# MedObvious: Exposing the Medical Moravec's Paradox in VLMs via Clinical Triage

Ufaq Khan*[1], Umair Nawaz[1], Lekkala Sai Teja[2], Numan Saeed[1], Muhammad Bilal[3], Yutong Xie[1], Mohammad Yaqub[1], and Muhammad Haris Khan[1]

[1] Mohamed bin Zayed University of Artificial Intelligence, UAE
[2] National Institute of Technology, Silchar
[3] Birmingham City University, UK
✉ ufaq.khan@mbzuai.ac.ae

Project Page: MedObvious-Website, Github: MedObvious

**Abstract.** Vision Language Models (VLMs) are increasingly used for tasks like medical report generation and visual question answering. However, fluent diagnostic text does not guarantee safe visual understanding. In clinical practice, interpretation begins with *pre-diagnostic sanity checks*: verifying that the input is valid to read (correct modality and anatomy, plausible viewpoint and orientation, and no obvious integrity violations). Existing benchmarks largely assume this step is solved, and therefore miss a critical failure mode: a model can produce plausible narratives even when the input is inconsistent or invalid. We introduce **MedObvious**, a 1,880-task benchmark that isolates input validation as a set-level consistency capability over small multi-panel image sets: the model must identify whether any panel violates expected coherence. MedObvious spans five progressive tiers, from basic orientation/modality mismatches to clinically motivated anatomy/viewpoint verification and triage-style cues, and includes five evaluation formats to test robustness across interfaces. Evaluating 17 different VLMs, we find that sanity checking remains unreliable: several models hallucinate anomalies on normal (negative-control) inputs, performance degrades when scaling to larger image sets, and measured accuracy varies substantially between multiple-choice and open-ended settings. These results show that pre-diagnostic verification remains unsolved for medical VLMs and should be treated as a distinct, safety-critical capability before deployment.

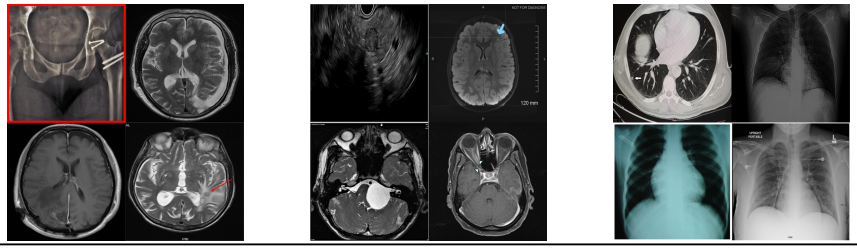
**Keywords:** Medical Vision-Language Models · Pre-diagnostic Sanity Checking · Clinical Triage · Visual Grounding · AI Safety

## 1 Introduction

Vision-Language Models (VLMs) are increasingly being used to interpret medical images. Recent systems can generate radiology-style descriptions, answer clinical

* Corresponding author

(a) Visual Referring (b) Detection MCQ (c) Detection MCQ



| | (a) | (b) | (c) |
|---|---|---|---|
| **Question** | Is this highlighted scan the clinical outlier; i.e., does it differ in modality, anatomy, or pathology from the others? Answer yes/no | You are reviewing a 2 × 2 grid of medical scans. One scan is the clinical outlier as it differs in modality, anatomy, or pathology. Which position contains the outlier? *A)* top-left *B)* top-right *C)* bottom-left *D)* bottom-right *E)* None | In this 2 × 2 medical image grid, one scan contains a clinical anomaly (chest CT). Which position is it? *A)* top-left *B)* top-right *C)* bottom-left *D)* bottom-right *E)* None |
| **Ground truth** | Yes | A (Top-left) | A (Top-left) |
| **LLaVA-1.5** | No (✗) | B (✗) | B (✗) |
| **Qwen3-VL** | Yes (✓) | D (✗) | D (✗) |
| **Lingshu** | Yes (✓) | B (✗) | A (✓) |

Fig. 1: **Qualitative MedObvious examples for pre-diagnostic visual triage.** Each column corresponds to a grid in (a)–(c). We report the task question, ground truth, and predictions from representative VLMs.

questions, and perform multi-step reasoning over images and text, driven by both general-purpose models such as GPT-4o [1], Flamingo [4] and LLaVA [20] and medical adaptations such as LLaVA-Med [15], RadFM [26], and others [22,13,21]. In parallel, these models are being explored as the core perception for *visual AI agents* [16,10] that can also interact with imaging software (e.g., navigating viewers, selecting series, adjusting visualization, and triggering downstream tools). This progress has motivated their potential use as assistants for clinical reporting and decision support. However, fluent language generation does not guarantee reliable visual perception. VLMs may produce coherent diagnostic narratives while failing basic sanity checks, such as detecting incorrect orientation, mismatched anatomy, unexpected modality, or physically implausible artifacts. We refer to this mismatch as the **Medical Moravec's Paradox**, extending Moravec's observation [2] that perception and spatial reasoning, trivial for humans, can be disproportionately difficult for machines even when higher-level outputs appear plausible. In medical imaging, this gap is consequential because failures occur before diagnosis: when the input is invalid or inconsistent, downstream reports become clinically uninterpretable.

Clinical interpretation begins with pre-diagnostic triage: clinicians first verify body part, view, modality, laterality, orientation, and basic image integrity, and they do not proceed to diagnosis if these checks fail. This requirement is amplified in *multi-image* and *AI-agentic* settings, where decisions depend on

consistency across a set of inputs, such as multiple fetal ultrasound views or long CT/MRI slice sequences and series. A single mis-acquired view, corrupted slice, mismatched series, or orientation error can compromise study-level reasoning, especially for models that aggregate evidence across images. Moreover, common tools such as 3D Slicer [11] and ITK-SNAP [28] support multi-panel layouts (e.g., axial, sagittal, coronal, and 3D) similar to a $2 \times 2$ display. VLM-based agents operating in these viewers must detect obvious panel-level inconsistencies to avoid acting on the wrong series, anatomy, or invalid inputs. Despite their importance, pre-diagnostic competencies are rarely evaluated explicitly. Standard medical VLM benchmarks such as VQA-RAD [14], PathVQA [12], PMC-VQA [29], VQA-Med [6], and SLAKE [17] primarily assess the correctness of final answers, while hallucination-focused benchmarks such as Med-HallMark [7] emphasize factual consistency of the generated text. These settings largely assume the input has been correctly perceived and can therefore miss failures on visually obvious sanity checks, allowing models to score well while remaining brittle and potentially unsafe in real multi-image or agentic workflows.

We introduce **MedObvious**, a benchmark for pre-diagnostic visual sanity checking in medical images. MedObvious asks a question that precedes diagnosis: *Is the input coherent and appropriate to interpret?* For this, we present small grids ($2 \times 2$ or $3 \times 3$) and require models to either identify an outlier panel or state that no outlier exists, as shown in Fig. 1. Although it is not a clinical reading interface, this provides a controlled probe of set-level consistency, mirroring requirements in multi-view ultrasound, multi-slice CT/MRI, and multi-panel viewer agentic workflows. MedObvious evaluates two axes. The *Clinical Safety* axis targets real failure modes that should be caught before any diagnostic verdict, including wrong body parts, flipped images, viewpoint/anatomy mismatches, and grossly apparent major pathology. The *Visual Grounding* axis uses synthetic inconsistencies (e.g., inserting modality-incompatible textures into an image) to test whether models check the visual input itself rather than relying on language priors for downstream tasks. MedObvious also includes explicit **negative controls** where all panels are consistent, so the correct response is that no outlier exists, directly measuring false alarms. Furthermore, it is organized into five progressive tiers (T1–T5) that increase in difficulty and clinical specificity as depicted in Fig. 2, ranging from basic orientation/modality mismatches to anatomy/viewpoint inconsistencies and high-saliency triage failures. In our zero-shot evaluation across *7 general*, *4 medical*, and *6 proprietary* VLMs, performance remains uneven as the best mean accuracy reaches 63.2%, yet negative-control accuracy spans a wide range, indicating that false alarms on normal inputs remain common. We also observe strong format sensitivity, with large gaps between multiple-choice and open-ended variants of the same underlying capability. Our main contributions are:

- We formalize the *Medical Moravec's Paradox* for medical VLMs, highlighting a gap between fluent diagnostic language and reliable pre-diagnostic visual sanity-checking, especially in multi-image and agentic-viewer settings.

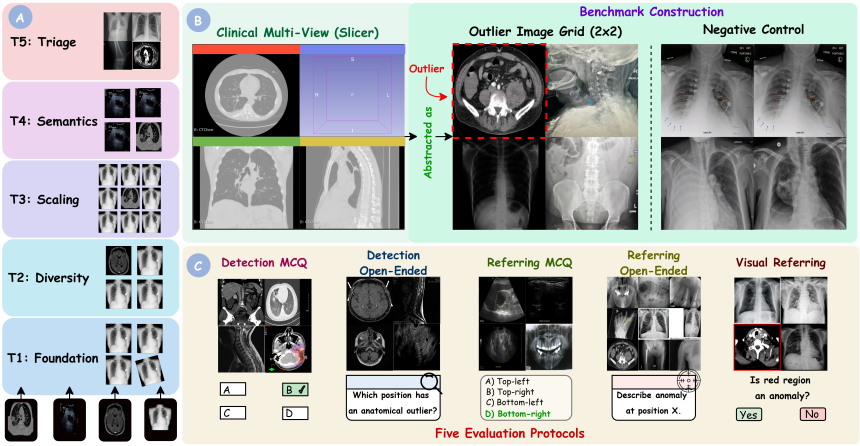



Fig. 2: **MedObvious overview.** (**A**) Five progressive tiers (T1–T5). (**B**) Construction: multi-view studies are abstracted into small grids with either a single outlier or a negative-control (no outlier). (**C**) Five evaluation protocols: Detection (MCQ/Open), Referring (MCQ/Open), and Visual Referring (Yes/No).

- We present first **MedObvious**, a 1,880 task benchmark spanning 5 progressive tiers, multiple grid configurations, five evaluation modes, and systematic negative controls, designed to evaluate pre-diagnostic visual triage independently of diagnosis.
- We benchmark 7 representative open-source, 6 closed-source, and 4 medically specialized VLMs under zero-shot inference.

## 2 MedObvious Construction

MedObvious is a benchmark designed to test whether medical VLMs can recognize *obvious* input-level inconsistencies before attempting any diagnosis, as inspired by [9]. It acts as a prerequisite for safe medical VLM deployment by *pre-diagnostic visual sanity checking*. Before interpreting pathology, clinicians first verify that an image is valid to read, including modality, anatomical region, viewpoint/orientation, and basic integrity. If these checks fail, the subsequent diagnosis is unreliable regardless of how fluent the generated text appears. MedObvious isolates this input-validation step and explicitly evaluates it.

**Motivation.** The clinical need is inherently *set-based*. Many studies are interpreted across multiple views, slices, or series, where a single inconsistent element can compromise the conclusions of the study. In fetal ultrasound, assessment of the fetal heart relies on multiple standard views. One mis-acquired plane or corrupted view may change the interpretation. In CT/MRI, clinicians reason over long slice sequences and multi-series studies. In longitudinal assessment, such

as reviewing the progression of Multiple Sclerosis from two or more brain MRI scans, an incorrectly positioned slice could lead to a completely different clinical decision. A flipped series, corrupted slice, or modality/anatomy mismatch is a coherence break that should be detected prior to diagnosis. This set-level checking is also reflected in common imaging software (e.g., 3D Slicer and ITK-SNAP), which presents multi-planar views in multi-panel layouts resembling $2 \times 2$ grid. Our grid-based tasks are therefore a controlled abstraction of this real requirement by detecting whether any element in a small visual set violates expected consistency before proceeding to downstream reasoning or agentic actions.

**Problem Formulation.** Each MedObvious instance presents a grid of $n$ images, $G = \{I_1, \ldots, I_n\}$ with $n \in \{4, 9\}$. The model must either identify the index of the inconsistent image or state that no outlier exists ($y \in \{1, \ldots, n\} \cup \{\varnothing\}$) where $y = k$ denotes $I_k$ as the outlier and $y = \varnothing$ denotes a clean and consistent grid. This setup evaluates input validity and set-level coherence, i.e., whether images that should belong to the same study context (across views, slices/series, or timepoints) are mutually consistent, rather than evaluating diagnosis.

Table 1: **MedObvious composition.** Tiers increase in clinical specificity.

| Tier | Name | Grid | #Mod. | Tasks | Pos. | Neg. | Clinical analog |
|---|---|---|---|---|---|---|---|
| T1 | Foundation | $2 \times 2$ | 2 | 440 | 275 | 165 | Basic QC (orientation, modality) |
| T2 | Diversity | $2 \times 2$ | 2 | 480 | 300 | 180 | Robustness across sources |
| T3 | Scaling | $3 \times 3$ | 4 | 360 | 225 | 135 | Multi-view/-slice comparison |
| T4 | Semantics | $2 \times 2$ | 3 | 320 | 200 | 120 | Anatomy/viewpoint verification |
| T5 | Triage | Mixed | 4 | 280 | 175 | 105 | "Stop and verify" safety cues |
| **Total** | | | **4** | **1,880** | **1,175** | **705** | |

**Datasets.** We primarily use curated subsets of ROCO [23] to construct anatomically and modality-defined tasks (e.g., chest radiographs, CT, MRI, ultrasound) using metadata filtering. To reduce modality shortcuts and enforce modality awareness, we additionally include non-radiological images, including endoscopy from Kvasir [24], to increase visual diversity and discourage reliance on generic grayscale radiology priors.

**Template-based generation.** Each task is generated from a shared template: we sample $n - 1$ *inlier* panels from a reference category (defined by modality, and when available, anatomy/viewpoint), and then either (i) insert one *outlier* from a different category or via a controlled integrity violation (e.g., orientation change or physically inconsistent composite), or (ii) create a *negative control* by sampling all $n$ panels from same reference category so the correct answer is $\varnothing$.

**Progressive tiers.** MedObvious comprises five tiers (1,880 tasks) with increasing clinical specificity (Table 1): T1 Foundation ($2 \times 2$; orientation/modality mismatches), T2 Diversity ($2 \times 2$; finer distinctions across sources), T3 Scaling ($3 \times 3$; many distractors and modality diversity), T4 Semantics ($2 \times 2$; anatomy/viewpoint mismatches and integrity violations), and T5 Triage; high-saliency failures (like gross abnormalities), and cross-modality coherence breaks.

**Evaluation Protocols.** Each grid is evaluated in 5 formats to separate visual capability from response-format effects: Detection MCQ (*pick the outlier*),

Table 2: **MedObvious: Per-task accuracy (%). Task A–E** correspond to Detection MCQ, Detection Open, Referring MCQ, Referring Open, and Visual Referring, respectively. "Positive (Pos)"/"Negative (Neg)" indicate accuracy on anomalous/non-anomalous samples. Avg is the macro-average of Task A–E, Pos(+), and Neg(-). Avg confidence intervals are 95% case-level bootstrap CIs. Best results per task are **bolded**.

| Model | Task-A | Task-B | Task-C | Task-D | Task-E | Pos(+) | Neg(−) | Avg [95% CI] | Var. ($10^{-4}$) |
|---|---|---|---|---|---|---|---|---|---|
| *General open-source VLMs* | | | | | | | | | |
| LLaVA-1.5-7B | 40.4 | 22.3 | 22.1 | 35.7 | 50.0 | 37.5 | 31.9 | 34.3 [32.4, 36.2] | 0.94 |
| Qwen2-VL-7B | 32.3 | 49.5 | 72.7 | 25.1 | 53.1 | 56.3 | 28.7 | 45.4 [43.3, 47.6] | 1.23 |
| Qwen2.5-VL-7B | **58.7** | **82.1** | 75.3 | 29.7 | **65.1** | 60.9 | **70.7** | **63.2 [60.5, 66.0]** | **1.98** |
| Qwen3-VL-8B | 31.2 | 32.7 | **80.8** | **38.7** | 52.9 | **68.9** | 2.9 | 44.0 [41.9, 46.3] | 1.30 |
| InternVL2.5-8B | 56.3 | 56.3 | 69.7 | 26.8 | 51.7 | 59.7 | 42.5 | 51.9 [49.3, 54.4] | 1.72 |
| InternVL3-8B | 38.5 | 43.6 | 80.4 | 27.6 | 50.2 | 64.6 | 16.6 | 45.9 [43.5, 48.4] | 1.60 |
| Pixtral-12B | 31.0 | 22.5 | 76.6 | 26.3 | 51.9 | 59.5 | 5.3 | 39.0 [37.0, 41.1] | 1.12 |
| *Medical open-source VLMs* | | | | | | | | | |
| LLaVA-Med-7B | 10.0 | 36.8 | 21.2 | 23.4 | 50.0 | 37.1 | 17.5 | 28.0 [26.1, 30.0] | 0.97 |
| Fleming-8B | 26.8 | 23.8 | 78.3 | 23.8 | 50.0 | 57.4 | 5.3 | 37.9 [36.1, 39.8] | 0.86 |
| Medgemma1.5-4B-IT | 23.6 | **86.1** | 43.4 | 19.5 | **66.6** | 44.6 | **64.1** | 49.7 [47.7, 51.8] | 1.07 |
| Lingshu-7B | **39.3** | 78.5 | **79.5** | **26.8** | 61.9 | **66.8** | 43.8 | **56.6 [54.6, 58.8]** | **1.15** |
| *Proprietary VLMs* | | | | | | | | | |
| Gemini-2.0-Flash | 54.2 | 42.7 | **85.9** | **35.7** | **69.3** | **75.4** | 25.6 | **55.5 [53.3, 57.9]** | **1.39** |
| Gemini-2.5-Flash | **67.2** | 45.5 | 80.4 | 31.9 | 55.9 | 74.1 | **26.3** | 54.4 [52.2, 56.8] | 1.35 |
| GPT-4o | 47.2 | **50.4** | 62.8 | 26.3 | 61.7 | 68.0 | 22.7 | 48.4 [46.1, 50.9] | 1.54 |
| GPT-4.1-nano | 25.3 | 16.8 | 26.3 | 18.3 | 55.1 | 34.9 | 21.4 | 28.3 [26.1, 30.6] | 1.29 |
| GPT-4.1-mini | 41.9 | 32.9 | 53.1 | 29.3 | 64.2 | 64.0 | 13.6 | 42.7 [40.4, 45.1] | 1.40 |
| GPT-5-nano | 43.4 | 41.7 | 82.5 | 28.9 | 63.4 | 73.1 | 14.8 | 49.6 [47.5, 51.7] | 1.14 |
| *Human expert* | 82.1 | 85.7 | 82.1 | 90.9 | 92.9 | 89.4 | 95.7 | 88.4 | – |

Detection Open (*state the outlier position*), Referring MCQ (*choose an outlier description given its position*), Referring Open (*describe the outlier given its position*), and Visual Referring (*Yes/No for highlighted region*). Using both multiple-choice and open-ended settings exposes selection biases and over-generation.

**Negative controls.** To measure false alarms, MedObvious includes explicit negative controls where all panels are consistent and no outlier exists (37.5% of tasks; 705/1,880). The correct label is $y = \varnothing$ (or "No" for binary verification).

## 3 Experiments and Results

MedObvious targets a pre-diagnostic requirement, i.e., before interpretation, the model must verify that the input is coherent and safe to reason over. We therefore evaluate VLMs as potential *pre-diagnostic gatekeepers* using the following clinically grounded research questions:

**RQ1 (Gatekeeping and false alarms).** Can models detect gross input violations (wrong modality or anatomy) and decide whether to proceed or abstain? Also, can models correctly determine that no anomaly is present when given normal, internally consistent inputs?

**RQ2 (Set-level consistency).** How does performance change as the candidate set grows, requiring systematic comparison across images?

Table 3: **Paired comparisons as statistical significance.** Avg is the mean of Task A–E, Pos(+), and Neg(-). $\Delta$ denotes Avg(Model A) − Avg(Model B), with paired case-level bootstrap 95% confidence intervals. Exact McNemar Holm-adjusted $p$-values are reported for binary item-level correctness comparisons.

| Model A | Model B | Avg A | Avg B | $\Delta$ Avg [95% CI] | Paired units | Holm $p$ |
|---|---|---|---|---|---|---|
| *Within-category leader vs. runner-up* | | | | | | |
| Qwen2.5-VL-7B | InternVL2.5-8B | 63.25 | 51.88 | 11.37 [9.12, 13.59] | 235 | $< 10^{-8}$ |
| Lingshu-7B | Medgemma1.5-4B-IT | 56.70 | 49.74 | 6.96 [4.62, 9.29] | 235 | $2.65\times10^{-4}$ |
| Gemini-2.0-Flash | Gemini-2.5-Flash | 55.61 | 54.51 | 1.10 [-0.81, 3.02] | 235 | 1.00 |
| *Cross-category leader comparisons* | | | | | | |
| Qwen2.5-VL-7B | Lingshu-7B | 63.25 | 56.70 | 6.56 [4.52, 8.53] | 235 | $4.60\times10^{-7}$ |
| Qwen2.5-VL-7B | Gemini-2.0-Flash | 63.25 | 55.61 | 7.65 [5.11, 10.18] | 235 | $1.92\times10^{-6}$ |
| Lingshu-7B | Gemini-2.0-Flash | 56.70 | 55.61 | 1.09 [-1.22, 3.41] | 235 | 1.00 |
| *Representative proprietary-model contrasts* | | | | | | |
| Gemini-2.0-Flash | GPT-5-nano | 55.61 | 49.63 | 5.98 [3.94, 8.04] | 235 | $5.10\times10^{-7}$ |
| GPT-5-nano | GPT-4o | 49.63 | 48.49 | 1.14 [-0.93, 3.25] | 235 | 1.00 |

**RQ3 (Clinical semantics).** Do models reliably detect clinically meaningful mismatches (e.g., anatomy/viewpoint) rather than relying on superficial cues, and do they confuse such mismatches with pathology?

**RQ4 (Grounding).** Under physically implausible or cross-modality inconsistencies, do models reject the input or rationalize it with plausible narratives?

**RQ5 (Interface robustness).** Are sanity-check decisions consistent across binary, localization, and free-text interfaces, or strongly format-dependent?

**Evaluation Pipeline.** All models are evaluated on the full MedObvious benchmark in a **zero-shot** setting, without fine-tuning, retrieval augmentation, or few-shot exemplars. We evaluate three model groups: **general open-source VLMs** (LLaVA-1.5-7B [19], Qwen2-VL-7B [25] Qwen2.5-VL-7B, Qwen3-VL-8B [5], Intern VL2.5-8B [8], InternVL3-8B [30], Pixtral-12B [3]), **medical open-source VLMs** (LLaVA-Med-7B [15], MedGemma1.5-4B-IT, Lingshu-7B [27]), Fleming-8B [18], and **proprietary VLMs** (Gemini-2.0-Flash, Gemini-2.5-Flash, GPT-4o, GPT-4.1-mini/nano, GPT-5-nano). Each instance is evaluated in five formats, each with format-specific prompts. Outputs are constrained and parsed into a closed label space (option letter, grid position, or Yes/No) to ensure consistent scoring across models. Moreover, open-source models are evaluated with a unified inference pipeline on NVIDIA A100 (40 GB) GPU. Proprietary models are queried via public APIs using the same prompts and output normalization.

**Evaluation Metrics.** We report accuracy for each format, as well as **Positive** accuracy (outlier present), **Negative** accuracy (no outlier), and **Overall** accuracy. Reporting Positive and Negative separately is clinically important as a model can appear strong on anomaly-present cases, yet remain unsafe due to false alarms on normal inputs.

**Results.** We summarize results by research question. Table 2 reports performance on the 5 evaluation modes, Table 4 reports performance on 5 tiers, and Table 3 reports the pairwise statistical signifcance of the benchmarked VLMs.

Table 4: Per-tiers overall accuracy (%) for different VLMs.

| Model | T1 | T2 | T3 | T4 | T5 | All |
|---|---|---|---|---|---|---|
| *General open-source VLMs* | | | | | | |
| LLaVA-1.5-7B | 35.2 | 40.4 | 21.6 | 42.5 | 36.7 | 35.2 |
| Qwen2-VL-7B | 50.9 | 47.2 | 37.2 | 52.8 | 39.6 | 45.5 |
| Qwen2.5-VL-7B | **68.8** | **67.2** | **50.5** | **84.0** | 49.2 | **63.9** |
| Qwen3-VL-8B | 47.2 | 48.9 | 34.7 | 53.4 | 32.8 | 43.4 |
| InternVL2.5-8B | 58.8 | 56.0 | 33.0 | 67.2 | **49.3** | 52.8 |
| InternVL3-8B | 50.4 | 50.4 | 37.2 | 57.1 | 33.9 | 45.8 |
| Pixtral-12B | 45.0 | 41.4 | 26.3 | 47.8 | 33.2 | 38.7 |
| *Medical open-source VLMs* | | | | | | |
| LLaVA-Med-7B | 32.7 | 32.5 | 28.8 | 26.8 | 25.0 | 29.1 |
| Fleming-8B | 41.8 | 43.9 | 29.1 | 39.0 | 31.4 | 37.0 |
| Medgemma1.5-4B | 59.7 | 53.7 | 37.2 | 53.7 | **53.5** | 51.5 |
| Lingshu-8B | **62.9** | **64.7** | **48.8** | **63.1** | 46.0 | **57.1** |
| *Proprietary VLMs* | | | | | | |
| Gemini-2.0-Flash | **59.3** | **61.0** | **56.3** | **66.8** | 34.6 | 55.6 |
| Gemini-2.5-Flash | 59.0 | 62.7 | 55.2 | 62.1 | 35.0 | 54.8 |
| GPT-4o | 54.3 | 56.0 | 45.2 | 59.6 | 34.6 | 49.9 |
| GPT-4.1-nano | 29.3 | 31.8 | 20.2 | 31.8 | **37.5** | 30.1 |
| GPT-4.1-mini | 47.7 | 52.2 | 36.1 | 50.9 | 33.5 | 44.1 |
| GPT-5-nano | 52.5 | 56.6 | 46.6 | 61.2 | 33.2 | 50.0 |

It shows consistent paired rankings: general open-source models often perform better, while closed models are not statistically separable.

***RQ1 (Gatekeeping and false alarms).*** Table 2 shows that overall accuracy is still far from a reliable pre-diagnostic gate, with large variance between models. The most safety-relevant signal is the **Pos(+)** and **Neg(−)** split. Several models achieve high Pos(+) accuracy while collapsing on Neg(−), indicating a "*always-find-something*" bias that would be unacceptable for a gatekeeper. In contrast, a smaller subset achieves substantially higher Neg(−), demonstrating that abstention is learnable, but not consistently present across model families. Importantly, the proprietary scale does not automatically fix this failure mode, as the negative accuracy remains modest for many such models, suggesting that normal-case calibration is a distinct problem from diagnostic fluency.

***RQ2 (Set-level consistency).*** The tier analysis in Table 4 reveals that the scaling from $2 \times 2$ to $3 \times 3$ is not a minor increase in difficulty but a **qualitative failure point**. The systematic drop in T3 (SCALING) suggests that many models do not reliably perform an exhaustive comparison across candidates. Instead, they appear to rely on a small number of salient cues. This matters clinically because multi-view and multi-slice studies require consistent reasoning across many related images (CT slices or multi-view fetal ultrasound), not just the detection of a single obvious frame.

***RQ3 (Clinical semantics).*** Models often rebound on T4 (SEMANTICS), which emphasizes anatomy/viewpoint mismatches. This indicates that clinically significant distribution shifts (e.g., chest vs. abdomen, frontal vs. lateral) can be easier to detect than distractor-heavy scaling, likely because they produce greater global changes in appearance. However, this "semantic strength" does not imply safety: a model can detect an anatomy mismatch yet still hallucinate anomalies.

***RQ4 (Grounding).*** We test whether models reject invalid inputs rather than rationalize them. High false-alarm rates on negative controls, together with strong interface sensitivity, suggest that models do not behave as conservative verifiers and often predict an outlier even when the correct response is "none".

***RQ5 (Interface robustness).*** Performance is strongly format-dependent. Many models are high on Referring MCQ but low on Referring Open, suggesting option selection is easier than producing grounded descriptions. Similarly, the reverse pattern also appears for some models. Overall, sanity checking is not interface-invariant, so models should be evaluated for consistency across binary, localization, and explanation-style outputs rather than a single prompt format.

**Discussion.** MedObvious shows that fluent report-style generation does not imply reliable pre-diagnostic verification. Across models, the main failures are **false**

**alarms** on normal grids, **degradation under scaling**, and strong **format sensitivity**. These failures directly limit the use of VLMs as sanity-check assistants. The implications are stronger for *agentic* medical AI, where models act on multi-panel, multi-image inputs and verify set-level consistency of modality, anatomy, and integrity before safe action. A model that hallucinates anomalies or fails under larger candidate sets can propagate errors while being confidently fluent. Overall, MedObvious isolates this prerequisite layer and highlights the need to reduce false alarms and improve set-level comparison under scaling.

**Limitations.** MedObvious is a controlled, template-based benchmark and therefore does not capture the full complexity of real clinical workflows. It also does not directly test downstream diagnostic performance or cover the full space of VLM-specific failure modes, which remain important directions for future work.

## 4 Conclusion

We present **MedObvious**, a benchmark for pre-diagnostic visual sanity checking in medical VLMs. Across progressive tiers, formats, and negative controls, we find that current models remain unreliable gatekeepers, with frequent false alarms, scaling degradation, and strong format sensitivity, motivating pre-diagnostic triage as a prerequisite for safe clinical and agentic deployment. Future work can extend to full multi-series volumes and interactive viewer-based evaluation.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Agrawal, K.: To study the phenomenon of the moravec's paradox. arXiv preprint arXiv:1012.3148 (2010)
3. Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., De Monicault, B., Garg, S., Gervet, T., et al.: Pixtral 12b. arXiv preprint arXiv:2410.07073 (2024)
4. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes. 9-12 September 2019 (2019)

7. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
9. Dahou, Y., Huynh, N.D., Le-Khac, P.H., Para, W.R., Singh, A., Narayan, S.: Vision-language models can't see the obvious. arXiv preprint arXiv:2507.04741 (2025)
10. Fallahpour, A., Ma, J., Munim, A., Lyu, H., Wang, B.: Medrax: Medical reasoning agent for chest x-ray. arXiv preprint arXiv:2502.02673 (2025)
11. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging **30**(9), 1323–1341 (2012)
12. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
13. Jiang, S., Wang, Y., Song, S., Zhang, Y., Meng, Z., Lei, B., Wu, J., Sun, J., Liu, Z.: Omniv-med: Scaling medical vision-language model for universal visual understanding. arXiv preprint arXiv:2504.14692 (2025)
14. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
16. Lin, K.Q., Li, L., Gao, D., Yang, Z., Wu, S., Bai, Z., Lei, S.W., Wang, L., Shou, M.Z.: Showui: One vision-language-action model for gui visual agent. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19498–19508 (2025)
17. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
18. Liu, C., Li, D., Shu, Y., Chen, R., Duan, D., Fang, T., Dai, B.: Fleming-r1: Toward expert-level medical reasoning via reinforcement learning. arXiv preprint arXiv:2509.15279 (2025)
19. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
21. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14788–14798 (2025)
22. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language

models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
23. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): A multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. pp. 180–189. Springer International Publishing, Cham (2018)
24. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 164–169 (2017)
25. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
26. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
27. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
28. Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G.: User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage **31**(3), 1116–1128 (2006)
29. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
30. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)