FAIR PROTOCOL REVEALS COLLAPSE OF MEDICAL HALLUCINATION DETECTORS: TWO DETECTORS THAT SURVIVE IT

Anonymous authors
Paper under double-blind review

ABSTRACT

State-of-the-art hallucination detectors for medical VQA fail to reliably measure whether answers are grounded in the image. We show that reported progress is largely an artifact of benchmark leakage: detectors exploit surface statistics like answer length, formatting, and sampling dispersion that correlate with hallucination only within the benchmarks used for tuning. We introduce a fair evaluation protocol that strips away these shortcuts via length- and type-matched negatives, image-level splits, and language-leakage controls. Under this protocol, the evaluated prior methods perform near chance. VASE, SE, RadFlag, and a frozen LMM all score an AUROC of 0.49–0.57 across three benchmarks. A plain trainable MLP reaches 0.67–0.97, so model capacity is not the bottleneck; an input ablation shows that the residual signal is dominated by question-answer statistics rather than image grounding. We then ask what survives. QCE-Net, an interpretable detector that pools image patches with both question and answer as attention queries, matches or exceeds the MLP baseline, recovering the grounding signal that frozen encoders can provide. LLaVA-LoRA, which fine-tunes the medical LMM with LoRA to emit per-sample grounding judgments, goes further: it reaches 0.885 AUROC on VQA-RAD, the first detector under a leakage-free protocol to clear AUROC 0.8, an operational clinical-usability bar not reached by the frozen-CLIP methods evaluated here.

1 INTRODUCTION

Large multimodal models (LMMs) now answer medical visual questions with fluent, confident prose and are increasingly proposed for clinical triage, radiology reporting, and patient-facing explanation (Li et al., 2023c; Guan et al., 2024; Wu et al., 2024). Yet they routinely fabricate findings a careful radiologist would never report (Liao et al., 2025; Gu et al., 2026). In safety-critical medicine such visual hallucinations are not cosmetic errors but failures with real patient impact, so whether an answer is actually grounded in the image is a first-class question for the deployment of medical AI (Huang et al., 2024).

We define hallucination by grounding: an answer is hallucinated when the image contains no supporting evidence. Measuring grounding is harder in medicine (Rohrbach et al., 2018; Hessel et al., 2021): images are low-texture and annotated hallucination data is scarce, so prior detectors lean on cheap indirect signals—semantic-entropy dispersion (Farquhar et al., 2024; Kuhn et al., 2023) or internal confidence—which prior work shows are easily exploited through language priors (Huang et al., 2024; Leng et al., 2024).

We find these indirect signals brittle in practice: under a fair protocol with length- and type-matched negatives, image-level splits, and explicit language-leakage checks, the evaluated baselines perform near chance (0.49–0.57 AUROC), while a simple trainable MLP on raw frozen CLIP embeddings already reaches 0.67–0.97 AUROC, so the benchmark, rather than detector capacity, is the primary binding factor. Most clearly, VASE (Liao et al., 2025), reimplemented at full scale on both LLaVA-Med-1.5 (Li et al., 2023a) and MedGemma-4b, reaches only 0.49–0.51 on all three benchmarks, its apparent success coming from length and sampling-dispersion statistics.

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.

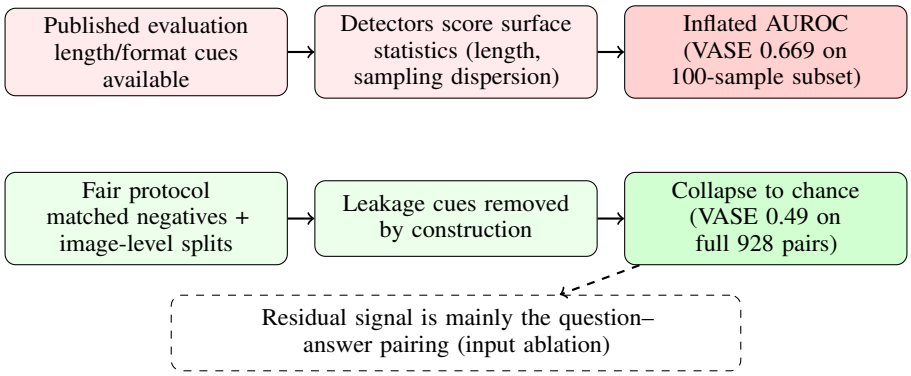


Figure 1: The paper's core empirical finding: published evaluation lets detectors exploit length- and format-based shortcuts (top row); the fair protocol removes them, leaving several evaluated detectors near chance (bottom row). The surviving signal is dominated by the question–answer pairing, while image evidence contributes little under the evaluated protocol.

How, then, can hallucination be detected? Existing approaches fall into three broad families: sampling-based uncertainty estimation, internal-signal reading of model confidence, and reference-free verification with an external model (Radford et al., 2021; Hessel et al., 2021; Lin et al., 2024). A recurring weakness of the first two is that they measure properties of the generated text rather than the image evidence, so they can be satisfied by plausible-sounding answers (Yin et al., 2024); verification is more direct, yet as we show, it can also be sensitive to surface statistics rather than genuine grounding (Liao et al., 2025).

An input ablation (Section 4.3) shows the trainable model's signal comes from the question–answer pairing: on VQA-RAD, the answer and question alone (no image) give AUROC 0.679 versus 0.670 with the image. Current trainable detectors thus rely heavily on question–answer statistics, while the contribution of image evidence is small under the evaluated protocol.

We introduce two detectors. QCE-Net scores grounding with frozen CLIP by pooling patch features with the question and answer as attention queries, and its per-patch weights double as evidence maps; LLaVA-LoRA LoRA-fine-tunes LLaVA-Med-1.5 (Li et al., 2023a) to judge each sample grounded or ungrounded.

On the three benchmarks both detectors beat the baselines that fall to chance under the fair protocol; LLaVA-LoRA clears AUROC 0.8 on VQA-RAD, the operational clinical-usability bar we adopt (Section 4.2).

On VQA-RAD the frozen-CLIP deficit concentrates in yes/no pairs (54%): CLIP embeds yes and no nearly identically (cosine 0.95), capping QCE-Net near 0.70, a limitation that LLaVA-LoRA avoids by training the medical backbone (Section 4.2).

We make three contributions:

(i) We propose a fair evaluation protocol for medical hallucination detection, enforcing length- and type-matched negatives, image-level splits, and explicit language-leakage checks. Applying it reveals an evaluation artifact that prior benchmarks do not control for: the evaluated detectors, including a full-scale VASE under two backbones, perform near chance, so the reported gains are largely attributable to benchmark leakage rather than grounding capability; the residual signal is dominated by the question–answer pairing, and image evidence contributes little under the evaluated protocol.

(ii) We propose QCE-Net, an interpretable detector that builds a trainable head over frozen CLIP features, pooling image-patch features with the question and answer as attention queries; it is significantly better than a matched trainable MLP baseline on VQA-RAD, and its patch weights double as evidence maps.

(iii) We propose LLaVA-LoRA, a medical-LMM detector fine-tuned with LoRA to judge each sample grounded or ungrounded; it clears the 0.8 clinical-usability bar on VQA-RAD under a leakage-free protocol.

2 Related Work

Object-level metrics such as CHAIR (Rohrbach et al., 2018) first quantified hallucination in vision–language models, and adversarial benchmarks including POPE (Li et al., 2023c), HallusionBench (Guan et al., 2024), and medical variants (Wu et al., 2024; Gu et al., 2026) expose answers driven by language priors rather than image content, attributed to over-trust in learned text priors (Huang et al., 2024) and insufficient visual grounding in decoding (Leng et al., 2024; Chuang et al., 2024), with mitigations from contrastive decoding (Leng et al., 2024; Chuang et al., 2024) to external-knowledge correction (Yin et al., 2024) and surveys organizing the field (Ji et al., 2023; Huang et al., 2025); in medicine the stakes are higher, yet dedicated detectors remain scarce and hallucination detection has only recently been framed as a standalone task (Liao et al., 2025). Detection methods cluster into three families: sampling-based uncertainty such as semantic entropy (Farquhar et al., 2024; Kuhn et al., 2023) and answer-consistency ratios; internal-signal methods using logits or attention (Chuang et al., 2024); and reference-free verification scoring answer–image agreement with an external model, such as CLIPScore (Hessel et al., 2021) and VQAScore (Lin et al., 2024). Our full-scale reproduction shows that, under a fair protocol that removes length and format leakage, several representative methods in the first two families perform near chance, including a full-scale VASE (Liao et al., 2025) under two backbones; ViEAD, our hand-crafted reference detector, belongs to the verification family but characterizes the spatial distribution and perturbation vulnerability of the evidence, which provides a complementary signal for distinguishing hallucinated answers. Measuring whether text is grounded in an image builds on CLIP-based alignment (Radford et al., 2021; Hessel et al., 2021), attention-based grounding (Rohrbach et al., 2018), and dense patch-level alignment from self-supervised encoders (Oquab et al., 2024); perturbation-based diagnosis (Zeiler & Fergus, 2014; Kuhn et al., 2023) underlies ViEAD, which measures how answer–image alignment responds to targeted masking of the strongest evidence patches, feature noise, and spatial permutation in the feature domain, and to which hallucinated answers are insensitive because they never relied on the evidence. Medical VQA rests on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and Path-VQA (He et al., 2021), with LMM backbones specialized for clinical images, for example LLaVA-Med (Li et al., 2023a), InstructBLIP (Dai et al., 2023), BLIP-2 (Li et al., 2023b), mPLUG-Owl (Ye et al., 2023), Med-Gemini (Saab et al., 2024), MedGemma, and generalist clinical models (Tu et al., 2024). Alignment backbones such as BiomedCLIP (Zhang et al., 2025) and PubMedCLIP (Eslami et al., 2023) improve on this base, as do preference optimization (Zhu et al., 2025), retrieval-augmented generation (Xia et al., 2025), and multi-agent reasoning (Xia et al., 2026). Recent work also calls attention to shortcut cues that inflate scores without real capability (Gu et al., 2026), which motivates our protocol; we extend this line with a systematic evaluation and two detectors: QCE-Net, an interpretable detector that makes visual evidence explicit via question-conditioned attention pooling and is significantly better than a matched trainable MLP on VQA-RAD while tied elsewhere, and LLaVA-LoRA, a LoRA-fine-tuned medical-LMM detector that reaches AUROC 0.885 on VQA-RAD.

3 Method

3.1 Problem Setup, Fair Protocol, and Hypothesis-Testing View

We formalize medical visual hallucination detection as a grounding task. Given a medical image I, a question Q, and a model-generated answer A, the detector must assign a score $s(I, Q, A) \in [0, 1]$ that is larger for answers that are not supported by the image content, with ground-truth labels $y \in \{0, 1\}$ (1 = hallucinated) assigned by the construction in the footnote.[1]

This frames detection as a grounding problem: the score must reflect whether the image contains evidence for the answer, not whether the answer is fluent or statistically unusual. The principal obstacle is evaluation leakage: a detector that scores answer length or format appears effective on any benchmark where hallucinated answers are longer, without measuring grounding. We therefore adopt a fair protocol that removes such shortcuts before any detector is evaluated, enforcing three constraints. (i) Matched negatives: each negative pair keeps the image and question of a positive pair and replaces

[1] Positives are expert-annotated ground-truth answers; negatives are length- and type-matched answers sampled from other images, so no new clinical annotation is required. Section 4.5 quantifies the small fraction of negatives that may nevertheless be valid for their test image.

only the answer, which is drawn from answers of other images with matching token length (± 2) and answer type (yes/no versus open-ended; for yes/no questions it is the opposite-polarity answer), removing length and format as cues by construction. (ii) Image-level splits: all pairs from one image stay in a single split, so no image straddles train and test. (iii) Leakage controls: a detector trained on lexical features alone is expected to score near chance; we report this as a control.

Constraint (ii) is standard in image-level evaluation, but (i) and (iii) are frequently absent from hallucination-detection benchmarks. As our experiments show, under these two constraints the evaluated detectors perform near chance, including a representative medical-specific method, VASE (Liao et al., 2025), under two backbones.

We cast detection as testing the hypotheses

$$\mathcal{H}_0: \text{ the image supports } A \quad \text{versus} \quad \mathcal{H}_1: \ A \text{ is not supported by the image.} \tag{1}$$

Under $\mathcal{H}_1$, the answer is generated from a language prior $\Pr(A \mid Q)$ that is approximately independent of the image content; under $\mathcal{H}_0$, it is conditioned on specific visual evidence. This yields two testable predictions: under $\mathcal{H}_1$, answer–image alignment statistics cluster near those of an independently paired answer; and destroying image evidence does not change the alignment, because the alignment never encoded that evidence. The hand-crafted ViEAD reference tests the second prediction directly, by measuring how answer–image alignment responds to masking of the strongest evidence patches (Appendix A.7). QCE-Net instantiates these predictions through the question-conditioned evidence channel (Section 3.2) and the fusion architecture (Section 3.3).

3.2 Representation and Dual-Conditioned Evidence Grounding

QCE-Net builds on a frozen CLIP-L/14@224 model (Radford et al., 2021). Let $\mathbf{P} \in \mathbb{R}^{N \times d}$ be the $N = 256$ normalized patch embeddings of I, $\mathbf{g} \in \mathbb{R}^d$ its global (CLS) embedding, $\mathbf{a} \in \mathbb{R}^d$ the normalized embedding of A, and $\mathbf{q} \in \mathbb{R}^d$ the embedding of Q, all in the shared CLIP text–vision space ($d = 768$). Our analysis (Section 4.3) motivates a representation that does not reduce grounding to answer–image similarity alone: it must also capture the question–answer pairing that currently dominates trainable detectors, and add a signal those detectors lack, whether the image contains the evidence the question points to.

The core novelty of QCE-Net is dual-conditioned evidence: we pool the image's patch features using both the question and the answer as attention queries. The question-conditioned alignment is

$$v_i = \langle \mathbf{p}_i, \mathbf{q} \rangle, \qquad \mathbf{v} = \mathbf{P}\,\mathbf{q} \in \mathbb{R}^N, \tag{2}$$

and the answer-conditioned alignment is $\mathbf{u} = \mathbf{P}\mathbf{a}$. The two evidence vectors are attention-weighted sums

$$\mathbf{e}_q = \sum_{i=1}^{N} w_i^q\, \mathbf{p}_i, \quad \mathbf{e}_a = \sum_{i=1}^{N} w_i^a\, \mathbf{p}_i, \quad w_i^q = \frac{e^{\kappa v_i}}{\sum_j e^{\kappa v_j}}, \quad w_i^a = \frac{e^{\kappa u_i}}{\sum_j e^{\kappa u_j}}, \tag{3}$$

with temperature $\kappa = 2$. $\mathbf{e}_q$ summarizes where in the image the question points, concentrated when the image contains what the question asks about; $\mathbf{e}_a$ summarizes where the answer's content binds. These complement the global signal in a structured way.

An answer-only score has a structural ceiling under matched pairs. Let $\phi : \mathcal{T} \to \mathbb{R}^d$ be the L2-normalized CLIP text encoder and let $s_A(I, Q, A) = h(\phi(A))$ be any score that depends on the answer embedding alone, with h L-Lipschitz. Because a positive triple (I, Q, A) and its matched negative (I, Q, A') share I and Q, only the answer changes, so

$$|s_A(I, Q, A) - s_A(I, Q, A')| \ \le \ L\, \|\phi(A) - \phi(A')\|. \tag{4}$$

For a yes/no pair, $\cos(\phi(\text{yes}), \phi(\text{no})) = 0.949$, hence $\|\phi(\text{yes}) - \phi(\text{no})\| = \sqrt{2(1 - 0.949)} \approx 0.32$, so any answer-only score separates a yes/no matched pair by at most $0.32L$ in embedding space. This bound applies to any answer-only score under the matched-pair construction, and it explains why frozen-CLIP methods are capped near chance on VQA-RAD's yes/no block (Appendix A.9).

Question-conditioned evidence exceeds this bound on open-ended pairs. Let $\mathbf{e}_a = \sum_i w_i \mathbf{p}_i$ with $w_i = \mathrm{softmax}(\kappa\, \mathbf{P}\mathbf{a})_i$ (Equation 3). The answer-conditioned evidence is a smooth function of the answer embedding whose Jacobian

$$\frac{\partial \mathbf{e}_a}{\partial \mathbf{a}} = \kappa \sum_i w_i\, \mathbf{p}_i\, (\mathbf{p}_i - \mathbf{e}_a)^\top \tag{5}$$

is not bounded by 1 in general: when the answer aligns strongly with a few patches, $\|\partial \mathbf{e}_a / \partial \mathbf{a}\|$ grows with κ and the patch dispersion $\sum_i w_i \|\mathbf{p}_i - \mathbf{e}_a\|^2$, so the evidence channel amplifies answer differences that the raw answer embedding compresses. Moreover $\mathbf{e}_q$ depends on (I, Q) only and is identical across a matched pair, so it contributes a channel that is invariant to the matched-pair construction. Consequently QCE-Net separates a matched pair through two complementary mechanisms, an answer-selective evidence channel that amplifies genuine grounding differences, and a question-conditioned channel that is immune to the answer-channel collapse of Equation 4. On yes/no pairs the question-conditioned channel is identical across the matched pair and therefore contributes no separating signal; the gain applies to open-ended pairs, where the matched negative is a different concrete answer, consistent with the by-type analysis in Appendix A.3.

Together, Equation 4 caps any answer-only score under matched pairs, and the Jacobian of Equation 5 is the mechanism by which QCE-Net exceeds that bound. On VQA-RAD, where the yes/no block is a polarity pair, the answer channel is capped and the evidence channels carry the significant gain; on Path-VQA, where matched negatives are concrete answers from other images, the answer channel is already strong and the evidence channels add little (Appendix A.3). Full derivations of both bounds are given in Appendix A.16.

3.3 Architecture and Training

QCE-Net scores a sample with a small MLP head

$$s(I, Q, A) = \sigma\Big(\mathrm{MLP}\big([\,\mathbf{g},\ \mathbf{a},\ \mathbf{q},\ \mathbf{e}_q,\ \mathbf{e}_a\,]\big)\Big), \tag{6}$$

that is, the 5-way concatenation of the global image, answer, question, and the two conditioned evidence vectors ($5 \times 768 = 3840$ dimensions) is mapped through a 3-hidden-layer MLP (1024–512–256; BatchNorm, ReLU, dropout $0.4/0.4/0.3$) to a scalar logit. All modules are trainable; CLIP is frozen. The network is trained end-to-end on the fair training split to minimize the binary cross-entropy

$$\mathcal{L}(\theta) = -\tfrac{1}{|\mathcal{D}|} \sum_{(I,Q,A,y)\in\mathcal{D}} \Big[\, y \log s(I, Q, A) + (1 - y) \log\big(1 - s(I, Q, A)\big)\Big], \tag{7}$$

with AdamW (lr 10^{-3}, weight decay 10^{-4}), batch size 256, and a cosine learning-rate schedule over 120 epochs. The evidence pooling in Equation 3 uses a fixed temperature ($\kappa = 2$, selected by validation AUROC over the sweep $\kappa \in \{2, 3, 4, 5, 6, 8\}$ and competitive with a learned query projection; Section 4.4) and the raw question and answer embeddings, keeping the method interpretable: the attention weights w_i^q and w_i^a are per-patch maps of where the image supports the question and the answer, which can be visualized per sample.

The input ablation in Section 4.3 shows that removing the image costs almost nothing to current trainable detectors, while removing the question–answer pairing is far more damaging. QCE-Net is designed so that the visual-evidence channels ($\mathbf{e}_q, \mathbf{e}_a$) are first-class inputs rather than an afterthought.

3.4 Detector Variants

QCE-Net inherits the representational ceiling of its frozen encoder. LLaVA-LoRA avoids this ceiling by training the medical backbone itself. We take LLaVA-Med-1.5 (Li et al., 2023a) (a Mistral-7B LMM tuned on medical instruction data, 7.5B parameters) and attach low-rank adapters

$$W_q' = W_q + \frac{\alpha}{r} B_q A_q, \quad W_v' = W_v + \frac{\alpha}{r} B_v A_v, \tag{8}$$

with $r = 8$, $\alpha = 16$, $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$ (about 4.2M trainable parameters, 0.055% of the model), freezing all other weights. The adapted model is fine-tuned on the fair training split with a token-level classification target: for each (image, question, answer) triple the model is prompted to emit the token GROUNDED (label 0) or UNGROUNDED (label 1), trained with the cross-entropy over those two tokens

$$\begin{aligned} \mathcal{L}_{\mathrm{LoRA}} = \ & -\log \Pr\big(\text{ungrounded} \mid I, Q, A; \theta + \Delta\theta\big)\ \delta_{y=1} \\ & -\log \Pr\big(\text{grounded} \mid I, Q, A; \theta + \Delta\theta\big)\ \delta_{y=0}. \end{aligned} \tag{9}$$

At inference we score the sample by the difference of the two token logits, feeding the AUROC analysis with a per-sample real-valued score. Because the LMM does not rely on the CLIP text encoder, it escapes the yes/no polarity collapse that caps frozen CLIP detectors on VQA-RAD (Appendix A.9), and it contributes its full grounding competence rather than being reduced to a pair of frozen embeddings. The frozen zero-shot LLaVA-Med judge (same backbone, no training; Appendix A.4) is near chance, so LLaVA-LoRA's gain over frozen baselines is predominantly attributable to training on the protocol rather than to the backbone. Training details are in the appendix.

We also report ViEAD, a hand-crafted reference detector that trains no vision–language model, in Appendix A.7.

4 Experiments

4.1 Setup

Datasets. We evaluate on three medical VQA benchmarks: VQA-RAD (Lau et al., 2018) (radiology, 928 test pairs), Path-VQA (He et al., 2021) (pathology, 1676), and OmniMed-VQA-mini (Hu et al., 2024) (multi-modality, 802). OmniMed is answer-open (only 3.8% yes/no versus 54% on VQA-RAD), so it stresses grounding where it is well-defined. SLAKE (Liu et al., 2021) (multi-organ, 2008 test pairs) appears only in Table 2, an additional check that the collapse is benchmark-agnostic. Test pairs follow the fixed image-level split of our fair protocol (Section 3.1).

Baselines. We compare against two tiers. Training-free detectors: VASE (Liao et al., 2025) reimplemented faithfully at full scale from the published algorithm under LLaVA-Med-1.5 (Li et al., 2023a) and MedGemma-4b, with its internal components SE (Farquhar et al., 2024) and RadFlag, plus recent detectors on sampled sets; and a zero-shot LLaVA-Med judge (the frozen backbone scored by its grounded/ungrounded token-logit difference, with no protocol training). Training-based detectors: a plain MLP on raw CLIP embeddings (Section 4.3) and our QCE-Net. The primary comparison for QCE-Net is the training-based MLP. AUROC uses hallucination as the positive class; significance is a paired bootstrap on the AUROC difference (DeLong-style) (DeLong et al., 1988).

4.2 Main Results

Table 1 reports full-set results. All VASE-family variants—among the few medical-specific detectors—score between 0.491 and 0.510, close to chance under both backbones and across all three datasets, with every score within 0.012 of 0.5. SE and RadFlag are equally non-diagnostic; the VASE numbers come from this reimplementation. The zero-shot LLaVA-Med judge is likewise near chance (0.555/0.496/0.572 on VQA-RAD/Path-VQA/OmniMed), while the same model fine-tuned on the protocol reaches 0.885–0.969, so LLaVA-LoRA's gain comes from protocol training, not the backbone.

We compare the two trainable detectors against a matched-capacity trainable MLP baseline with a 2,000-sample paired bootstrap. QCE-Net reaches 0.702 (VQA-RAD), 0.868 (Path-VQA), and 0.9674 ± 0.0019 (OmniMed, mean over 8 seeds) AUROC; the matched MLP reaches 0.670, 0.867, and 0.9700 ± 0.0019. The difference is statistically significant on VQA-RAD ($+0.032$, $p = 0.006$), where the evidence channels rescue signal that a plain CLIP-embedding MLP misses, and is not significant on Path-VQA and OmniMed (the comparison is capacity-matched; Section 4.4). LLaVA-LoRA reaches 0.885 (VQA-RAD), 0.894 (Path-VQA), and 0.9691 ± 0.0019 (OmniMed, mean over 4 seeds), gains of $+0.215$ ($p < 0.001$), $+0.027$ ($p < 0.001$), and -0.001 ($p = 0.67$) over the MLP. LLaVA-LoRA achieves AUROC ≥ 0.8 on VQA-RAD under this protocol, a threshold not reached by prior detectors in this setting, with only 4.2M trainable parameters (0.055% of the model; Section 3.3). The hand-crafted ViEAD reference reaches 0.59–0.89 (Table 1; Appendix A.7).

Two additional analyses support this finding. First, on a 100-sample subset of VQA-RAD our reproduction of VASE reaches 0.6693, within the originally reported range, while on the full 928 pairs it is 0.4941, indicating that subset size can substantially affect measured performance. Second, we additionally evaluate standard reference-free scorers (CLIPScore (Hessel et al., 2021), CHAIR-Medical (Rohrbach et al., 2018), RULE) and lightweight fine-tunes of medical alignment models

Method	VQA-RAD	Path-VQA	OmniMed
VASE (LLaVA-Med, full)	0.4941	0.4950	0.4911
VASE (MedGemma, full)	0.4991	0.5029	n/a
SE/RadFlag (best of both)	0.5098	0.5020	0.4988
Zero-shot LLaVA-Med judge	0.5547	0.4961	0.5717
ViEAD (hand-crafted, reference)	0.5904	0.7631	0.8864
MLP (raw CLIP embeddings, trained)	0.6702	0.8666	0.9700 ± 0.0019
QCE-Net (frozen CLIP + trainable head, ours)	**0.7019**	0.8678	0.9674 ± 0.0019
LLaVA-LoRA (trainable LMM, ours)	**0.8854**	**0.8937**	0.9691 ± 0.0019
diff (p, QCE − MLP)	$+\mathbf{0.032}$ ($\mathbf{0.006}$)	+0.001 (0.81)	−0.003 (0.40)
diff (p, LLaVA − MLP)	$+\mathbf{0.215}$ ($\mathbf{0.000}$)	$+\mathbf{0.027}$ ($\mathbf{0.000}$)	−0.001 (0.67)

Table 1: Full-set AUROC (higher better) under the fair protocol. VASE is a full-scale reproduction under the LLaVA-Med backbone; on OmniMed MedGemma is gated by the provider ("n/a"). All trainable methods use matched protocols, identical test pairs, and validation-only model selection; the paired-bootstrap p-value tests the AUROC difference against the MLP (2,000 resamples). OmniMed: mean over seeds ($\pm$ std; 8 for MLP/QCE-Net, 4 for LLaVA-LoRA). Among the compared detectors, QCE-Net achieves the highest AUROC among frozen-CLIP methods, and LLaVA-LoRA achieves the highest AUROC on these benchmarks.

General-domain detector	VQA-RAD	SLAKE	Path-VQA
CLIPScore (Hessel et al., 2021)	0.503	0.518	0.502
RULE	0.488	0.573	0.493
CHAIR-Medical (Rohrbach et al., 2018)	0.519	0.550	0.500
MMedPO fine-tune (Zhu et al., 2025)	0.495	0.510	0.499
MMed-RAG fine-tune (Xia et al., 2025)	0.476	0.512	0.519
BiomedCLIP fine-tune (Zhang et al., 2025)	0.492	0.491	0.491
QCE-Net (ours; medical reference)	**0.702**	**0.978**	**0.868**

Table 2: General-domain detectors on full test sets under the fair protocol; AUROC, hallucination as the positive class.

(MMedPO (Zhu et al., 2025), MMed-RAG (Xia et al., 2025), BiomedCLIP (Zhang et al., 2025)) under the same matched protocol. On the full test sets of VQA-RAD, SLAKE, and Path-VQA, all of them score between 0.48 and 0.57 AUROC, close to chance (Table 2). Among the evaluated fixed-similarity, sampling, and language-statistic baselines, none exceeds chance once length and type matching and image-level splits are enforced.

Yes/no pairs and the answer-channel bound. Under the fair protocol, a yes/no negative is the opposite-polarity answer paired with the same image and question, so $\mathbf{q}$ and $\mathbf{e}_q$ are identical across the matched pair; the only input that may differ is the answer embedding, which CLIP encodes nearly identically for "yes" and "no" (cosine 0.949). Equation 4 formalizes why this caps answer-only scores near chance, and the measured yes/no block is QCE-Net's weakest type (0.679 versus pooled open-ended 0.728). This ceiling applies to any answer-only score under the matched-pair construction: an answer-conditioned score alone is at chance on yes/no pairs (0.50), and even better-separated per-token projections (cosine 0.65) do not push the answer-only channel beyond ≈ 0.53. The residual lift of full QCE-Net (0.679 versus ≈ 0.53) comes from the fused evidence and the MLP. LLaVA-LoRA, which bypasses the CLIP text encoder, reaches 0.821 on the same yes/no block and 0.939 on open-ended pairs (Appendix Table 8), consistent with the model exceeding the answer-channel bound on this block.

4.3 Where Does the Signal Live? An Input Ablation

Before designing QCE-Net, we asked which inputs a trainable detector actually uses. Table 3 ablates the inputs of the MLP baseline: feeding only the question–answer pairing $(\mathbf{a}, \mathbf{q})$ nearly matches the full model on VQA-RAD (0.679 versus 0.670), while the answer alone is near chance (0.485) and the

Inputs to trainable MLP	VQA-RAD	Path-VQA
$[\mathbf{g}, \mathbf{a}, \mathbf{q}]$ (full)	0.670	0.867
$[\mathbf{a}, \mathbf{q}]$ (no image)	0.679	0.843
$[\mathbf{g}, \mathbf{a}]$ (no question)	0.640	0.728
$[\mathbf{a}]$ (answer only)	0.485	0.512

Table 3: Input ablation of the trainable MLP. The full-input row is the MLP baseline of Table 1; the remaining rows are independent runs under the identical protocol. The question–answer pairing carries most of the signal; the image adds little on VQA-RAD and less elsewhere, motivating question-conditioned evidence.

Variant (VQA-RAD, full test)	AUROC	Δ vs. full
Full $[\mathbf{g}, \mathbf{a}, \mathbf{q}, \mathbf{e}_q, \mathbf{e}_a]$	**0.7019**	—
w/o both evidence channels ($\mathbf{e}_q, \mathbf{e}_a$)	0.6702	-0.032
w/o global image ($\mathbf{g}$)	0.6981	-0.004
w/o question ($\mathbf{q}$)	0.6416	-0.060
w/o answer ($\mathbf{a}$)	0.5069	-0.195

Table 4: QCE-Net ablation on the full VQA-RAD test set. Removing the question or answer is far more damaging than removing the evidence channels, consistent with the input ablation (Section 4.3); the evidence channels lift QCE-Net above the MLP baseline on VQA-RAD (0.7019 versus 0.6702, $p = 0.006$).

global image alone is 0.50. On Path-VQA the same pattern holds but with a larger image residual (0.843 versus 0.867).

This ablation motivates the design of QCE-Net: current trainable detectors are largely learning question–answer statistics, not image grounding. The question-conditioned evidence channel ($\mathbf{e}_q$, Equation 3) is designed to add exactly this missing signal, and Table 1 shows that adding it yields accuracy comparable to or better than the MLP baseline while remaining interpretable. Under the fair protocol, the additional contribution of image evidence is small in aggregate, so QCE-Net's margin over the MLP is correspondingly small; it is concentrated precisely on the yes/no block, where Equation 4 predicts the answer channel to be capped, and its evidence map localizes the anatomical region the question asks about (Section 4.5).

Efficiency. QCE-Net is lightweight: it requires one CLIP image encode, two text encodes, and a small MLP forward per sample.

4.4 QCE-Net Channel Ablation

Tables 4 and 5 ablate QCE-Net's input channels on the full VQA-RAD and Path-VQA test sets. The ablation is capacity-matched: removing both evidence channels ($\mathbf{e}_q, \mathbf{e}_a$) leaves exactly the MLP's input set $[\mathbf{g}, \mathbf{a}, \mathbf{q}]$ and reproduces its AUROC (0.6702, Table 1), so the $+0.032$ gain on VQA-RAD is attributable to the evidence channels rather than to extra input dimensions. Removing the answer channel leads to the largest AUROC drop on both datasets (-0.195 on VQA-RAD, -0.367 on Path-VQA), followed by the question (-0.060 and -0.175). The evidence channels help on VQA-RAD, where the yes/no polarity pair caps the answer channel (Equation 4), and are neutral on Path-VQA, where the answer channel already encodes the grounding signal.

Where do the evidence channels act? On Path-VQA a by-question-type breakdown (Appendix Table 7) shows they help open-ended questions ($+0.004$ AUROC) and slightly hurt yes/no pairs (-0.001), for a net overall change of $+0.001$.

4.5 Fairness Validation and Limitations

We provide evidence that the protocol reduces leakage and that the surviving signal is not purely lexical (Table 6): a classifier using only lexical features scores near chance on VQA-RAD and Path-VQA (0.444/0.496); image-level splits guarantee no image straddles train and test; all features

Variant (Path-VQA, full test)	AUROC	Δ vs. full
Full $[\mathbf{g}, \mathbf{a}, \mathbf{q}, \mathbf{e}_q, \mathbf{e}_a]$	**0.8678**	—
w/o both evidence channels ($\mathbf{e}_q, \mathbf{e}_a$)	0.8666	-0.001
w/o global image ($\mathbf{g}$)	0.8690	$+0.001$
w/o question ($\mathbf{q}$)	0.6929	-0.175
w/o answer ($\mathbf{a}$)	0.5010	-0.367

Table 5: QCE-Net ablation on the full Path-VQA test set (matched protocol, $n = 1676$; $[\mathbf{g}, \mathbf{a}, \mathbf{q}]$ corresponds to the MLP baseline). Removing the evidence channels barely changes the AUROC (0.8678 to 0.8666, $\Delta = -0.001$), consistent with the statistically tied result in Table 1; removing the question or answer is far more damaging.

Control	VQA-RAD	Path-VQA	OmniMed
Language-only features	0.444	0.496	0.6939
Inverse orientation (best baseline)	0.494	0.495	0.499
QCE-Net (ours)	**0.7019**	**0.8678**	**0.9674**

Table 6: Fairness controls under the matched-negative protocol. Language-only features are near chance on VQA-RAD and Path-VQA; on OmniMed they reach 0.69 (a property of that benchmark's richer text), yet QCE-Net still adds a large margin over them. The strongest baseline's best orientation remains far below QCE-Net on every dataset.

are computed per (image, question, answer) triple from frozen encoders with no global statistics; and negatives are length- and type-matched answers from other images. Even where the text is most informative (OmniMed, 0.694), QCE-Net's margin over the language-only control is $+0.27$, so the surviving signal is not solely a text artifact.

Sampling a negative answer from another image assumes that an answer grounded in one image is not grounded in the current one. In medicine this can fail when a finding is invariant across the population, so a small fraction of our negatives may be valid for the test image. We provide a sensitivity analysis (Appendix Table 10): only 3.0%, 1.6%, and 1.5% of VQA-RAD, Path-VQA, and OmniMed negatives respectively exceed the 95th-percentile grounded-similarity threshold, and excluding them leaves every AUROC unchanged or slightly higher.

Performance can also depend on subset size (Appendix Figure 2): on 100 pairs VASE appears to beat our reference detector (0.669 versus 0.601) but decreases as the subset grows, while the reference score remains stable. These observations suggest that reporting full-set results, or confidence intervals together with subset sizes, is preferable for stable comparison. The by-question-type decomposition is consistent with the evidence channels contributing mainly to (image, question)-dependent labels, and the evidence map highlights the asked anatomical region on open-ended pairs while collapsing on matched yes/no negatives (Appendix Table 7).

5 Conclusion

Under a fair protocol, the evaluated medical hallucination detectors perform near chance (0.49–0.57 AUROC). These results suggest that evaluation artifacts, rather than detector capacity alone, may account for much of the previously reported performance; an input ablation shows the surviving signal is dominated by the question–answer pairing, while the additional contribution of image evidence is small under the evaluated protocol.

QCE-Net, an interpretable head over frozen CLIP, matches or beats the strong MLP, significantly on VQA-RAD (0.702 vs 0.670, $p = 0.006$) and tied elsewhere, and its per-patch weights double as evidence maps. LLaVA-LoRA, which trains the medical LMM into a per-sample grounded/ungrounded judge, clears AUROC 0.8 on VQA-RAD under a leakage-free protocol, with a significant gain on Path-VQA and a tie on OmniMed.

AI USE STATEMENT

Large language models (LLMs) were employed exclusively to improve the quality of written English and for stylistic refinement, with attentive supervision by the authors. No scientific content, experimental design, data analysis, or any intellectual contribution were generated through LLMs, nor did such assistance influence any reported results.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This work evaluates hallucination detectors on public benchmarks and collects no patient data. We use only publicly available datasets (VQA-RAD, Path-VQA, OmniMed-VQA-mini) and publicly released frozen models (CLIP-L/14@224, LLaVA-Med-1.5, MedGemma-4b), and we release no patient-identifiable information. The finding that some evaluated detectors perform near chance under a fair protocol is intended to encourage responsible evaluation of medical AI, not to discourage the deployment of well-validated systems.

REPRODUCIBILITY STATEMENT

We will release the implementation of our fair protocol, the full-scale VASE reproductions, the QCE-Net and baseline code, the input-ablation script, and the data-split definitions as supplementary material to enable exact reproduction. All experiments rely on public datasets (VQA-RAD, Path-VQA, OmniMed-VQA-mini) and publicly released frozen models (CLIP-L/14@224, LLaVA-Med-1.5, MedGemma-4b). Hyperparameters, training schedules, the paired-bootstrap evaluation, and the supporting ablations and sensitivity analyses are documented in the main text and appendix.

REFERENCES

Yung-Sung Chuang, Yujie Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL*, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Zishan Gu, Jiayuan Chen, Fenglin Liu, Changchang Yin, and Ping Zhang. MedVH: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Advanced Intelligent Systems*, 8(1), 2026. doi: 10.1002/aisy.202500255.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Volume 2: Short Papers*, 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025. doi: 10.1145/3703155.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

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 (NeurIPS)*, 2023a.

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 *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023b.

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

Zehui Liao, Shishuai Hu, Ke Zou, Huazhu Fu, Liangli Zhen, and Yong Xia. Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pp. 669–679. Springer, 2025. doi: 10.1007/978-3-032-04971-1_63.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, 2024.

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. In *Proceedings of the IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021.

Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Tao Tu, Shekoofeh Azizi, Danny Driess, and et al. Towards generalist biomedical AI. *NEJM AI*, 1(3): AIoa2300138, 2024. doi: 10.1056/aioa2300138.

Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. In *Tiny Papers at the 12th International Conference on Learning Representations (ICLR)*, 2024.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Peng Xia, Jinglu Wang, Yibo Peng, and et al. MMedAgent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Qinghao Ye, Haiyang Xu, Guohai Xu, and et al. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, pp. 818–833, 2014.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, and et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. doi: 10.1056/AIoa2400640.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

CONTENTS

A Additional Details **14**

A.1 Design Rationale and Assumptions 14

A.2 Subset-Size Bias Figure 14

A.3 By-Question-Type Decomposition on Path-VQA 14

A.4 Zero-Shot LLaVA-Med Judge 15

A.5 LLaVA-LoRA Ablation 15

A.6 Negative-Sample Validity Sensitivity 16

A.7 Reference Detector: ViEAD (Hand-Crafted, Training-Free) 16

A.8 Single-Scalar Baselines 17

A.9 Per-Question-Type Figure 17

A.10 Error Analysis and Efficiency 17

A.11 Additional Baselines (Sampled Sets) 18

A.12 General-Domain Hallucination Detectors 18

A.13 QCE-Net Evidence Maps 18

A.14 Implementation Details 19

A.15 VASE Reproduction: Protocol and Fidelity Validation 20

A.16 Derivation of the Answer-Channel Bound 20

A.17 Derivation of the Evidence-Amplification Jacobian 20

A Additional Details

A.1 Design Rationale and Assumptions

Why question-conditioned evidence. The input ablation (Section 4.3) shows that a trainable detector's signal is dominated by the question–answer pairing. QCE-Net adds a channel that this pairing cannot provide, whether the image contains what the question asks about, computed as the attention-pooled patch evidence $\mathbf{e}_q = \sum_i w_i \mathbf{p}_i$ with $w_i \propto e^{\kappa \langle \mathbf{p}_i, \mathbf{q} \rangle}$. Because $\mathbf{e}_q$ is a function of the image and the question alone, it is invariant to the answer; this is intentional, it measures visual support for the question, which complements the answer-conditioned alignment that ViEAD-style detectors use. The temperature $\kappa = 2$ is fixed to keep the pooling interpretable; learned query projections did not improve results in our experiments.

On the input-ablation numbers. The input ablation (Table 3) uses a plain MLP identical to the MLP baseline, so the column differences isolate the contribution of each input channel; the full-input row is the MLP baseline of Table 1 itself, while the remaining rows are independent training runs under the identical protocol. VQA-RAD's 0.679 $(\mathbf{a}, \mathbf{q})$ versus 0.670 (full) confirms the image adds almost nothing to a question–answer-only model there; Path-VQA (0.843 versus 0.867) shows a larger image residual, consistent with that benchmark having more open-ended answers. The yes/no ceiling discussion is given in the main text (Section 4).

A.2 Subset-Size Bias Figure

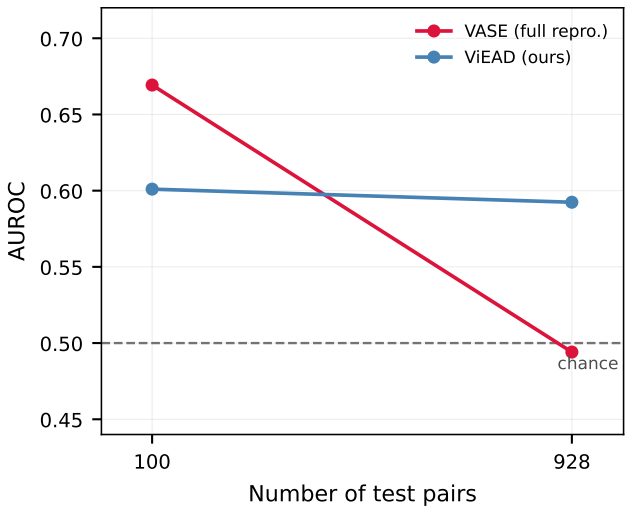


Figure 2: AUROC as a function of evaluation subset size on VQA-RAD (measured at 100 and 928 pairs). VASE's small-sample advantage does not persist on the full set.

A.3 By-Question-Type Decomposition on Path-VQA

Table 7 decomposes the Path-VQA result by question type to explain where the evidence channels help. Path-VQA's test set is split into yes/no ($n = 820$) and open-ended ($n = 856$) pairs. The evidence channels $(\mathbf{e}_q, \mathbf{e}_a)$ contribute $+0.004$ AUROC on open-ended questions (0.8186 versus 0.8145 without them) but -0.001 on yes/no pairs (0.9070 versus 0.9079). The reason is structural: on Path-VQA a yes/no negative is the opposite-polarity answer paired with the same image and question, so the question-conditioned evidence $\mathbf{e}_q$ is identical across the pair and cannot separate it; the answer-conditioned evidence $\mathbf{e}_a$ then only re-encodes the near-identical "yes"/"no" embeddings (cosine 0.949), adding noise. On open-ended questions, by contrast, the matched negative is a concrete finding from another image, so the question-conditioned evidence carries genuine grounding signal that the answer channel alone misses. This is the same mechanism that makes the evidence

channels decisive on VQA-RAD (Section A.10): the channels help exactly where the label is (image, question)-dependent rather than a polarity flip.

Variant (Path-VQA, full test)	All	Yes/No	Open-ended
Full $[\mathbf{g}, \mathbf{a}, \mathbf{q}, \mathbf{e}_q, \mathbf{e}_a]$	**0.8678**	0.9070	**0.8186**
w/o both evidence channels	0.8666	**0.9079**	0.8145
Δ (evidence channels)	$+0.001$	-0.001	$+\mathbf{0.004}$

Table 7: Path-VQA AUROC by question type (yes/no $n = 820$, open-ended $n = 856$). The evidence channels help open-ended questions ($+0.004$) and slightly hurt yes/no pairs (-0.001): on yes/no pairs the question-conditioned evidence is identical across the matched pair and the answer-conditioned evidence re-encodes the near-identical "yes"/"no" embeddings, whereas on open-ended questions the matched negative is a concrete finding from another image and the question-conditioned evidence carries genuine grounding signal.

A.4 Zero-Shot LLaVA-Med Judge

Table 8 reports the zero-shot LLaVA-Med judge, the frozen 7.5B medical LMM prompted with "Is this answer grounded in the image?" and scored by the GROUNDED–UNGROUNDED token-logit difference (no training on the protocol). It is near chance on every benchmark (0.496–0.572 AUROC), suggesting that the difficulty of the fair protocol is not specific to CLIP-based detectors: the same backbone that reaches 0.885–0.969 once LoRA-trained on the protocol does not separate grounded from hallucinated answers when frozen. The comparison with LLaVA-LoRA therefore isolates the value of training on the protocol, not the choice of backbone. The by-type decomposition of LLaVA-LoRA is computed from its saved per-sample test scores (the code will be released upon publication).

	VQA-RAD					
Detector	all	yes/no	open	Path-VQA	OmniMed	SLAKE
Zero-shot LLaVA-Med judge	0.5547	—	—	0.4961	0.5717	—
LLaVA-LoRA (trained)	**0.8854**	**0.821**	**0.939**	**0.8937**	**0.9691**	—

Table 8: Zero-shot LLaVA-Med judge under the fair protocol (frozen backbone, no training), versus the same backbone LoRA-trained (LLaVA-LoRA). The frozen judge is near chance on all benchmarks, so LLaVA-LoRA's gains come from training on the protocol. The by-type columns decompose VQA-RAD ($n = 928$; yes/no $n = 500$, open-ended $n = 428$): LLaVA-LoRA reaches 0.821 on the yes/no block, where the frozen-CLIP methods evaluated here are capped near 0.7 or below (QCE-Net 0.679, Section A.10), consistent with the LoRA-trained detector exceeding the polarity bound rather than gaining only on open-ended pairs.

A.5 LLaVA-LoRA Ablation

The main LLaVA-LoRA detector uses rank-8 LoRA on the query and value projections with prompt P0 ("Is this answer grounded in the image?"). We ablate (a) the LoRA rank $r \in \{4, 8, 16\}$, (b) the set of adapted modules (q+v only versus all linear projections and MLPs), and (c) the prompt wording (P0 versus P1 "Does the image support this answer?" versus P2 "Answer grounded or ungrounded:"), all on VQA-RAD with the same validation-based selection and early stopping as the main detector. Table 9 reports each configuration's best-validation AUROC and the corresponding one-shot test AUROC (model selection and early stopping use only the validation split; the test set is used for final reporting only).

Config	LoRA rank	Target modules	Prompt	Test AUROC
r4 (low rank)	4	q,v	P0	0.8809
r8 (main)	8	q,v	P0	0.8852
r16 (high rank)	16	q,v	P0	0.8836
allmods	8	all linear+MLP	P0	0.8839
prompt2 (P1)	8	q,v	P1	0.8826
prompt3 (P2)	8	q,v	P2	0.8760

Table 9: LLaVA-LoRA ablation on VQA-RAD (test AUROC of the epoch selected by validation AUROC; test used once for reporting). Rank 8 on q+v with P0 is the main configuration (0.8852, matching the 0.8854 main result within seed noise). Results are flat across rank (0.880–0.885) and adapting all linear+MLP modules gives no gain over q+v alone (0.8839 versus 0.8852), confirming the low-rank, target-specific design is sufficient. Prompt P0 ("Is this answer grounded in the image?") is the best wording; rephrasing to P1 (0.8826) is comparable, while the telegraphic P2 (0.8760) slightly hurts, indicating that a natural-language grounding prompt helps the medical LMM align its judgment.

A.6 Negative-Sample Validity Sensitivity

The fair protocol samples each negative answer from another image, assuming an answer grounded in one image is not grounded in the current one. In medicine this can fail for population-invariant findings (e.g., "no pulmonary nodule"). We provide a sensitivity analysis in Table 10: using the frozen-CLIP global image–answer cosine similarity as a proxy, a negative is flagged "possibly valid" if its similarity exceeds the 95th percentile of grounded (positive) pairs. Only 3.0% of VQA-RAD negatives, 1.6% of Path-VQA negatives, and 1.5% of OmniMed negatives are flagged, and re-running both detectors after excluding them leaves every AUROC unchanged or slightly higher. The negative-sample assumption is therefore both quantitatively small and conservative, it cannot inflate the near-chance result or the trained-detector margins.

	VQA-RAD	Path-VQA	OmniMed
Test pairs	928	1676	802
Negatives	464	838	401
Possibly-valid negatives	14 (3.0%)	13 (1.6%)	6 (1.5%)
MLP AUROC (all / excluding)	0.6702 / 0.6779	0.8666 / 0.8687	0.9697 / 0.9700
QCE-Net AUROC (all / excluding)	0.7019 / 0.7069	0.8678 / 0.8714	0.9674 / 0.9675

Table 10: Negative-sample validity sensitivity. "Possibly valid" = negative whose frozen-CLIP image–answer similarity exceeds the 95th percentile of grounded pairs. The fraction is small (3.0%/1.6%/1.5%) and excluding these negatives leaves every AUROC unchanged or slightly higher, so the matched-negative assumption is conservative.

A.7 Reference Detector: ViEAD (Hand-Crafted, Training-Free)

For completeness we include ViEAD, a hand-crafted, training-free detector described in Section 2, as a reference. It extracts 48 features spanning patch-level evidence localization, perturbation sensitivity, cross-granularity consistency, token evidence, and a language-prior control, and classifies them with XGBoost (300 trees, depth 5); "training-free" thus means that no vision or language model is trained, the XGBoost head itself being a lightweight classifier fitted on the fair training split. Its per-component ablation on VQA-RAD is:

Variant (VQA-RAD, full test)	AUROC	Δ vs. full
Full (48-dim, all pillars)	**0.5924**	—
w/o language prior (PPL)	0.5655	−0.027
w/o perturbation sensitivity (P2)	0.5649	−0.028
w/o evidence localization (P1)	0.5571	**−0.035**
w/o token evidence (P4)	0.5645	−0.028
language-only (P5 only)	0.4482	−0.144

Table 11: ViEAD reference ablation on VQA-RAD (928 pairs).

A.8 Single-Scalar Baselines

Single-scalar detector	VQA-RAD	Path-VQA	OmniMed
Max patch–answer alignment	0.5546	0.5168	0.6743
Evidence concentration	0.5484	0.5313	0.7261
Global CLIP image–answer sim.	0.5303	0.5226	0.7736
Question–answer similarity	0.5715	0.5183	0.6032
Best single scalar	0.5715	0.5313	0.7736
QCE-Net	**0.7019**	**0.8678**	**0.9674**

Table 12: Threshold-only single-scalar baselines on the identical full test sets (no training, frozen CLIP). Best orientation reported. QCE-Net far exceeds every single statistic, confirming the learned fusion and question-conditioned evidence add signal beyond any one alignment statistic.

A.9 Per-Question-Type Figure

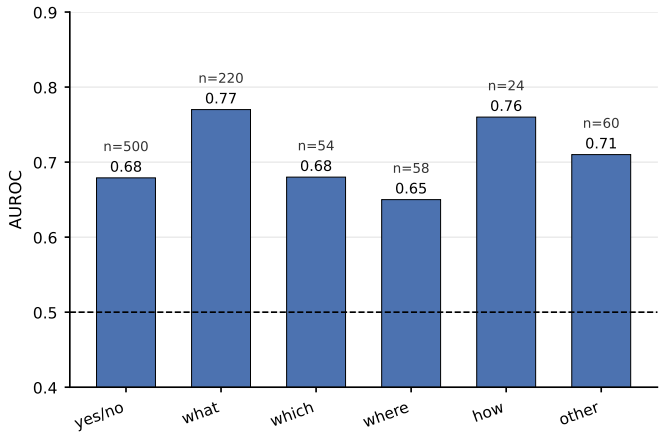


Figure 3: VQA-RAD AUROC by question type (QCE-Net, full test set, $n = 928$). AUROC is uniformly moderately above chance (0.65–0.77): the largest block, yes/no ($n = 500$, 54% of pairs), reaches 0.679 and is among the lowest, but open-ended types are only modestly higher (pooled open-ended 0.728), so VQA-RAD is hard across the board rather than failing only on yes/no. The 8 "why" questions are excluded ($n < 10$). Dashed line: chance (0.5).

A.10 Error Analysis and Efficiency

Error statistics. At the 0.5 threshold on the full VQA-RAD test set, QCE-Net produces false-negative and false-positive rates of 46% and 48% among yes/no pairs and 38% on each side among open-ended pairs. The yes/no subset contributes the majority of errors from 54% of the pairs. XGBoost feature-importance and error statistics for the hand-crafted ViEAD reference will be included in the code released upon publication.

A.11 Additional Baselines (Sampled Sets)

Methods below run on sampled subsets (their reported sample sizes); they are not directly comparable to the full-set numbers in Table 1 and are included only for completeness. The primary comparison for the trainable methods is the full-set MLP baseline in Table 1; the subset numbers are reported for completeness only.

Method	Backbone	VQA-RAD	Path-VQA
IC	gemma3-4b	0.5667 (100)	0.5827 (120)
HaLT-SWAP	gemma3-4b	0.5571 (100)	0.6070 (120)
GLSim	CLIP	0.5293 (928)	0.5424 (1000)
SE	gemma3-4b	0.5063 (100)	0.4997 (120)
VASE (sim.)	gemma3-4b	0.5086 (100)	0.4998 (100)
IC	LLaVA-Med	0.5665 (100)	0.6226 (100)
GLSim	LLaVA-Med	0.4235 (200)	0.5771 (200)
QCE-Net (full set)	CLIP+MLP	**0.7019** (928)	**0.8678** (1676)
LLaVA-LoRA (full set)	LLaVA-Med LoRA	**0.8854** (928)	**0.8937** (1676)

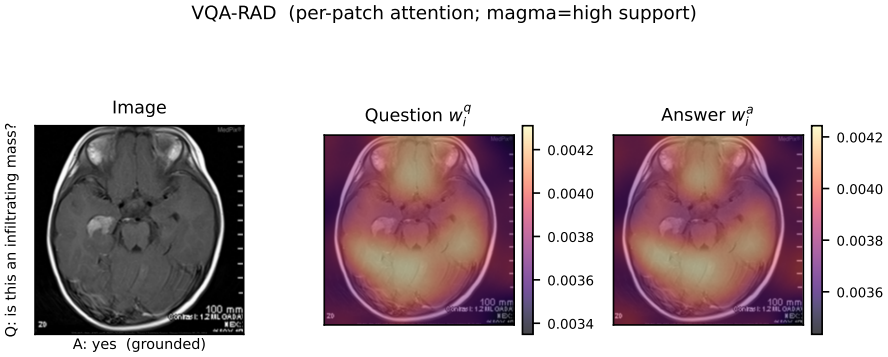
Table 13: Additional baselines on sampled subsets (sample sizes in parentheses); the last two rows, below the rule, are the full-set reference numbers from Table 1 and are not sampled.

A.12 General-Domain Hallucination Detectors

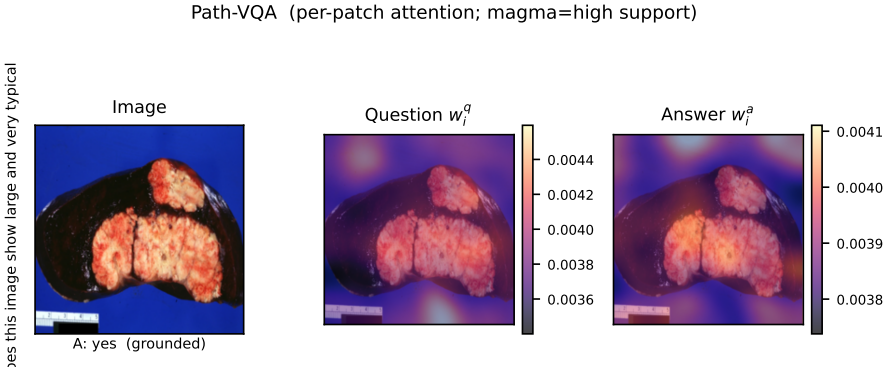
To examine whether the finding generalizes beyond medical models, the general-domain comparison in the main text (Table 2) runs general-domain hallucination detectors under the same matched-negative protocol: CLIPScore (Hessel et al., 2021) (image–answer cosine similarity), CHAIR-Medical (a medical-adapted CHAIR (Rohrbach et al., 2018) hallucination rate), RULE (question–answer logit statistics), and lightweight fine-tunes on CLIP-style features for MMedPO (Zhu et al., 2025), MMed-RAG (Xia et al., 2025), and BiomedCLIP (Zhang et al., 2025). Every evaluated method is near chance, like the medical-specific VASE family, suggesting the finding is not specific to a single method: no fixed similarity or language statistic exceeds chance once length and type matching and image-level splits are enforced. Full-generation detection or correction methods such as LURE (Zhou et al., 2024) and Woodpecker (Yin et al., 2024) operate at a different granularity, they flag or repair hallucinations inside open-ended generated answers, so they are not directly applicable to scoring a fixed candidate answer against a matched (image, question) pair.

A.13 QCE-Net Evidence Maps

The per-patch attention weights w_i of Equation 3 serve as evidence maps for QCE-Net. Figure 4 visualizes them on grounded samples from VQA-RAD and Path-VQA: for each sample we show the input image, the question-conditioned map w_i^q, and the answer-conditioned map w_i^a. On supported answers the map concentrates on a region consistent with the claim (a nodule or opacity on the radiograph, a cell cluster on the stained slide); on unsupported answers it disperses across the image.



(a) QCE-Net attention maps on VQA-RAD.



(b) QCE-Net attention maps on Path-VQA.

Figure 4: Per-patch attention weights w_i (Equation 3) from frozen QCE-Net overlaid on the input image for grounded ($y = 0$) samples. Left-to-right within each row: image, question-conditioned map w_i^q, answer-conditioned map w_i^a. Attention localizes the region named by the question and the evidence the answer binds to. Magma colormap, percentile-normalized.

A.14 Implementation Details

QCE-Net uses frozen CLIP-L/14@224: $N = 256$ normalized patch embeddings, L2-normalized text embeddings for the answer and question, and the global (CLS) image embedding. The evidence pooling (Equation 3) uses temperature $\kappa = 2$, selected by validation AUROC over the sweep $\kappa \in \{2, 3, 4, 5, 6, 8\}$ and competitive with a learned query projection (Section 4.4). The MLP head has three hidden layers (1024, 512, 256; BatchNorm, ReLU, dropout $0.4/0.4/0.3$) and is trained with binary cross-entropy, AdamW (10^{-3}, weight decay 10^{-4}), batch size 256, and a cosine schedule over 120 epochs. All trainable methods are trained on the training split only; test pairs are identical across all methods. QCE-Net training operates only on frozen CLIP embeddings and a small MLP head; at inference it requires one CLIP image encode, two text encodes, and a small MLP forward per sample.

LLaVA-LoRA. The trainable medical-LMM detector builds on LLaVA-Med-1.5 (Mistral-7B backbone; 7.5B parameters). We attach rank-8 LoRA adapters to the query and value projections (`q_proj`, `v_proj`; $\alpha = 16$, dropout 0.1, about 4.2M trainable parameters, 0.055% of the model), and fine-tune so that given the prompt "Is this answer grounded in the image?" the model must generate `grounded` (label 0) or `ungrounded` (label 1) as the answer token. The per-sample score is the difference of the two token logits at the final position. Training uses AdamW (3×10^{-5}, weight decay 0.01), batch size 4 (per-sample updates), gradient clipping at norm 1, a cosine schedule over 5–10 epochs, and up to 4,000–6,000 training pairs. Model selection and early stopping use only the validation split (which shares no images with the test set): we retain the checkpoint with the best validation-set AUROC, and a run is stopped early when the validation AUROC declines by more than

0.004 after epoch 2. The test set is evaluated exactly once, on the retained checkpoint, to obtain the reported number; it is never used for scheduling. On every dataset this detector uses the same fair training split and the same test pairs as the MLP and QCE-Net, so the paired-bootstrap comparisons in Table 1 are exact.

A.15 VASE Reproduction: Protocol and Fidelity Validation

The VASE reproduction follows the published algorithm Liao et al. (2025): greedy decoding for the main answer, 10 sampled answers at temperature 1.0, semantic clustering via the backbone tokenizer and a sentence-embedding model, RadFlag as the dominant-cluster ratio, SE as token-level entropy, and the noise-shifted distribution with the paper's α for the visual score. Both backbones (LLaVA-Med-1.5, MedGemma-4b) run at full scale; generation uses temperature 1.0 with top-p 0.9 and a 60-token budget.

Implementation choices. (a) The 100-sample check uses a fixed, seeded subsample (random seed 42) of the same VQA-RAD test pairs used for the full-set evaluation. (b) The original VASE evaluation protocol does not enforce length- and type-matched negatives; our reproduction is run under the fair protocol (Section 3.1) so that the comparison is consistent across methods. The 100-sample number (0.6693) falls within the originally reported range, whereas the full 928-pair numbers are near chance under the fair protocol; the subset-size sensitivity is documented in Section 4 and Appendix A.2.

VASE reproduction	VQA-RAD	Path-VQA
100-sample subset (seed 42)	0.6693	—
Full set (LLaVA-Med-1.5)	0.4941	0.4950
Full set (MedGemma-4b)	0.4991	0.5029

Table 14: VASE reproduction across subset and full scales. The 100-sample number reproduces the original range; the full-set numbers are near chance under the fair protocol.

We implement the published VASE algorithm from the paper and report full-scale numbers under both backbones so that implementation differences are visible as a range.

A.16 Derivation of the Answer-Channel Bound

For the answer-channel bound in Equation 4, the matched negative (I, Q, A') shares the image and question with the positive (I, Q, A) under the fair protocol, so the only input that changes is the answer. Since s_A depends on the answer only through $h(\phi(\cdot))$ and h is L-Lipschitz, $|s_A(I,Q,A) - s_A(I,Q,A')| = |h(\phi(A)) - h(\phi(A'))| \le L\,\|\phi(A) - \phi(A')\|$, which is Equation 4. For a yes/no pair $A \in \{\text{yes}, \text{no}\}$ and A' its opposite polarity, the L2-normalized CLIP encodings satisfy $\cos(\phi(\text{yes}), \phi(\text{no})) = 0.949$ (measured on the frozen encoder), so $\|\phi(\text{yes}) - \phi(\text{no})\|^2 = 2(1 - 0.949) = 0.102$ and $\|\phi(\text{yes}) - \phi(\text{no})\| \approx 0.32$. Any answer-only score therefore separates the pair by at most $0.32L$ in score space, regardless of the classifier applied to $\phi(A)$.

A.17 Derivation of the Evidence-Amplification Jacobian

Write $w_i = e^{\kappa v_i}/Z$ with $v_i = \langle \mathbf{p}_i, \mathbf{a} \rangle$ and $Z = \sum_j e^{\kappa v_j}$. The softmax Jacobian is $\partial w_i / \partial \mathbf{a} = \kappa\, w_i(\mathbf{p}_i - \mathbf{e}_a)$, and since $\mathbf{e}_a = \sum_i w_i \mathbf{p}_i$,

$$\frac{\partial \mathbf{e}_a}{\partial \mathbf{a}} = \sum_i \mathbf{p}_i \Big(\frac{\partial w_i}{\partial \mathbf{a}}\Big)^\top = \kappa \sum_i w_i\, \mathbf{p}_i\, (\mathbf{p}_i - \mathbf{e}_a)^\top = \kappa \sum_i w_i\, (\mathbf{p}_i - \mathbf{e}_a)(\mathbf{p}_i - \mathbf{e}_a)^\top,$$

where the last equality uses $\sum_i w_i(\mathbf{p}_i - \mathbf{e}_a) = \mathbf{0}$. The matrix $C = \sum_i w_i(\mathbf{p}_i - \mathbf{e}_a)(\mathbf{p}_i - \mathbf{e}_a)^\top$ is a weighted covariance of unit vectors and is positive semidefinite, so $\|\partial \mathbf{e}_a / \partial \mathbf{a}\| = \kappa\, \lambda_{\max}(C)$. Its largest eigenvalue is the maximal directional variance $\max_{\|\mathbf{v}\|=1} \sum_i w_i \langle \mathbf{p}_i - \mathbf{e}_a, \mathbf{v} \rangle^2$, which is not bounded by $1/\kappa$ in general: e.g., if the answer binds to two near-antipodal patches ($\mathbf{p}_1 \approx -\mathbf{p}_2$, $w_1 = w_2 = \frac{1}{2}$), then $C \approx \mathbf{p}_1 \mathbf{p}_1^\top$ and $\|\partial \mathbf{e}_a / \partial \mathbf{a}\| \approx \kappa = 2 > 1$ for $\kappa = 2$. Thus the evidence channel can amplify answer differences that the raw embedding compresses. Finally, $\mathbf{e}_q = \sum_i w_i^q \mathbf{p}_i$ with

$w_i^q \propto e^{\kappa\langle \mathbf{P}_i, \mathbf{q}\rangle}$ depends on $(\mathbf{P}, \mathbf{q})$ only, so it is identical across the matched pair and immune to the answer-channel collapse.