the field has documented multiple high-impact attack vectors specific to clinical AI (Finlayson et al., 2019; Bagdasaryan et al., 2023; Clusmann et al., 2025). We additionally draw methodological inspiration from the broader LLM adversarial-evaluation literature: Anil et al. (2024) show that adversarial pressure can be amplified power-law-style by stuffing the context with misleading exemplars, and Hubinger et al. (2024) construct minimal-edit triggers that survive safety training. We construct three types of adversarial probes inspired by these threat models:

Confusing image pairs. Following the methodology of MediConfusion (Sepehri et al., 2025), we identify pairs of visually similar medical images from different diagnostic categories within OmniMedVQA and SLAKE. A model that relies on superficial visual features rather than diagnostically relevant patterns will confuse these pairs. We will generate 2,000 confusing pairs (pre-registered target; not built in v0) and measure the drop in accuracy compared to the original (non-confusing) test set.

Image perturbations. We apply clinically plausible image perturbations—Gaussian noise (simulating low-dose acquisition), contrast adjustments ($\pm$30%), rotation ($\pm$15°), and JPEG compression (quality 10–30)—to each test image. The adversarial robustness score is the fraction of originally correct predictions that remain correct after perturbation, averaged across perturbation types.

Misleading text prompts. We prepend misleading context to questions (e.g., "A colleague mentioned this looks like pneumothorax. [original question]") and measure the rate at which models change correct answers to incorrect ones. This tests susceptibility to anchoring bias, a well-documented cognitive bias in clinical medicine.

S17.3 Dimension 2: Calibration

A well-calibrated model's stated confidence should match its empirical accuracy: if a model reports 80% confidence across a set of predictions, approximately 80% of those predictions should be correct. Calibration is especially important in medicine, where downstream clinical decisions depend on how much to trust a model's output (Guo et al., 2017). We follow the P(True) self-evaluation lineage of Kadavath et al. (2022) for closed-form probes, the verbalized-confidence protocol of Tian et al. (2023); Si et al. (2023); Xiong et al. (2024) for open-ended probes, and the generative-verifier formulation of Zhang et al. (2025) as a soft-confidence baseline. Med-Gemini (Saab et al., 2024) is the closest medical precedent: it uses the entropy of sampled answers to decide when to trigger web search, improving MedQA accuracy.

We measure calibration using three complementary metrics:

- **Expected Calibration Error (ECE)** (Pakdaman Naeini et al., 2015): We partition predictions into $M = 15$ equal-width confidence bins and compute ECE $= \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|$, where B_m is the set of predictions in bin m.
- **Brier Score**: BS $= \frac{1}{N} \sum_{i=1}^{N} (f_i - o_i)^2$, capturing both calibration and refinement.
- **Adaptive ECE (AECE)**: Using equal-mass (rather than equal-width) bins to mitigate bin-count sensitivity.

For closed-source models that do not expose logits, we will employ a verbalized confidence elicitation protocol (Tian et al., 2023; Xiong et al., 2024): after generating an answer, the model is prompted to state its confidence on a 0–100 scale. We pre-register a validation of this approach that compares verbalized and logit-based confidence for open-source models where both are available, with a pre-registered target rank correlation of $\rho \geq 0.6$ and an analysis of absolute-scale offsets (we expect verbalized confidence to over-estimate accuracy in the high-confidence regime (Xiong et al., 2024)). This calibration dimension is deferred to v1 and is not measured in the v0 CIPE release.

S17.4 Dimension 3: Selective Abstention

In clinical practice, recognizing the limits of one's competence is a fundamental professional obligation. A trustworthy medical VLM should be able to abstain from answering when it is likely to

be wrong or when a question falls outside its competence (Geifman & El-Yaniv, 2017; Yang et al., 2024; Wen et al., 2025). We would adopt the ternary grading of SimpleQA (Wei et al., 2024) (and Med-HALT's fake-question test (Pal et al., 2023)), treating *not-attempted* as a first-class outcome rather than equating it with an incorrect answer: a prediction is *correct*, *incorrect*, or *abstained*, and a perfect-precision selective predictor minimizes incorrect at the cost of fewer attempts.

We evaluate selective abstention through two sub-tasks:

Out-of-distribution detection. We present models with images from medical domains not represented in their training data (e.g., unusual imaging modalities, rare pathologies synthesized from MedTrinity (Xie et al., 2025)) and measure whether they appropriately express low confidence or explicitly decline to answer. We compute the AUROC of the model's confidence scores for distinguishing in-distribution from OOD inputs.

Unanswerable question detection. We will construct 3,000 unanswerable questions (pre-registered target; not built in v0) by (a) pairing questions with unrelated images, (b) asking about anatomical structures not visible in the image, and (c) posing questions whose answer requires information beyond what is depicted. A correct response is any form of abstention ("I cannot determine this from the image," "insufficient information," etc.). We report the abstention rate on unanswerable questions (sensitivity) alongside the false-abstention rate on answerable questions (specificity).

S17.5 Dimension 4: Consistency

A trustworthy model should produce semantically equivalent answers to semantically equivalent questions, and should appropriately change its answer when the question semantics change. Methodologically, our paraphrase consistency rate generalises PARAREL (Elazar et al., 2021) to medical VQA, and our negation-flip and leading-belief tests draw on the sycophancy-evaluation methodology of Sharma et al. (2024).

Paraphrase invariance. For each question in the test set, we will generate three paraphrases using a separate LLM with manual quality review, targeting 12,000 paraphrased question sets (pre-registered target; not built in v0). Consistency is measured as the fraction of question groups where all paraphrases elicit the same answer as the original question. We report both exact-match consistency and semantic consistency (assessed via BERTScore).

Negation sensitivity. We will negate 2,000 yes/no questions (pre-registered target; not built in v0) (e.g., "Is there a fracture?" $\rightarrow$ "Is there no fracture?") and measure whether the model appropriately flips its answer. A model that answers "yes" to both the original and negated question is inconsistent. We report the negation flip rate (fraction of correctly inverted answers).

S17.6 Dimension 5: Visual Faithfulness

Medical VLMs should base their answers on the visual evidence in the provided image, not on textual priors or statistical shortcuts. This dimension assesses whether models genuinely "look" at the image, building on the object-hallucination evaluation lineage of CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023c), which we adapt from natural images to medical findings. Med-PaLM M's radiology evaluation (Tu et al., 2024) has radiologists count omissions separately from erroneous findings; the design would keep the same split when scoring open-ended faithfulness probes.

Text-only baseline comparison. We evaluate each model twice on the same questions: once with the image provided and once with the image replaced by a blank (white) image or omitted entirely. The *visual reliance score* is defined as the accuracy gain from including the image: $\Delta_{\text{vis}} = \text{Acc}_{\text{with-image}} - \text{Acc}_{\text{text-only}}$. A low Δ_{vis} indicates the model is relying on textual priors rather than visual evidence.

Shuffled image test. We pair each question with a randomly selected image from a different patient and different pathology. The *visual faithfulness rate* is the fraction of questions where the

model changes its answer when presented with a mismatched image. A model that gives the same answer regardless of the image is not visually faithful. We additionally measure the accuracy drop under image shuffling: a large drop indicates genuine image utilization.

S17.7 Dimension 6: Factual Grounding

Beyond providing correct answers, models should be able to ground their responses in specific visual evidence. We evaluate factual grounding through spatial localization tasks, complementing the radiology factuality literature: RadGraph (Jain et al., 2021) for entity-relation extraction, RaTEScore (Zhao et al., 2024) for entity-aware scoring with negation sensitivity, and the grounded report-generation methodology of MAIRA-2 / RadFact (Bannur et al., 2024) for sentence-level entailment plus spatial precision/recall.

For a planned subset of 1,500 images with bounding-box annotations (pre-registered target; drawn from SLAKE and supplemented with annotations generated via MedSAM (Ma et al., 2024) with expert review; not built in v0), we prompt models to identify the region of the image relevant to their answer. We evaluate grounding quality using:

- **Pointing accuracy**: Whether the model's indicated region (center point or bounding box) overlaps with the ground-truth annotation (IoU > 0.5).
- **Grounding consistency**: Whether the grounded region is consistent with the textual answer (e.g., if the model says "left lung consolidation," the grounding should be in the left lung).
- **Hallucinated grounding rate**: The fraction of incorrect answers where the model nonetheless provides a confident spatial grounding, indicating confabulation.

S17.8 Composite Trustworthiness Score

While MedTrust is fundamentally multi-dimensional, practitioners frequently need a single ranking number for procurement or model selection. We define the *composite trustworthiness score* (CTS) as the geometric mean of the six normalized dimension scores plus accuracy, with calibration entered as $1 - \text{ECE}$:

$$\text{CTS}(m) = \exp\left(\frac{1}{7}\sum_{d\in\mathcal{D}\cup\{\text{acc}\}} \log s_{m,d}\right), \qquad s_{m,d} \in [0,1]. \tag{5}$$

The geometric mean has two desirable properties for trustworthiness summaries that an arithmetic mean does not: (i) it is dominated by the *worst* axis, so a model that catastrophically fails on a single dimension cannot mask the failure with strength on the others; and (ii) it is unit-free, supporting future axis additions without re-normalisation. We treat CTS as an *advisory ranking only*—the dimension breakdown remains the authoritative output of MedTrust and is what we recommend reporting in any deployment decision.

S17.9 Data Statistics

Table 16 summarizes the composition of MedTrust across source datasets and trustworthiness dimensions.

The benchmark covers imaging modalities including X-ray, CT, MRI, ultrasound, fundoscopy, dermoscopy, and histopathology. Question types span closed-ended (yes/no, multiple choice) and open-ended formats, with closed-ended questions predominating in the calibration and consistency dimensions to enable unambiguous automated scoring.

S18 Evaluated and Deferred Model Panels

What was evaluated in this paper. The single source of truth for the models actually run in this paper is the open-source panel of Table 13 together with the recovered readout-artefact cluster (Table 20; model-wise diagnostics in Figure 8)—50 unique open-source VLMs over 63 evaluations in total, of which those two tables name 48; the per-evaluation summary

Table 16: **Pre-registered design targets for the deferred MedTrust benchmark (not built).** Numbers are pre-registered QA-pair targets per source dataset and trustworthiness dimension, not measured composition of a finalised benchmark. Targets were set prior to any model evaluation to prevent post-hoc tuning. The v0 release reports CIPE measurements (Section 3) on 200 VQA-RAD closed-form yes/no samples per model (across 63 model evaluations; no trust-dimension stratification in v0), pending full curation.

Source	Adv. Rob.	Calibration	Abstention	Consistency	Vis. Faith.	Grounding
SLAKE	2,500	1,800	1,000	3,200	1,800	800
VQA-RAD	1,800	1,200	600	2,400	1,200	400
OmniMedVQA	4,000	3,500	1,000	4,500	3,500	200
PathVQA	2,200	2,000	400	2,400	2,000	100
Total (pred.)	10,500	8,500	3,000	12,500	8,500	1,500

Underlined purple values are pre-registered design targets (not measured); see Table 13 for measured CIPE scores on the v0 VQA-RAD sample (200 samples per model, three † rows at n=500).

`outputs/cipe_v0_summary_latest.csv` is the complete record (§3)—and the two medical specialists MedGemma-4B-it and MedGemma-27B-it (§4.1). No proprietary API model and no medical-specialist model other than MedGemma was evaluated for any number reported in this paper.

Deferred v1 panel (not evaluated here). The original benchmark design envisaged a broader panel adding proprietary APIs (GPT-5, OpenAI o3, Claude 4.1, Gemini 2.5 Pro, GLM-4.5V) and further medical specialists (LLaVA-Med, BiomedGPT, Med-Flamingo, RadFM, CheXagent), together with the six-dimension protocol (adversarial robustness, calibration, selective abstention, consistency, visual faithfulness, factual grounding) of Appendix S17. *None of this panel or protocol was run for the present paper*; it is future scope, and we report no snapshot, cost, or metric for it. Readers should weigh only the actually-evaluated panel above.

S19 Per-architecture-family CIPE visualisation

S20 Anatomy- and modality-stratified CIPE (descriptive)

The panel-level CIPE statistics in Table 13 aggregate over anatomy and modality. We compute a heuristic anatomy/modality stratification by keyword-tagging the VQA-RAD questions of the panel runs (regex over {head/neck, chest, abdomen, pelvis, spine, limb} for anatomy and {CT, MRI, X-ray, ultrasound} for modality; `outputs/cipe_per_anatomy_modality.csv`, 1,899 rows). These questions come from the split the harness loads by default, which is `train`; every VQA-RAD receipt we hold locally records `split: train`, but the general-panel dumps are not held locally, so we cannot confirm the split for every row. Restricting to bins with $n \geq 10$ items per cell yields 252 usable (model, anatomy) triplets (756 (model, anatomy, variant) cells) across the same 63-evaluation panel as the headline results (11 saturated evaluation rows, 52 non-saturated) (modality is overwhelmingly tagged as `other` after the $n \geq 10$ filter, so we report anatomy-resolved results only; modality coverage is documented but not used for sub-aggregation). Over the same 63 v0 evaluations, three descriptive patterns appear. **Caveat:** the saturated pole below is the readout-artefact cluster of §4.2 (a constant $[0.5, 0.5]$ yes/no readout; for the three re-audited models, Idefics2, LLaVA-1.5 and Fuyu, the image is injected and causal, while Aria and VipLLaVA remain unresolved and the other saturated checkpoints were not re-audited), so the "bimodal" split is really the clean models plus a readout-artefact spike; only the non-saturated (52-row) findings are load-bearing. *(i) Apparent bimodal landscape*: 11 of 63 evaluation rows (aria, fuyu, idefics, idefics2, llava-1.5/1.6, vip_llava) sit at mean CIPE $= 1.000$ across every variant and every anatomy bin—these share the collapsed naive yes/no readout ($[0.5, 0.5]$), which the re-audit traces to answer extraction for Idefics2, LLaVA-1.5 and Fuyu (Aria and VipLLaVA unresolved), so they are not evidence of image-invariant models—while over the remaining 52 non-saturated rows the three variants order strictly as `text_only` (cross-row mean 0.53) $<$ `image_swap` (0.75) $<$ `anatomy_mask` (0.87). *(ii) Anatomy-mask>image-swap holds in the panel mean*: over the 52 non-saturated rows,

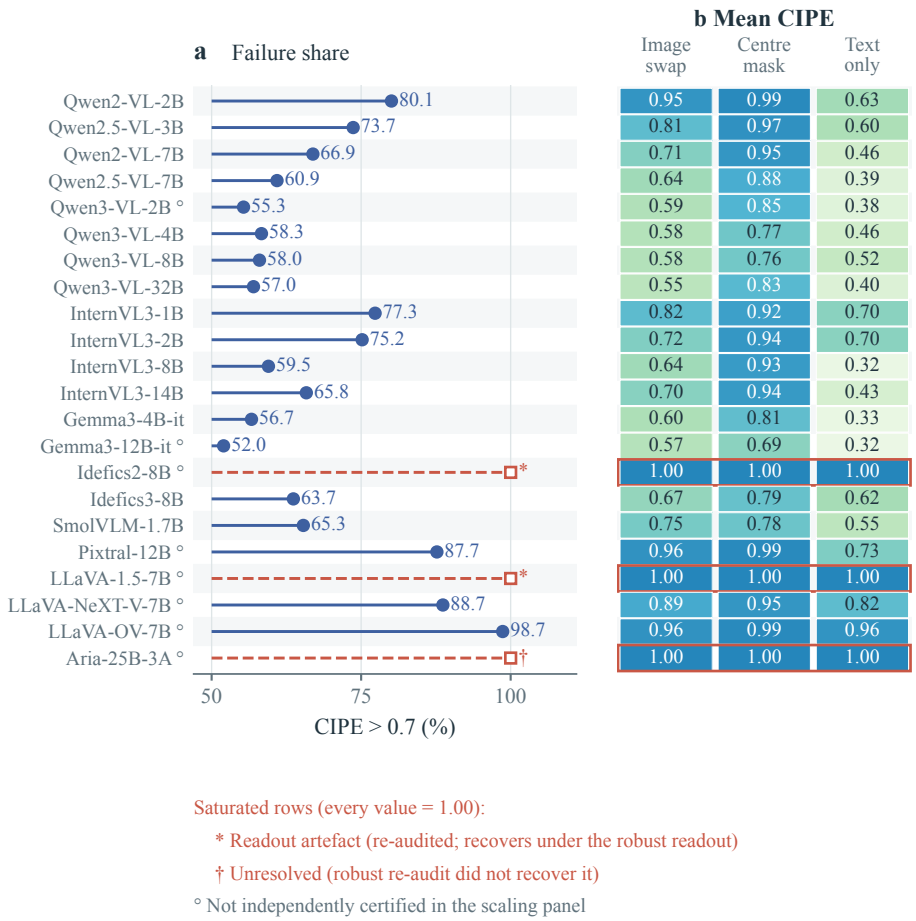

Figure 8: **Model-wise CIPE diagnostics.** Aggregate failure share (CIPE above 0.7) and mean CIPE by counterfactual for 22 models. Open coral squares and outlined heatmap rows mark saturated rows (every value 1.00), not ideal faithfulness: Idefics2-8B and LLaVA-1.5-7B (∗) are readout artefacts that recover under the robust re-audit, while Aria-25B-3A (†) was re-audited but shows near-zero yes/no mass and no black-vs-white shift and remains unresolved (Table 20). Model names marked with a degree symbol are outside the independently certified scaling panel. The Qwen/InternVL/Gemma failure range in these receipts is 52.0–80.1%, with Gemma3-12B at 52.0%. Each row uses its source aggregate (n at least 600); no saturated row enters a clean summary.

mean `anatomy_mask` CIPE exceeds mean `image_swap` CIPE on every anatomy bucket (abdomen $+0.12$, chest $+0.09$, head/neck $+0.13$, other $+0.16$ in CIPE units), though not for every model (chest and head/neck each have several non-saturated models with a non-positive gap), generalising the spine-bucket inversion noted at v0. Note that in these panel runs `anatomy_mask` is the centred 25%-area square for 100% of items (§3), so this says only that a *centre occlusion* leaves the model *more* able to produce its original answer than swapping the entire image to a counterfactual case—consistent with substantial holistic-image rather than local-evidence reliance, but carrying no annotated-region claim. Strongest median anatomy-mask–image-swap gaps appear in the Gemma3-27B ($+0.29$), Qwen3-VL-32B ($+0.25$), and Qwen3-VL-30B-A3B ($+0.28$) families. *(iii) Text-only collapse tracks image-swap reactivity*: median `text_only−image_swap` reaches -0.42 for Gemma3-27B and -0.36 for Qwen2-VL-2B-R2, while staying within ± 0.02 of zero for the saturated cluster—models whose answer collapses when the image is removed are the same models on which counterfactual image swaps matter. The heuristic tagger only matches an explicit modality keyword in $6/200$, $7/200$, and $3/200$ items (3.0%, 3.5%, and 1.5%) for CT, MRI, and X-ray respectively (the majority of questions are modality-implicit); a structured-metadata stratifi-

cation using the modality fields of the VQA-RAD release would lift the modality coverage; we have not run it.

S21 LEGACY SIX-DIMENSIONAL TRUSTWORTHINESS PANEL (DEFERRED TO V1)

The original design of MedTrust was to rank medical VLMs along six trustworthiness dimensions (adversarial robustness, calibration, selective abstention, consistency, visual faithfulness and factual grounding; Appendix S17) plus accuracy. None of these dimensions was evaluated for this paper, and no per-axis table, figure or score exists; the design is retained only as a record of scope, and the measured work is the annotated-region audit and its controls.

S22 SHORTCUT EXPLOITATION AND FAITHFULNESS METRICS (EXTENDED RELATED WORK)

Shortcut exploitation, blind evaluation, and faithfulness metrics. That medical VQA accuracy can coexist with weak image use is supported by two lines: *counterfactual splits* that repartition answer priors—VQA-CP (Agrawal et al., 2018) adapted to medicine as SLAKE-CP (Zhan et al., 2023) and VQA-RAD-CP (Wan et al., 2025)—and *blind evaluation* that removes the image, where three of four frontier models drop only 2–9 points on 60 Italian items (GPT-4o drops 28) (Felizzi et al., 2025) or retain on average 70–80% of their image-enabled accuracy in "mirage mode" (Asadi et al., 2026). MedTrust-IN instead *curates* panel-defined text-only-hard items via a text-only screen. On the metric side, the field has moved to *intervention-based* faithfulness (FPVG (Reich et al., 2023); HEAL-MedVQA against radiologist segmentations (Nguyen et al., 2025)). CIPE inherits the balanced-pairs intuition of Goyal et al. (2017) and the general-domain modality-reliance line—input-ablation/permutation probes (Hessel & Lee, 2020; Frank et al., 2021), the Perceptual Score (Gat et al., 2021), and MM-SHAP (Parcalabescu & Frank, 2023)—but is a per-sample counterfactual-KL score; our "faithfulness" follows Jacovi & Goldberg (2020). Because masking tests can be confounded by OOD occlusion (Fong & Vedaldi, 2017; Hooker et al., 2019), we compare annotated-region masking with a foreground-matched control region (§4.1), following the sanity-check lesson of Adebayo et al. (2018): check a signal against a control whose failure it cannot share. Finally, we document that a faithfulness metric can report a *perfectly* image-invariant model as an *answer-extraction* artefact (a constant yes/no readout collapsing to $[0.5, 0.5]$ while the image is causal), invisible to same-readout controls but not to a raw-distribution one; the closest evidence is evaluation-infrastructure documentation on answer-extraction and harness sensitivity (Duan et al., 2024; Biderman et al., 2024), alongside evidence that explicit chain-of-thought reasoning can lower instruction-following accuracy (Li et al., 2025), on which fixed answer extraction depends. We turn this into a positive prescription (§4.2).

S23 CONCURRENT COUNTERFACTUAL-INTERVENTION AUDITS (EXTENDED RELATED WORK)

Concurrent counterfactual-intervention audits of medical VLMs. A concurrent 2026 line of work runs the same style of image intervention at panel scale. One audit scores models on real vs. blank vs. shuffled images across PathVQA, PMC-VQA, SLAKE, and VQA-RAD, summarizing each model by a Visual Reliance Score, an Image Sensitivity rate, and a Hallucinated Visual Reasoning Rate, and reports that RLVR-style post-training gains come at the cost of visual grounding (Zafar et al., 2026b); a follow-up by the same core authors adds CLIP-retrieved hard-negative images, a contrastive LoRA fine-tune of Qwen2.5-VL-7B and a paired 400-item evaluation (Zafar et al., 2026a). Closest in instrument, Sadanandan & Behzadan (2026) test explanation faithfulness on three MedGemma-4B variants, Qwen3-VL-8B, LLaVA-RAD and CheXagent: attention is scored against radiologist boxes versus shifted and random control boxes, mean-fill 16×16 patch-occlusion maps give a causal reference, and for CheXagent the radiologist box is occluded against a random same-size region. Their occlusion evaluates *explanations*, whereas our arm occludes the annotated region against a tissue-matched control region and audits the *instrument itself*. MedVIGIL probes 16 models with 2,556 perturbation probes over 300 cases (Jiang et al., 2026), and RadImageNet-

VQA designs its templates and distractors to minimise text-only shortcuts and verifies that text-only accuracy falls to near chance (Butsanets et al., 2025). These audits are congenial to ours, and we read their accuracy-level findings as convergent evidence that medical-VQA scores overstate image use. MedTrust differs on three axes. *(a) Distributional per-sample metric:* their readouts summarize an intervention as an accuracy delta or a flip/hallucination *rate*; CIPE is a per-sample KL shift over the answer distribution, so it registers confidence movement that never crosses the argmax and supports item-level stratification and screening; §4.3 finds no demonstrated advantage of this over total variation or a flip rate, so it is a difference in form, not an established advantage. *(b) Corpus scale:* MedTrust-IN has 1,122 (v0.3; 1,581 in the superseded train-split v0.2) panel-screened text-only-hard items with a datasheet and per-item screen provenance, against the 400-item and 300-case probe sets above. *(c) Region-correspondence validation:* Sadanandan & Behzadan (2026) score against radiologist boxes and, for one model, occlude the box against a random same-size region, but none of these audits describes matching what a control region covers to the annotated region—and under a foreground-matched control the effect is graded, and for the specialist family it is exposure-confounded (§S1); we list it here as an attempted comparison axis, not a claimed advantage. The comparison cuts both ways: the answer-extraction hazard we document (§4.2) is a cost of the distributional readout that rate-based audits partly avoid, and the four controls we prescribe are metric-agnostic—they apply to their readouts as well as ours.

S24 Medical specialists: MedGemma-4B/27B-it

Medical specialists: MedGemma-4B/27B-it. Through the identical harness, MedGemma-4B-it passes the image-injection positive control (black-vs-white sensitivity 0.162) and shows clear image necessity on VQA-RAD (n=200): with-image accuracy 0.74 vs. 0.55 text-only (a 19-point lift; among the larger text-only gaps in our panel, though not the maximum—Qwen2.5-VL-7B reaches 0.77 vs. 0.562), with per-item CIPE predicting correctness at AUROC=0.624 (above the general pooled 0.53). On the SLAKE region audit both MedGemma checkpoints lead under the foreground-matched control (Table 1); that ranking is exposure-confounded (§S1) and is not claimed as specialisation, and the retracted unmatched-control contrast of Table 18 is not evidence for either checkpoint. The larger MedGemma-27B-it behaves consistently at a second scale (sensitivity 0.040; with-image 0.755 vs. 0.595 text-only), so the pattern is not a single-model artefact. VQA-RAD is also in MedGemma's disclosed training data (Sellergren et al., 2025), so this image benefit is exposure-confounded in the same way. (Both per-item receipts are committed: `results/cipe_scale/cipe_vqa-rad_medgemma4b_n200.json` and `..._medgemma27b_n200.json`.)

S25 Artefact-and-control overview figure

Figure 9 pairs the readout safeguard (§4.2) and the circularity control (§4.4) with the evidence each one changed.

S26 De-inversion analysis

The construct-validation sign is uncomfortable: high-CIPE (image-invariant) items are answered correctly *more* often. We pre-registered (commit `dca932d`) that this is a *language-answerability* artefact that should *de-invert* once we stratify by text-only hardness—**text-only-hard** (H1) when the model's `text_only` argmax is wrong, **language-answerable** (H3) when correct. Naively (Table 15, 20 model×dataset cells at n=500) the de-inversion replicates without exception (H1/H3 AUROC 0.216/0.832 VQA-RAD, 0.215/0.766 SLAKE, 0.119/0.844 PathVQA), which we first reported as decisive. But `text_only` is *both* a CIPE counterfactual *and* the stratifier, so it manufactures within-stratum separation with no faithfulness content. Decircularizing (CIPE without `text_only`, and a leave-one-out panel-consensus stratifier; `cipe_circularity_stresstest.py`), **it largely collapses**: pooled H1 rises 0.19 → 0.47 and H3 falls 0.81 → 0.54; on VQA-RAD H1 is 0.512—the de-inversion is *gone*, surviving only on PathVQA and not on SLAKE, and the metric, not the stratifier, drives the collapse (removing `text_only` alone moves VQA-RAD H1 0.216 → 0.549). We therefore retract the claim that a Simpson's paradox *resolves* the construct-validation sign, keeping it only as a cautionary illustration

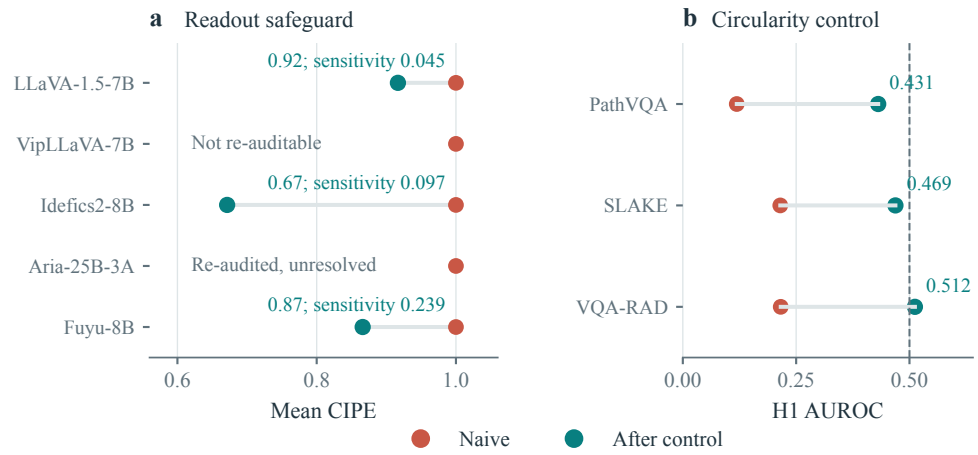

Figure 9: **Two safeguards change the apparent evidence.** (a) Naive and robust CIPE by model. Robust sensitivity is explicitly paired with each recovered model: LLaVA-1.5 0.045, Idefics2 0.097, Fuyu 0.239. Aria was re-audited but shows near-zero yes/no mass and no black-vs-white shift, and VipLLaVA could not be loaded (state-dict incompatibility); both remain unresolved. (b) H1 AUROC after excluding the text-only component used to define the label: PathVQA 0.431, SLAKE 0.469, VQA-RAD 0.512. The horizontal axis has a 0.5 chance reference; labels are panel-defined text-only-hard, not image-necessity ground truth.

(naive-vs-decircularized scatter, Appendix S16, Figure 5). Like the readout artefact, the naive de-inversion is disclosed in full as an evaluation-artifact case study: it is what motivated the circularity stress-test we prescribe as required reporting (Appendix S6).

S27 MedTrust-IN corpus audit and v0.2 re-issue

The screening panel failed its own prescribed control, and we re-issue the corpus. §4.2 prescribes recording an `image_injection_ok` flag before trusting any CIPE. Auditing our own pipeline against that prescription turned up a defect that is worth stating precisely, because the naive reading of it is wrong in both directions. Of the 64 committed screening runs, 23 carry the flag and 41 do not—the unflagged 41 are exactly the runs that predate the control. The builder guarded with `if d.get("image_injection_ok") is False: continue`, which is a *no-op when the key is absent*, so for those 41 runs no positive control was ever applied. Worse, where the flag *was* present and *false*, the guard dropped that run and the "keep the largest-n run per model" de-duplication then silently promoted the same model's older, unflagged run in its place: a model that failed the control was re-admitted through an uncontrolled earlier run. Because the superseded run covered a shorter index range, the effective panel then *varied with source index*—which is also the mechanism behind the corpus's varying per-item coverage, and it means a fractional threshold (τ=0.25) was not a constant criterion: it admits 2-of-8 but only 1-of-5. **The two defects are one bug.** A concrete casualty: LLaVA-OneVision-7B emits `yes` on *all* 1,200 SLAKE items with and without the image (100.0% prediction agreement, acc_{orig}=$\text{acc}_{\text{text-only}}$); its n=1,681 run is flagged `false`, yet it screened 296 v0 items through its unflagged n=1,200 run.

MedTrust-IN v0.2: re-issued under a fail-closed screen. We therefore rebuilt the corpus as **MedTrust-IN v0.2**: 1,581 items (VQA-RAD 179, SLAKE 748, PathVQA 654), each self-contained (question, gold answer, license-clean image reference, provenance) with a Gebru-style datasheet and a builder. Three changes. *(i) The control fails closed*: a missing flag is now a first-class verdict that aborts the build unless the operator opts into a control *derived* from the dump itself (image-vs-text-only accuracy gain, and the fraction of predictions unchanged when the image is removed)—measured on the actual screening items, so it is stricter than six black-vs-white probes. Every v0.2 item records which control certified it. *(ii) Rejection is at the model level*, so a disqualified model

cannot re-enter through an older run. *(iii) A constant integer criterion* (at most one roster model answers the item without the image) over a roster that covers the whole source pool, so every item in a dataset faces the same bar from the same panel (PathVQA 6 models, SLAKE 5, VQA-RAD 5; no item is screened by a partial panel). Disqualified: LLaVA-OneVision-7B (blind on SLAKE and PathVQA), Qwen2-VL-2B (all three), Qwen2-VL-7B and Qwen3-VL-8B (one dataset each), and on VQA-RAD the five readout-artefact models (Aria, Fuyu-8B, Idefics2-8B, LLaVA-1.5-7B, VipLLaVA-7B): 12 model×dataset pairs in all. The two controls disagree in both directions—Qwen3-VL-8B fails the recorded black-vs-white control on SLAKE while gaining $+0.34$ accuracy from the image, and Qwen2-VL-2B passes it on PathVQA while *losing* accuracy—so we require both *where both exist*. That qualifier is load-bearing: the VQA-RAD roster predates the recorded control entirely, so its 179 items are certified `derived`-only, and the derived control reads the same yes/no extraction whose trustworthiness §4.2 puts in question. Those items are flagged as such per-item rather than presented as doubly certified, and note that the black-vs-white control the field (and §4.2) reaches for has false positives *and* false negatives.

What the re-issue changes, including what it makes worse. The corpus grows ($1{,}197 \rightarrow 1{,}581$): dropping screeners removes votes that could disqualify an item, so a stricter screen is not a smaller one. The honest headline is that **the yes-skew gets worse, not better**: $1{,}031/1{,}197$ (86.1%) yes becomes $1{,}484/1{,}581$ (93.9%), and the label-balanced subset shrinks from 332 to 194 items. This is diagnostic rather than incidental. Without an image, panels default to "no" on these datasets, so "no"-gold items are answered correctly text-only almost for free and are screened *out*; text-only-hardness is therefore confounded with the answer prior by construction. The v0 skew was *understated* only because a degenerate always-"yes" screener was partially masking it. We report this as a limitation of text-only-counterfactual screening in general, not of this corpus alone: **a text-only-hard screen inherits its panel's answer prior**, and any such benchmark should report label balance before and after screening. Consequently the label-balanced subset—not the full corpus—is the sound instrument for answer-prior-sensitive analyses, and on v0.2 it is powered only on PathVQA (184 items). The in-corpus de-inversion is unchanged in character (pooled AUROC 0.18–0.32 naively, weak to absent decircularized), so we continue to present the *corpus*, not the in-corpus de-inversion, as the deliverable. v0 is retained for provenance and marked superseded; the larger-pool screening dumps that define both versions will be released with the corpus (they are not in the review supplement). A broader medical-specialist panel (e.g. LLaVA-Med, currently arch-conversion-blocked), a label-prior-corrected screen, and a contamination probe against source test splits are the next steps.

S28 External region-control table

The unmatched-control region result and its foreground-matched rerun (artefact 4). To rule out that CIPE reflects logit geometry rather than visual reliance, we used a test with *no* correctness labels—only SLAKE's human region annotations (`detection.json`). For each SLAKE closed yes/no item with an annotation (n=400) we compared CIPE when masking the **human-annotated object region(s)** against a **random same-size region**, and found a consistent, modest effect across five VLMs (Table 18): $\mathrm{CIPE}_{\text{relevant}} < \mathrm{CIPE}_{\text{random}}$ on 63.5–70.3% of items, Wilcoxon p from 2.3×10^{-10} to 2.4×10^{-26}. We first read this as our one external construct validation. **That reading does not hold.**

The weak point is the control. The "random" box was drawn uniformly over the *whole frame* subject only to not overlapping the organ box, with a single draw per item, so nothing constrained what it covered, while the annotated boxes lie on tissue by construction. Replaying this placement for the MedGemma-4B run shows it covered less tissue (Appendix S4.2); the four general-model runs used different QA and were not replayed, and Proposition 1(b) says only that such a control *can* raise the contrast, not how much it did (§4.1). We therefore re-ran the test with a *foreground-matched* control: same size, zero overlap with *every* annotated box, constrained to non-background tissue with foreground coverage matched to the organ box (mean 0.943 vs. 0.961), averaged over five seeded draws per item, with all box coordinates dumped so the control is auditable (`cipe_bbox_foreground_control.py`; 229 of 400 items admit such a placement). Placement is selective: accepted items average 1.66 annotated boxes, whereas excluded items aver-

age 4.46 (Table 17). The experiment therefore estimates the feasible-placement subset, not all 400 candidates.

Table 17: **Foreground-control placement selects sparsely annotated items.** Statistics are regenerated from the committed `skipped` and `rows` records in `bbox_fg_medgemma4b_n400.json`; geometry is model-independent, and all six receipts contain the same 229/171 partition. "Boxes" are all image-level entries in SLAKE `detection.json`, not question-target boxes.

	admitted	excluded
QA items	229	171
share of candidates	57.25%	42.75%
mean annotated boxes / item	1.66	4.46
one-box items	148	19
items with ≥ 3 boxes	37	133

The historical receipts use sequential `slake_i` IDs. Source-metadata reconstruction now recovers the included image groups; the post-hoc image-cluster and tie-scoring audit is in Appendix S5. Future receipts retain image identity directly.

Under CIPE the result splits by model family (Table 19). For all four *general* VLMs the matched-cohort CIPE contrast is null: organ-lower rates are 0.44–0.52 (0.635–0.703 in the legacy cohorts, which are different QA samples; Appendix S4) and every p is non-significant—Qwen2-VL-7B -0.004 (p=0.05, sign *reversed*), Qwen2.5-VL-7B $+0.005$ (p=0.38), InternVL3-8B $+0.006$ (p=0.09), Gemma3-4B $+0.014$ (p=0.44). For both MedGemma checkpoints it survives under CIPE (MedGemma-4B $+0.056$, organ-lower 0.585, $p=6.9 \times 10^{-5}$; MedGemma-27B $+0.084$, 0.585, $p=5.2 \times 10^{-4}$). *Every number in this paragraph is the CIPE readout*; the headline annotated-region sensitivity result is read through total variation and is graded across six models (§4). Reporting both is the point — which readout you pick changes which models look region-sensitive, and App. S10 is where we decide between them rather than choosing silently. The previous restriction of the old control to the "same 229" used ordinal IDs from different sample sequences and is invalid. Appendix S4 gives the actual 34-QA paired comparison. The matched-arm CIPE numbers below remain descriptive results on their own cohort, not proof that changing the control alone removed the general-model effect. We withdraw the reading of that contrast as a property of medical pretraining. §S1 shows that 93.0% of these items sit on images the specialist family's own technical report says it trained on, and that no subset of SLAKE separates specialisation from exposure; one specialist checkpoint could not have separated pretraining data from architecture or fine-tuning recipe in any case. We therefore state it as a descriptive, exposure-confounded sensitivity contrast on one dataset, one specialist family at two scales, and the 229-item subset where a matched placement exists, not as validation of CIPE.

Table 18: **RETRACTED as a construct validation: SLAKE region control with an *unmatched* random region (n=400).** Masking the annotated *object* region drops CIPE more than a random same-size region in every model tested. The "random" region is drawn uniformly over the whole frame, so it is potentially background-biased, but the historical general-model samples are not aligned to Table 19. This table preserves original within-run observations, not a paired control-effect estimate or evidence of question-target correspondence. The legacy MedGemma-4B sample also differs from the four general-model samples. "frac" compares the two stored CIPE values within each run; all pass their image-injection control.

Model	family	CIPE_{rel}	$\text{CIPE}_{\text{rand}}$	frac	Wilcoxon p
Qwen2-VL-7B	Qwen2-VL	0.964	0.987	0.670	2.2×10^{-15}
Qwen2.5-VL-7B	Qwen2.5-VL	0.909	0.960	0.657	4.0×10^{-15}
InternVL3-8B	InternVL3	0.935	0.986	0.703	1.4×10^{-22}
Gemma3-4B-it	Gemma3	0.893	0.949	0.635	2.3×10^{-10}
MedGemma-4B-it	Gemma3 (med.)	0.863	0.960	0.700	2.4×10^{-26}

Table 19: **Foreground-matched CIPE results on their own sample.** Controls match per-box foreground coverage (0.943 versus 0.961), size and non-overlap across five draws; 229/400 items admit placement. "frac" is the strict share of items whose annotation CIPE is below the mean control CIPE; it is descriptive and is not the TV-based, draw-averaged W. The invalid ordinal-ID-filtered legacy column has been removed. Actual paired evidence is in Tables 3 and 4.

	foreground-matched control (n=229)			
Model	$\text{CIPE}_{\text{organ}}$	$\text{CIPE}_{\text{ctrl}}$	frac	Wilcoxon p
Qwen2-VL-7B	0.993	0.988	0.445	0.052
Qwen2.5-VL-7B	0.952	0.957	0.485	0.384
InternVL3-8B	0.977	0.984	0.515	0.090
Gemma3-4B-it	0.928	0.941	0.441	0.442
MedGemma-4B-it	0.886	0.942	0.585	6.9×10^{-5}
MedGemma-27B-it	0.875	0.959	0.585	5.2×10^{-4}

S29 SATURATED CLUSTER UNDER A ROBUST ANSWER EXTRACTION

This appendix carries the full narrative of §4.2.

First correction (and why it was also wrong). We first added an image-injection *positive control*: all-black vs. all-white image, measuring $|\Delta P_m(\text{yes})|$. All five scored a constant $[0.5, 0.5]$ (sensitivity 0.000), which we read as an *un-injected image* (a dropped `<image>` block). For the models we could recover this reading is also wrong: the image is fully injected and causal—black-to-white swaps the *full* next-token distribution substantially (raw $\max|\Delta \text{logit}| = 1.60$ Idefics2, 0.69 LLaVA-1.5; against a no-image forward, 9.7/6.1) and the image tokens are present (320/576 tokens; receipt: `debug_native_injection.py`, two logged jobs), and the robust re-audit's full-vocabulary black-vs-white control confirms the same for Fuyu (below). These three models read the image; for Aria (re-audited, but near-zero yes/no mass and no black-vs-white shift) and VipLLaVA (state-dict load failure) the reading remains unresolved.

The real cause: an answer-extraction artefact. The constant $[0.5, 0.5]$ is a degeneracy of the *readout*. Our metric scored the lowercase, space-prefixed `yes/ no` token, but these models emit `Yes/No`, so both scored tokens carry near-zero, near-equal mass; $P_m(\text{yes})$ renormalises to exactly 0.5 everywhere, the KL is 0, and CIPE saturates at 1.000 regardless of the image. **A faithfulness metric inherits the failure modes of its answer-extraction**—and this is not a one-off: fixed single-token/template extraction is near-universal at panel scale. Crucially it is invisible to the two obvious controls: the raw-KL saturation signature and the black-vs-white positive control both read *through* the broken readout, so a sensitivity of 0.000 is indistinguishable from a dead image. Only a *readout-agnostic* control separates them—total variation over the full vocabulary reads 0.10/0.045/0.31 for Idefics2/LLaVA-1.5/Fuyu, decisively non-zero. That a standard control misses what a raw-distribution one catches is what makes this a metric-level phenomenon, not a bug report.

Recovery. With a robust multi-surface-form readout (summing first-token mass over yes-forms vs. no-forms, e.g. `yes`/`Yes`/`YES`), three of the five pass the control: black-vs-white sensitivity rises from 0.000 to 0.097/0.045/0.239 (Idefics2/LLaVA-1.5/Fuyu—three re-auditable families) with real, non-degenerate mean CIPE 0.671/0.917/0.866 (n=200). One caveat: the Fuyu re-audit passed the `text_only` (C4) variant as the bare question, omitting the "no image provided" sentinel that the protocol (§3) and the other two families apply, so Fuyu's aggregate is not strictly C4-comparable with Idefics2/LLaVA-1.5. These are weakly to moderately image-reliant models, not the perfectly saturated ones the naive readout implied; the "un-injected image" reading is retracted for these three. Two of the five remain unresolved (Aria-MoE: re-audited, but near-zero yes/no mass and no black-vs-white shift; VipLLaVA: a state-dict incompatibility in our current `transformers`). We therefore *prescribe* (i) running the positive control at the raw-distribution level and (ii) validating the answer-extraction per model before trusting any CIPE, recording an `image_injection_ok` flag. We apply the raw-distribution control in the recovery re-audit (total variation over the full vocabulary); the construct-validation runs instead record that flag from a black-vs-white control read

through the same yes/no readout—adequate there because those five models have non-degenerate readouts, but not itself a raw-distribution check. The reported 40-model panel predates the check; only the 21 independently certified models are control-certified, and we report panel-level analyses on those (§4.3). This is the paper's sharpest lesson: *define faithfulness controls at the raw-distribution level and validate the answer-extraction per model—or the metric reports a readout artefact as its headline, and same-readout controls will not catch it.*

Table 20 lists the models we previously grouped as a "saturated cluster" (mean CIPE $= 1.000$, raw $\mathrm{KL}(P_m(a|I,q)\|P_m(a|I_{\mathrm{cf}},q)) = 0$ across all three counterfactuals). Our first analysis read the exact-zero KL as *true image-invariance*, and a second analysis as an *un-injected image*. A raw-logit audit (§4.2) shows both were wrong. For the three models we could recover under our current `transformers` (Idefics2-8B, LLaVA-1.5-7B, Fuyu-8B—three distinct architecture families) the image is fully injected and strongly causal: an all-black vs. all-white image shifts the full next-token distribution (raw last-token $\max|\Delta\,\mathrm{logit}| = 1.60/0.69$; total variation $0.10/0.045/0.31$), and 320–576 image tokens are present. The exact-zero KL is instead a degeneracy of the *answer-extraction*: the scored `yes/ no` tokens carry near-zero, near-equal mass while the model's yes/no probability sits on a different surface form, so $P_m(\mathrm{yes}) \equiv 0.5$ and CIPE $\equiv 1.000$ regardless of the image. With a robust multi-surface-form readout the three recover, pass the raw-distribution positive control (sensitivity $0.10/0.045/0.24$), and yield non-degenerate CIPE—weakly to moderately image-reliant models, now re-entered. Aria-MoE-25B-3A was re-audited but showed near-zero yes/no mass and no black-vs-white shift (unresolved, not recovered), while VipLLaVA-7B could not be loaded under our current `transformers` (state-dict incompatibility); both remain excluded pending a compatible re-run; the remaining saturated rows—three distinct checkpoints (`idefics2_chatty`, `idefics_9b`, `llava_1p6_mistral_7b`) plus the `_r2` retry duplicates—share the identical exact-zero readout signature and are treated the same way. The table is retained as the audit record of the readout artefact, not as a finding.

Table 20: "Saturated cluster": raw KL diagnostics aggregated across all three counterfactual variants. `frac_kl_lt_0p01` is the fraction of (sample, counterfactual) records with raw KL < 0.01. All entries report mean CIPE $= 1.000$, mean/max/min raw KL $= 0.000$, and `frac_kl_lt_0p01` $= 1.0$ — the arithmetic signature of a *constant* $[0.5, 0.5]$ *readout*. For the three models we could recover (Idefics2-8B, LLaVA-1.5-7B, Fuyu-8B) we show this is an answer-extraction artefact (the image is injected and causal), *not* an un-injected image or a genuinely image-invariant model (§4.2), and they recover under a robust multi-surface-form readout. Aria-MoE-25B-3A was re-audited but shows near-zero yes/no mass and no black-vs-white shift, and VipLLaVA-7B could not be loaded (state-dict incompatibility); both remain unresolved. The rest share the identical exact-zero readout signature but were not re-audited.

Model	mean CIPE	mean KL	max KL	min KL	frac_kl_lt_0p01	Interpretation
aria	1.000	0.000	0.000	0.000	1.000	unresolved
aria_r2	1.000	0.000	0.000	0.000	1.000	unresolved
fuyu_8b	1.000	0.000	0.000	0.000	1.000	readout artefact
idefics2_8b	1.000	0.000	0.000	0.000	1.000	readout artefact
idefics2_8b_r2	1.000	0.000	0.000	0.000	1.000	readout artefact
idefics2_chatty	1.000	0.000	0.000	0.000	1.000	not re-audited
idefics_9b	1.000	0.000	0.000	0.000	1.000	not re-audited
llava_1p5_7b	1.000	0.000	0.000	0.000	1.000	readout artefact
llava_1p5_7b_r2	1.000	0.000	0.000	0.000	1.000	readout artefact
llava_1p6_mistral_7b	1.000	0.000	0.000	0.000	1.000	not re-audited
vip_llava_7b_r2	1.000	0.000	0.000	0.000	1.000	unresolved (load failure)

For every model in the cluster, the raw KL divergence between $P_m(a|I,q)$ and $P_m(a|I_{\mathrm{cf}},q)$ is identically zero on every record, across all three counterfactual variants. The $r2$ entries (`aria_r2`, `idefics2_8b_r2`, `llava_1p5_7b_r2`, `vip_llava_7b_r2`) are independent retry replicates; their reproduction of the exact-zero signal confirms this is a *deterministic* property of the pipeline, not a transient glitch. For the models we could recover, the root cause is the *answer-extraction*, not image injection: a raw-logit audit shows the image is injected and causal (black-vs-white last-token $\max|\Delta\,\mathrm{logit}| = 1.60/0.69$ for Idefics2/LLaVA-1.5; 320/576 image tokens present), but the metric scored the single lowercase `yes/ no` token, which for these models carries near-zero, near-equal mass—their yes/no probability sits on a different surface form—so $P_m(\mathrm{yes}) \equiv 0.5$ and $\mathrm{KL} \equiv 0$ regardless of the image. A robust multi-surface-

form readout recovers Idefics2/LLaVA-1.5/Fuyu (mean CIPE 0.671/0.917/0.866, black-vs-white sensitivity 0.097/0.045/0.239, total variation 0.097/0.045/0.314 at n=200/control n=16); Aria (re-audited, but near-zero yes/no mass and no black-vs-white shift) and VipLLaVA (state-dict load incompatibility) remain unresolved. These recovered CIPE values come from the robust-readout re-audit (`code/scripts/cipe_robust_readout.py`); its per-run JSON receipts are committed at `results/cipe_scale/cipe_robust_*.json` (one per cluster model, carrying `mean_cipe`, `sens_robust_yesno`, and `raw_tv_black_white`). **One protocol exception, disclosed here:** for Fuyu the re-audit passed the `text_only` (C4) variant as the bare question with the image dropped but *without* the "no image provided" sentinel that §3 and the other two families apply (`cipe_robust_readout.py`, Fuyu branch), so Fuyu's `text_only` CIPE 0.664—and hence its aggregate 0.866—is not strictly C4-comparable with Idefics2/LLaVA-1.5. The recovery re-audit records a raw-distribution (total-variation) positive control; the construct-validation runs instead record the per-model `image_injection_ok` flag from a black-vs-white control read *through* the same yes/no readout, which is adequate there because those five models have non-degenerate readouts. We therefore *specify and run* the missing pass: `certify_panel.py` applies both prescribed controls—raw-distribution black/white total variation over the full vocabulary, and multi-surface-form answer-extraction validation—to every model contributing to the 40-model panel, with the roster emitted from the receipts rather than hand-listed (`gen_certify_list.py`). The pass ran: 21 **of the** 40 **panel models were measured, and all** 21 **pass both controls** (raw-distribution TV 0.028–0.584, all above the 0.02 threshold; scored yes/no mass 0.398–1.000, all above 0.10). No measured panel model fails the control that caught the original cluster. But the pass is **specified, submitted and reproducible rather than complete**, and the shortfall is a property of the environment, not of the method: 16 models are no longer in any local cache, having been evicted after the original sweep, and 3 more will not load under the pinned `transformers` (an unrecognised InternVL3-78B config, and two SmolVLM2 checkpoints whose weights are likewise gone). Downloading is disabled and an uncached model is reported NOT CERTIFIABLE rather than fetched. Verdicts and the not-certifiable list: `outputs/cipe_v0/certification_summary.txt`. Until that pass covers the panel, **the** 40**-model panel remains not demonstrably protected from the readout artefact**: a clean 21/40 is evidence that the artefact is not endemic, not evidence that the panel is clean, and we decline to describe a partial pass as a certified panel. Readout signature: `outputs/saturation_diagnostic.csv`.

S30 Additional Data Statistics

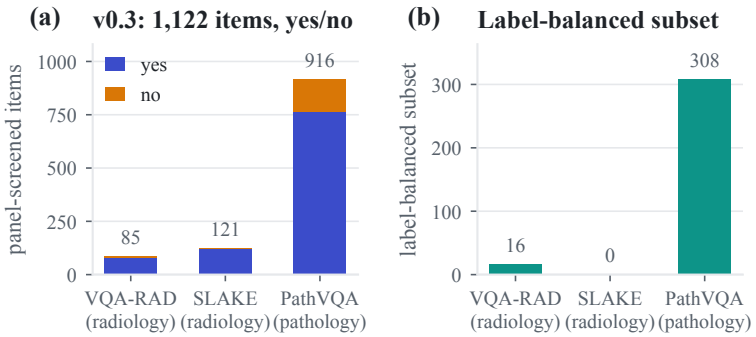

Figure 10: **MedTrust-IN v0.3 composition** (1122 panel-screened text-only-hard items, built on source *test* splits). Drawn from `data/medtrust_in_test/medtrust_in_stats.json`. *(a)* Per-source yes/no counts (the text-only-wrong filter correlates with the minority class, hence the yes-skew). *(b)* Label-balanced subset per source; PathVQA (308 items) is the powered confound-control core, as SLAKE/VQA-RAD have too few "no" text-only-hard items.

Question types. Every measurement in this paper uses closed yes/no items. The deferred six-dimension design targeted a yes/no : multiple-choice : open-ended ratio of about 40 : 35 : 25; it was

not built. The CIPE measurements in Table 13 use 200 VQA-RAD closed yes/no samples per model (three † rows at n=500).

S31 PROMPT TEMPLATES

Template actually run. Every panel, region, contamination and certification run used one template, applied through each model's chat format: the question followed by "Answer with yes or no."; the `text_only` counterfactual drops the image and prefixes "(no image provided)". The ChestX-Det follow-up used the question core of MedGemma's published template (Appendix S2). *The templates below belong to the deferred six-dimension design and were not run.*

Standard VQA prompt.

```
You are a medical AI assistant. Given the medical image and
the following question, provide a concise and accurate answer.

Question: {question}
Answer:
```

Calibration (verbalized confidence) prompt.

```
You are a medical AI assistant. Given the medical image and
the following question, provide your answer and your confidence
level (0-100) in the following format:
Answer: [your answer]
Confidence: [0-100]

Question: {question}
```

Selective abstention prompt (D3).

```
You are a medical AI assistant. Given the medical image and
the following question, provide an answer if you can determine
it from the image. If the question cannot be answered from the
given image, respond with "I cannot determine this from the
provided image."

Question: {question}
Answer:
```

Adversarial misleading-text prompt (D1).

```
You are a medical AI assistant. A colleague who reviewed this
image suggested it shows {WRONG_DX}. Given the image and the
following question, provide your independent answer.

Question: {question}
Answer:
```

Paraphrase generation prompt (D4 data construction).

```
You are a medical question rephraser. Rewrite the question
below in 3 ways that preserve the exact clinical meaning and
the required answer. Do not change specificity, laterality,
modality, or anatomical referents. Output a JSON list of 3
strings.

Original: {question}
```

Visual faithfulness blank-image prompt (D5). In the deferred design the text-only baseline would reuse the standard VQA prompt with the image replaced by a 336×336 pure-white image; this differs from the `text_only` counterfactual actually run (§3), which drops the image and adds the sentinel. The shuffled-image variant pairs the same question with a randomly drawn image from a different patient and a different modality, sampled with a fixed seed to ensure reproducibility.

Grounding prompt (D6).

```
You are a medical AI assistant. Given the medical image and
the following question, provide your answer and identify the
region of the image relevant to your answer as a bounding box
in normalized coordinates [x1, y1, x2, y2] where each value is
in [0, 1].

Format:
Answer: <your answer>
Bounding box: [x1, y1, x2, y2]

Question: {question}
```

S32 DATASHEET

This appendix first datasheets **MedTrust-IN v0.3** (to be released upon publication) (§S32.1), then retains v0.2 only as superseded provenance (§S32.2). §S32.3 describes the *pre-registered, deferred* six-axis design, which is not a released artefact.

S32.1 MEDTRUST-IN V0.3: PANEL-DEFINED TEXT-ONLY-HARD CORPUS

Source of truth: `data/medtrust_in_test/{medtrust_in_stats.json, medtrust_in_v0.jsonl,DATASHEET.md}` (in the review supplement; public release upon publication); the Markdown datasheet is generated from the first two files by `gen_datasheet_v02.py`.

Composition.
1,122 closed yes/no items on source test splits: PathVQA 916, SLAKE 121, VQA-RAD 85; 960 yes / 162 no. The balanced subset contains 324 items, 308 from PathVQA and none from SLAKE.

Screen semantics.
An item is retained when at most one roster model answers correctly without the image. This is a *panel-defined text-only-hard* label, not an image-sufficiency claim or human adjudication of visual necessity.

Image-supported answerability.
The screen does not require success with the image. Existing per-item receipts show $59/1{,}122$ items are missed by every roster model even with the image; $1{,}063/1{,}122$ have at least one image-conditioned success and $871/1{,}122$ are solved by a roster majority. These are useful hard cases, not automatically invalid questions; downstream users can filter on `frac_image_correct` when demonstrated image-supported answerability is required.

Controls and provenance.
Every item is `recorded+derived`; each carries the constant screening roster, source split, seed, shuffled index and source reference. The builder fails closed on mixed splits and contradictory `image_ref`.

Licenses.
Items inherit their source licenses: SLAKE CC BY 4.0, VQA-RAD CC0 1.0; PathVQA items are distributed by reference only (`image_ref`, answer and a hash of the question, with a script that restores the text from the public mirror), because its dataset card states that copyright in the source images and captions stays with the publishers. Images are ref-

erenced, not redistributed; attribution for each source is in the datasheet and the supplement README.

Use boundary.
The full corpus has an 85.6% yes prior, so accuracy comparisons should use `balanced=true`; even then, the balanced core is effectively PathVQA.

S32.2 MEDTRUST-IN V0.2 (SUPERSEDED BY V0.3; RETAINED FOR THE RECORD)

Source of truth: `data/medtrust_in_v02/{medtrust_in_stats.json, medtrust_in_v0.jsonl, curated_deinversion.txt}`; builder `code/scripts/build_medtrust_in.py`. The superseded v0 artefact is retained under `data/medtrust_in/` for provenance; §4.4 explains why it was re-issued and what changed.

Composition.
1,581 closed-form yes/no medical-VQA items — VQA-RAD 179, SLAKE 748, PathVQA 654 — screened from source pools of 500, 1,681, and 2,400 items respectively (train splits, shuffled at seed 42). Each item carries question, gold answer, a license-clean `image_ref` into the source dataset (images are referenced, not re-hosted), provenance, and the curation fields `n_text_only_correct`, `frac_text_only_correct`, `n_models`, `screening_roster`, `positive_control`, `mean_cipe`, `balanced`.

Screen (panel-defined text-only-hardness).
An item is kept when at most *one* model of that dataset's screening roster answers it correctly with the image removed — a constant *integer* criterion, because a fractional threshold is not constant across roster sizes (τ=0.25 admits 2-of-8 but only 1-of-5, so under v0 the bar an item faced tracked its source index). The roster is the set of models that (a) pass the positive control on that dataset and (b) cover the whole source pool: PathVQA 6, SLAKE 5, VQA-RAD 5 (from Qwen2-VL-2B/7B, Qwen2.5-VL-7B, Qwen3-VL-8B, InternVL3-8B, Gemma-3-4B/27B, LLaVA-OneVision-7B). Per-item `n_models` is therefore constant within a dataset, and no item is screened by a partial panel. **The screen is defined by this panel and is not model-agnostic**: no human adjudicated whether a question is genuinely unanswerable without the image (annotator study deferred, Appendix S36), so the corpus is precisely a *panel-defined text-only-hard* set, and an item easy for a model outside the roster may survive the screen.

Positive control (new in v0.2).
Each dump is certified two ways before it may define any label: the recorded black-vs-white `image_injection_ok` flag, and a control derived from the dump itself (accuracy gain from the image, and the fraction of predictions unchanged when the image is removed). A model failing either on a dataset is disqualified *for that dataset entirely*, including its other runs. Disqualified: LLaVA-OneVision-7B (SLAKE, PathVQA), Qwen2-VL-2B (all three), Qwen2-VL-7B (VQA-RAD), Qwen3-VL-8B (SLAKE), and on VQA-RAD the five readout-artefact models (Aria, Fuyu-8B, Idefics2-8B, LLaVA-1.5-7B, VipLLaVA-7B): 12 model×dataset pairs in all. Per item, `positive_control` records which control certified its roster (1,402 items `recorded+derived`, 179 `derived`). **Read before use:** the 179 VQA-RAD items are `derived`-only because that roster predates the recorded black-vs-white control, and the derived control is computed from the same yes/no extraction whose trustworthiness §4.2 questions. They are the weakest-certified stratum in the corpus.

Label skew (read before use).
1,484 yes / 97 no (93.9% yes), *worse* than v0's 1,031/1,197 (86.1%). This is structural, not a regression: with the image hidden, panels fall back on "no" for these datasets, so "no"-gold items are answered correctly text-only almost for free and are screened out. Text-only-hardness is thus confounded with the answer prior by construction, and v0's milder skew was an artefact of a degenerate always-"yes" screener masking it. A constant-"yes" predictor scores 0.939 on the full set, so **no accuracy comparison should use the full corpus** — use `balanced=true`.

Balanced subset skew (read before use).
The balanced subset is 194 items (97 yes / 97 no), and its *source* composition is lopsided: PathVQA 184, SLAKE 2, VQA-RAD 8. PathVQA is the only source with enough "no" text-only-hard items to balance (92/92), so **the label-balanced core of MedTrust-IN is effectively a PathVQA (pathology) corpus**, and the SLAKE/VQA-RAD balanced cells (2/8) cannot power a comparison. The balanced subset is smaller than v0's 332 for the same reason the skew worsened.

Modality coverage.
Radiology (VQA-RAD), mixed radiology (SLAKE), pathology (PathVQA). No dermatology, ophthalmology, or other modalities.

Other cautions.
The `anatomy_mask` counterfactual on this corpus uses the centred 25%-area occlusion fallback (no per-item bbox), so it carries no annotated-region claim (§3).

Uses. Intended: evaluating visual grounding / faithfulness of medical VLMs; the population on which CIPE and the H1/H3 stratification are meaningful. *Not* intended: training, or clinical decision-making. Inherits any bias of the source datasets.

Distribution / maintenance.
To be released as `medtrust_in_v0.jsonl` with its datasheet and `medtrust_in_stats.json`; images via the cited source datasets. Reproducible from sources at seed 42. v0; a decircularized re-curation, per-item bboxes, balanced-"no" mining, and a medical specialist in the scoring roster are future work.

S32.3 Pre-registered six-axis design (deferred; not a released artefact)

The fields below describe the *pre-registered target* design, retained for transparency about scope. The numbers marked purple are design targets, not measurements, and no artefact corresponding to them has been built or released.

We adopt the datasheet template of Gebru et al. (2021). Selected fields of the design are summarised below.

Motivation.
MedTrust was created to evaluate the trustworthiness of medical vision-language models along six clinically grounded axes that go beyond accuracy. It was created by the authors of this paper.

Composition.
The design targets the per-dimension question–answer counts of Table 16, organised across six trustworthiness dimensions (the v0 work reports CIPE measurements on 200 VQA-RAD closed-form yes/no samples per model across 63 model evaluations, not a dimension-stratified subset). Each instance would be an image, a question, an optional reference answer, a dimension label, and provenance metadata (source dataset, modality, augmentation type).

Collection process.
All base images and questions would be derived from four publicly released datasets: VQA-RAD (Lau et al., 2018) (CC0 1.0), and SLAKE (Liu et al., 2021), OmniMed-VQA (Hu et al., 2024) and PathVQA (He et al., 2020), each released under the license stated by its original authors. Augmented probes (paraphrases, perturbations, OOD pairings, unanswerable questions) would be generated programmatically, and any LLM-generated content filtered before use.

Preprocessing.
Images are resized to a per-model native resolution. We do not perform colour normalisation or histogram equalisation, to avoid removing diagnostic signal.

Uses. Intended uses: trustworthiness evaluation of medical VLMs and methodology research on trustworthiness training objectives. *Not intended* for: training models to maximise the benchmark score (gaming risk), prospective clinical decision-making, or as a substitute for regulatory clearance.

Distribution.
Planned public release upon publication. Intended terms: code Apache 2.0; augmented data CC BY-NC 4.0; base images under their original dataset licenses.

Maintenance.
Planned annual versioned updates; no private split, issue tracker or leaderboard exists for the review version.

S33 Model Snapshots and Inference Settings

Table 21 reports the exact versions of the models actually evaluated in this paper. Unless otherwise stated, decoding uses temperature 0, top-p 1.0, max new tokens 256, and a 5-attempt exponential backoff on transient errors. All entries are open-weight and identified by their HuggingFace repository (exact commit revisions were not pinned in v0); no proprietary API model was evaluated (see Appendix S18).

Table 21: Model snapshots for the actually-evaluated panel. All open-weight entries are identified by their HuggingFace repository in `code/configs/models.yaml`; exact commit revisions were not pinned in v0. The proprietary panel of the deferred v1 design (Appendix S18) is not listed because it was not evaluated.

Model group (open-weight, evaluated)	Pinning	Decoding
Qwen2-VL / Qwen2.5-VL / Qwen3-VL (all scales, incl. Thinking)	HF repo (rev. not pinned)	T=0, bf16
InternVL3 (1B–38B)	HF repo (rev. not pinned)	T=0, bf16
Gemma3 / Gemma3n (4B–27B)	HF repo (rev. not pinned)	T=0, bf16
Idefics3 / SmolVLM / SmolVLM2	HF repo (rev. not pinned)	T=0, bf16
LLaVA-NeXT / LLaVA-OneVision / LLaVA-1.5 / Pixtral / Mistral-Small-3.1 / LFM2-VL	HF repo (rev. not pinned)	T=0, bf16
Idefics2-8B / Fuyu-8B (recovered readout-artefact cluster)	HF repo (rev. not pinned)	T=0, bf16
MedGemma-4B-it / MedGemma-27B-it (medical specialist)	HF repo (rev. not pinned)	T=0, bf16

S34 Compute and Cost Disclosure

We summarise the compute footprint of the evaluation *actually run* in this paper. All inference is local, open-weight, and zero-cost (no API spend): the CIPE sweep is a small number of single forward passes per (model, item, counterfactual) tuple at temperature 0, over closed-form yes/no candidate logprobs. The panel of 50 open-weight models plus MedGemma-4B/27B was evaluated at 200–500 VQA-RAD items $\times$ three counterfactuals each, plus the SLAKE n=400 region-validation runs, the n=500 per-dataset de-inversion runs, the full-pool test-split screening (e.g. PathVQA 3,362 items), the ChestX-Det runs, the 48 contamination runs and the certification pass. No response cache was in force for the reported runs (see the Reproducibility Statement), so reruns are full recomputation. Models up to 13B ran on a single A100/L40S-class GPU; 24–38B models on two. The total sweep is on the order of a few hundred GPU-hours of local inference; no proprietary-API queries, no paraphrase set, and no full six-dimension benchmark were run (those belong to the deferred v1 design, Appendix S18). A precise energy figure is not separately metered for the v0 sweep.

S35 Maintenance plan

The corpus accompanying this paper is **MedTrust-IN v0.3**, built on source *test* splits (1,122 items); v0.2 and v0 are retained and marked superseded, and §S11 tabulates all three builds. Releases

will follow semantic versioning, with `v1.0.0` tagged at publication. Versioned updates, a public issue tracker and a leaderboard are planned; none exists for the review version, and we make no commitment here about their schedule.

S36 Annotator Instructions and Inter-Rater Agreement

Recruitment (planned for camera-ready). Two medical annotators (one board-certified radiologist, one pathologist) will be recruited from a partner medical school via departmental email at the camera-ready phase; institution and per-hour compensation will be disclosed when the curated benchmark is released, and will be set above local minimum wage. The v0 manuscript does not yet rely on the human-annotation deliverables; the CIPE measurements use programmatic counterfactuals.

Tasks (planned). Annotators will perform two tasks: (i) bounding-box review for the D_6 grounding subset, where they accept, correct, or reject MedSAM-generated boxes, and (ii) verification of $\sim$500 randomly sampled augmented probes per dimension to confirm clinical plausibility (e.g., that a paraphrase preserves the required answer; that an unanswerable question is genuinely unanswerable from the image alone).

Agreement (pre-registered target). For the camera-ready release, inter-rater agreement will be computed on a subset of 200 items per task. The pre-registered *publication gate* is $\kappa \geq 0.75$ on all three tasks (bounding-box acceptance, paraphrase semantic equivalence, unanswerability); the camera-ready will report the achieved κ values. Disagreements are resolved by a third annotator (a senior physician); items without three-way agreement are dropped from the benchmark.

Annotator-facing instructions (excerpt).

```
Task: bounding-box verification.
1. View the image with the candidate bounding box overlaid.
2. Decide:
   ACCEPT  - the box tightly encloses the region named in the answer
   ADJUST  - the box overlaps the region but is loose; provide a
             tighter box
   REJECT  - the box is in the wrong region or off-image
3. If REJECT, provide a free-text reason.
4. Take no more than 60 seconds per item; flag for senior review
   if uncertain.
```

S37 Dataset Release and Reproducibility Checklist

We summarise compliance with community dataset-release and reproducibility best practices.

- **Datasheet:** Provided in Appendix S32, following Gebru et al. (2021).
- **License and intended use:** Intended terms: code Apache 2.0; augmentations CC BY-NC 4.0; base images under their original dataset licenses (Appendix S32).
- **Hosting and access:** Public release upon publication (planned hosting via a code repository, Hugging Face Datasets and Zenodo); for review, the v0.3 items with their statistics and datasheet (question text for SLAKE and VQA-RAD, PathVQA items by reference, no pixels) and an anonymized receipt subset are in the supplementary material (Reproducibility Statement).
- **Croissant metadata:** *Not yet released.* No Croissant record ships with v0; it is a camera-ready deliverable.
- **Reproducibility:** Evaluation code and prompts will be released upon publication; the region-audit receipts and verifier are in the supplementary material. Two corrections: model snapshots are identified by `model_path` but exact commit revisions were *not* pinned in v0, and there is *no* response cache in the pipeline that produced the reported numbers (Appendix S33; Reproducibility Statement).

- **Author responsibility:** The authors confirm that they bear all responsibility in case of license violations and that the artefact to be released does not contain personally identifiable information beyond what was already public in the source datasets.
- **Maintenance:** Plan in Appendix S35.
- **Limitations and broader impact:** Discussed in §5 and in the Broader Impact statement.
- **Compute disclosure:** Reported in Appendix S34.
- **Annotator compensation and consent:** No annotators were used in v0; the planned camera-ready protocol (recruitment, compensation above local minimum wage, consent to research use) is documented in Appendix S36.
- **TRIPOD-LLM:** A point-by-point checklist against the TRIPOD-LLM reporting guideline (Gallifant et al., 2025) is a camera-ready deliverable; we do not claim compliance for this version.

S38 Per-Model Detailed Results

The six-axis tables above describe a deferred design, not a released or committed camera-ready result. The current paper reports the executed CIPE measurements (Table 13) and the annotated-region audit, with their controls and negative outcomes. CIPE is not a validated cross-dimension diagnostic; total variation is the retained descriptive readout. No per-dimension coverage beyond the measured scope is claimed.

S39 Interval Estimates: Bootstrap Confidence Intervals and Permutation Tests

The legacy panel and construct-validation tables report point estimates. Here we attach 95% bootstrap confidence intervals and, where a table claims a null, a permutation test. All intervals are computed *post hoc from the committed per-item raw result files* (no re-running of any model) by `code/scripts/cipe_ci_appendix.py`, which writes the receipt `outputs/cipe_v0/ci_appendix.json`. Bootstrap: 2000 percentile resamples; a fixed per-table seed makes the receipt byte-reproducible. Point estimates match the corresponding Table 13 rows to rounding; the four bottom rows below are the n=500 sample-size-robustness runs, of which three are the † rows of Table 13 and Qwen2.5-VL-3B is a *separate* n=500 run (Table 13 lists Qwen2.5-VL-3B at n=200).

Table 22: **Bootstrap 95% CIs on mean CIPE.** The three per-counterfactual rows resample over items (one record per item, n=500); the aggregate row and the four bottom panel rows resample the flat $(\text{sample}, \text{counterfactual})$ records independently (n=1500), which does not preserve within-item dependence. *Top:* Table 12, Qwen2-VL-7B on VQA-RAD (n=500 samples per counterfactual). *Bottom:* four n=500 sample-size-robustness runs on Qwen2/2.5-VL backbones (n=1500 records each); three are the † rows of Table 13, and Qwen2.5-VL-3B is a separate n=500 run (listed at n=200 in Table 13).

Table	Row	mean CIPE	95% CI
12	image_swap (C1)	0.712	[0.677, 0.746]
12	anatomy_mask (C2)	0.950	[0.939, 0.959]
12	text_only (C4)	0.463	[0.432, 0.497]
12	aggregate	0.708	[0.690, 0.727]
13	Qwen2-VL-2B†	0.855	[0.843, 0.868]
13	Qwen2.5-VL-3B	0.794	[0.779, 0.808]
13	Qwen2-VL-7B†	0.708	[0.689, 0.726]
13	Qwen2.5-VL-7B†	0.633	[0.612, 0.654]

Reproducibility, Ethics and Disclosure Checklist

1. **Claims. [Yes]** The abstract and introduction state what the body supports: an annotated-region sensitivity audit with a foreground-matched control, a panel-defined text-only-hard corpus, and

Table 23: **Region-annotation construct signal (Table 18): CI on the paired effect and a permutation test of the null.** Δ=$\mathrm{CIPE}_{\text{random}}-\mathrm{CIPE}_{\text{relevant}}$ per item (>0 means masking the annotated object region drops CIPE more); 95% CIs and $\texttt{frac}(\mathrm{CIPE}_{\text{rel}}<\mathrm{CIPE}_{\text{rand}})$ are bootstrapped over the n=400 paired items. *perm.* p is a paired sign-flip permutation test of exchangeability (H_0: $\mathbb{E}[\Delta]=0$), 2000 shuffles; $<5\times10^{-4}$ denotes $p<1/2000$. Every Δ CI excludes 0. These intervals describe the retracted unmatched-control contrast (Table 18); they are not evidence of annotated-region sensitivity.

Model	Δ mean	Δ 95% CI	frac	frac 95% CI	perm. p
Qwen2-VL-7B	0.023	[0.012, 0.034]	0.670	[0.620, 0.715]	$<5\times10^{-4}$
Qwen2.5-VL-7B	0.051	[0.033, 0.069]	0.657	[0.613, 0.705]	$<5\times10^{-4}$
InternVL3-8B	0.051	[0.037, 0.067]	0.703	[0.660, 0.745]	$<5\times10^{-4}$
Gemma3-4B-it	0.056	[0.028, 0.087]	0.635	[0.588, 0.680]	$<5\times10^{-4}$
MedGemma-4B-it	0.097	[0.066, 0.131]	0.700	[0.652, 0.743]	$<5\times10^{-4}$

Table 24: **De-inversion collapse (Table 15): CI and permutation test on the text-only-hard (H1) AUROC.** Per-model $\mathrm{AUROC}(\mathrm{CIPE}\rightarrow\text{correct})$ on the text-only-hard stratum, averaged over the m models; 95% CI is a cluster (over-model) bootstrap and *perm.* p is a per-model label-permutation test of the null AUROC=0.5 (2000 shuffles). Under the *naive* (circular) readout the de-inversion is strongly below chance everywhere ($p<10^{-3}$); after *decircularization* it is indistinguishable from chance on VQA-RAD and SLAKE (CI straddles 0.5, p non-significant) and survives only weakly on PathVQA—exactly the honest reading in §4.4.

Dataset	Readout	H1 AUROC	95% CI (m models)	perm. p vs 0.5
VQA-RAD	naive	0.216	[0.125, 0.311] (m=7)	<0.001
VQA-RAD	decirc.	0.512	[0.445, 0.574] (m=7)	0.616
SLAKE	naive	0.215	[0.165, 0.271] (m=7)	<0.001
SLAKE	decirc.	0.469	[0.407, 0.524] (m=8)	0.140
PathVQA	naive	0.119	[0.069, 0.166] (m=6)	<0.001
PathVQA	decirc.	0.431	[0.413, 0.446] (m=6)	<0.001

an audit of failed or unidentifiable interpretations. SLAKE boxes are explicitly image-level rather than question-target links; only $229/400$ items admit a matched placement; the MedGemma ranking is exposure-confounded; CIPE loses to simpler readouts; and the amended ChestX-Det follow-up is labelled hypothesis-generating rather than confirmatory.

2. **Limitations. [Yes]** §5, plus per-result caveats in situ. The foreground control is complete but placement-selective; source image IDs were omitted from old receipts and recovered conditionally from metadata. The post-hoc cluster audit retains specialist effects but shows general-model significance is tie-convention-dependent (Appendix S5). MedTrust-IN v0.3 is rebuilt from test splits and positive-controlled, but test membership does not prove non-exposure and text-only-hardness does not imply image sufficiency.
3. **Theory assumptions and proofs. [Yes]** Proposition 1 states its assumptions in §3.1: a monotone foreground-dose shift, exchangeable noise independent of placement and a continuous symmetric F, plus a Lipschitz h and a bounded density for part (c). The proof and its scope remarks are in Appendix A. The proposition and proof were drafted by an LLM agent and checked by a separate LLM agent.
4. **Experimental result reproducibility. [Yes]** with a stated scope. The metric, counterfactuals, prompts, decoding settings, seed (42) and model paths are specified (§3, App. S33); per-record dumps will be released upon publication, and the region-run receipts are in the supplementary material. Exact model commit revisions were not pinned in v0, so results are reproducible up to HuggingFace checkpoint drift. The Reproducibility Statement states that no reported number used a response cache and that no held-out split exists.
5. **Open access to data and code. [No]** for review; code, v0.2/v0 of the corpus, the builder, the per-record result dumps and the larger-pool screening dumps (SLAKE n=1,681, PathVQA n=2,400) that defined 43% of the v0 corpus will be released upon publication. The review supplement contains MedTrust-IN **v0.3** (1,122 test-split items; question text for SLAKE and VQA-RAD, PathVQA items by reference; no pixels) with its datasheet, the region-audit receipts and an offline verifier, the cited registrations, and the replay and ChestX-Det receipts (Reproducibility Statement). No LICENSE file exists yet; the intended terms are stated in App. S32 (code Apache 2.0, augmentations CC BY-NC 4.0, source images under their original licenses) and the file is a camera-ready deliverable. No Croissant record ships with v0.
6. **Experimental setting / details. [Yes]** §3 and App. S33 give the counterfactual definitions, answer extraction, sample sizes, and per-model inference settings.
7. **Experiment statistical significance. [Yes]** App. S39 reports 95% bootstrap CIs and permutation tests; the scaling fit is a mixed-effects model with the pseudo-replication correction applied (certified 21: MixedLM p=0.033, OLS p=0.237, specification-dependent and not claimed; all-40 p=0.064 for comparison only); the region audit reports item-bootstrap BH q-values with post-hoc image-cluster intervals and sign-flip tests. The construct-validation AUROC is reported per dataset because it reverses sign by dataset, and is recomputed after excluding the 4 of 21 evaluations that fail the positive control (unchanged: $0.5291 \to 0.5295$). The contamination probe bootstraps each train-vs-test gap (10,000 resamples) and reports a 95% CI per cell.
8. **Experiments compute resources. [Yes]** App. S34; no reported number relies on cached responses, so reruns are full recomputations.
9. **Code of ethics. [Yes]** The work conforms to the ICLR Code of Ethics; see the Ethics Statement.
10. **Broader impacts. [Yes]** Broader Impact statement: the intended use (routine auditing of image use) and its limits (no clinical evidence; the corpus's answer prior).
11. **Safeguards. [NA]** The release is an evaluation corpus of already-public, de-identified medical VQA items and per-record logprob summaries; it carries no model weights, no generative capability, and no new patient data, so no release safeguards are required. The corpus is explicitly *not* a clinical clearance instrument (Broader Impact).
12. **Licenses for existing assets. [Yes]** SLAKE (CC BY 4.0), VQA-RAD (CC0 1.0), PathVQA (recorded as MIT-licensed by its public mirror; the original terms could not be verified at source) and ChestX-Det (annotations under Apache 2.0; images are a subset of NIH ChestX-ray14) are cited and used under these licenses. No image pixels are redistributed; the review supplement carries question/answer text and annotation metadata from these sources with per-source license and

attribution notices in its README. PathVQA's dataset card notes that copyright in its source images and captions stays with the textbook publishers and PEIR, so PathVQA items are distributed by reference only (image reference, answer and a question hash, with a script that restores the text from the public mirror).

13. **New assets documentation. [Yes]** MedTrust-IN v0.3 will ship (upon publication; see item 5) with a Gebru-style datasheet generated from the build receipt, a builder, per-item provenance (source dataset, split, seed, shuffled index), per-item screening-roster and positive-control fields, and a build-time gate that refuses to emit a corpus in which an item's image reference contradicts its source split (a defect that shipped once).
14. **Crowdsourcing and research with human subjects. [NA]** No human subjects and no annotators were used for any reported v0 result; all counterfactuals are programmatic. The annotator protocol in App. S36 is deferred design, not conducted work.
15. **IRB approvals. [NA]** No human-subjects research was conducted. All source data are public, de-identified, previously released datasets.
16. **Declaration of LLM usage. [Yes]** See the Use of AI Assistants statement.