HARNESSING MEDICAL VISUAL-LANGUAGE MODELS WITH AGENTIC EVIDENCE MANAGEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision–language agents increasingly rely on segmentation tools, and recent work has focused on acquiring the right mask, often treating a good mask as the end point of tool use. However, do these tool-augmented agents actually make effective use of the evidence once it is obtained? To answer this question, we need a benchmark that separates evidence use from segmentation quality, spans multiple medical image types, and allows the same mask to be presented in different ways. Existing medical VQA datasets do not jointly provide these properties. We therefore construct **Mask-Evidence MedVQA**, a 2.5K-example benchmark spanning four image types, where each question is paired with an answer-relevant ground-truth mask and multiple possible renderings. Our study reveals two limitations of direct tool use. First, **evidence presentation** matters. Different renderings of the same fixed mask can substantially change model behavior. Second, masks introduce an **evidence trade-off**. Highlighting a relevant region can expose diagnostic evidence the model previously missed, but the added visual cue can also induce incorrect associations. As a result, even a ground-truth mask can correct some answers while turning initially correct ones into incorrect ones. Together, these findings show that acquiring the right mask is not enough; an agent must also decide how to present the evidence and whether it should change the answer. We therefore introduce **TUNE** (**T**ool **U**se is **N**ot **E**nough), an evidence-management harness that explicitly manages what evidence to acquire, how to present it, and whether to accept the resulting answer update. Across three standard medical VQA benchmarks, TUNE improves its backbone's mean no-tool accuracy by **11.5** points and exceeds the previous state-of-the-art medical agent by **3.5** points.

1 INTRODUCTION

Medical vision–language model (VLM) agents increasingly consult specialist tools before answering questions about medical images (Li et al., 2024; Fallahpour et al., 2025; Du et al., 2026; Jiang et al., 2025a; Wang et al., 2026). Segmentation models such as MedSAM and BiomedParse are particularly useful because they return an explicit mask of the region relevant to the question (Ma et al., 2024; Zhao et al., 2025). Accordingly, recent work has focused on improving evidence acquisition: deciding when to invoke a tool, what finding or structure to segment, and how to obtain the relevant region reliably (Jiang et al., 2025a; Du et al., 2026; Wang et al., 2026). The field has thus improved mask acquisition while largely treating the returned mask as the end point of tool use. But do these tool-augmented agents actually make effective use of the evidence once it is obtained?

Answering this question requires **multiple medical image types** to test generality across domains. It also requires **masks as additional evidence**, so no-tool answers exist and mask-induced changes can be measured. Finally, **multiple renderings per mask** are needed to vary presentation while holding localization fixed. Existing medical VQA benchmarks provide at most two of the three, as shown in Table 1. We therefore construct Mask-Evidence MedVQA, with 2,494 questions across ultrasound, chest X-ray, endoscopy, and CT, each paired with a relevant ground-truth mask. Each mask can be rendered as an outline, calipers, a crop, an inset, or another style used in medical figures. Using ground-truth masks factors mask quality out of tool-use analysis, allowing us to measure how evidence use varies across image types and renderings.

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.

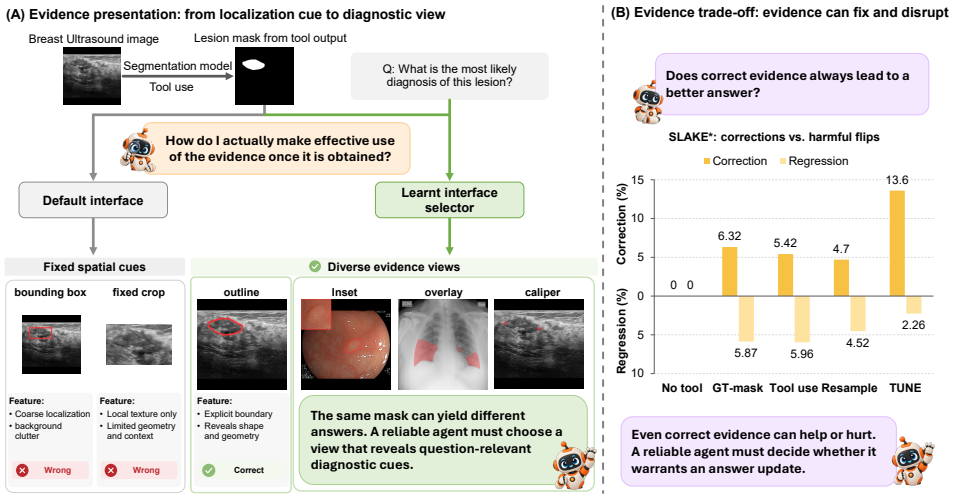

Figure 1: Tool use is not enough: reliable evidence use requires choosing how to present a mask and whether to update the answer. (A) Evidence presentation. The same mask can lead to different predictions depending on its rendering. In the ultrasound example, bounding-box and fixed-crop views yield incorrect answers, whereas an outline exposes lesion shape and yields a correct answer. Adaptive presentation selects among complementary views to reveal question-relevant diagnostic cues. **(B) Evidence trade-off.** Even ground-truth mask evidence can correct errors or disrupt initially correct predictions. On SLAKE*, upward bars show corrections (incorrect-to-correct changes) and downward bars show regressions (correct-to-incorrect changes), both relative to no-tool predictions. TUNE achieves the highest correction rate (13.6%) and the lowest regression rate (2.26%) among the compared evidence-using methods, illustrating the benefit of managing both evidence presentation and answer updates.

First, the same localized evidence can lead to different answers depending on how it is presented. In Fig. 1(b), the underlying ground-truth mask is identical, but different renderings, such as a crop or an outline, yield different predictions. This suggests that the model does not use a mask merely as a localization cue. An outline emphasizes boundary shape, calipers emphasize size, and a crop or inset changes the balance between local detail and surrounding context. Evidence presentation is thus part of the reasoning process: holding localization fixed does not hold the effective evidence seen by the model fixed. This interpretation is consistent with prior findings on visual prompting and the diagnostic conventions used in medical figures (Shtedritski et al., 2023; Jambor et al., 2021).

Second, correct evidence is not uniformly beneficial. In Fig. 1(a), the model answers correctly from the original image but changes to an incorrect answer after seeing a ground-truth mask. This is not an isolated failure: across Mask-Evidence MedVQA, added mask evidence both corrects previously incorrect answers and overturns previously correct ones. Highlighting can make question-relevant evidence easier to access, but it also changes visual salience and context and may introduce presentation-specific shortcuts. This exposes an evidence trade-off between informativeness and presentation-induced interference. Making relevant evidence more explicit can help the model discover missed information, but can also disrupt an otherwise correct prediction. Such failures are consistent with prior observations on visual cue-following and spurious correlations in medical imaging (Shtedritski et al., 2023; Jabbour et al., 2020). Therefore, a reliable agent should compare the evidence-conditioned answer with its original answer and decide whether the new evidence warrants an update.

These findings suggest that reliable tool use requires managing what evidence is acquired, how it is presented, and whether it should change the answer. We therefore introduce **TUNE** (**T**ool **U**se is **N**ot **E**nough), a harness that makes these decisions explicit from question to final prediction. TUNE consists of three stages. Building on prior work in evidence acquisition, the **Target Planner** converts the question into an executable request for answer-relevant visual evidence. To address the effects of presentation, the **Evidence Presenter** selects how the returned mask should be rendered. Because even correct evidence can hurt, the **Answer Critic** compares the original and evidence-conditioned predictions and decides whether the new evidence warrants an update. Together, these stages turn

Table 1: Medical VQA benchmarks and support for evidence evaluation. # QA denotes the number of question–answer pairs; image types denote imaging modalities. Mask as evidence means the original image alone supports an answer, with the mask providing additional evidence. Multiple renderings are alternative views of the same mask.

Dataset	# QA	Multiple image types	Mask as evidence	Multiple renderings
VQA-RAD (Lau et al., 2018)	0.45K	✓	✗	✗
SLAKE (Liu et al., 2021)	1.06K	✓	✗	✗
PathVQA (He et al., 2020)	6.76K	✗	✗	✗
PMC-VQA (Zhang et al., 2023)	2.00K	✓	✗	✗
Halt-MedVQA (Wu et al., 2024)	2.36K	✓	✗	✗
HEAL-MedVQA (Nguyen et al., 2025)	2.85K	✗	✓	✗
Mask-Evidence MedVQA-Eval (Ours)	**2.49K**	✓	✓	✓

one-shot "retrieve-and-answer" tool use into explicit evidence management, deciding what evidence to request, how to communicate it, and whether to act on it. We further co-evolve TUNE and the agent by coupling harness optimization with model adaptation. The first refined harness generates evidence-management trajectories to update the model. The updated model then informs a second harness refinement, whose trajectories support final model training.

We evaluate TUNE on three public medical VQA benchmarks, SLAKE, VQA-RAD, and OMVQA-3k (Liu et al., 2021; Lau et al., 2018; Hu et al., 2024), and the four datasets of Mask-Evidence MedVQA. Together, these datasets span nine image modalities and multiple organs. TUNE improves Qwen3.5-9B accuracy over direct tool use on all seven datasets, by 9.1 percentage points on average. On the three public benchmarks, TUNE raises the backbone's mean accuracy from 75.2% to 86.7%, a gain of 11.5 percentage points. This exceeds the strongest prior medical agent's mean accuracy by 3.5 percentage points (Du et al., 2026). In summary, our main contributions are:

- **A new benchmark.** Mask-Evidence MedVQA combines multiple medical image types, ground-truth masks as evidence, and multiple renderings per mask to measure evidence use while factoring out mask quality.
- **An evidence-management harness.** We introduce TUNE to manage what evidence to acquire, how to present it, and whether that evidence warrants an answer update.
- **Co-evolutionary training.** We co-evolve TUNE and the agent: trajectories from a refined harness support model adaptation, while the updated model informs further harness refinement before final training.
- **A stronger medical VLM agent.** TUNE improves over direct tool use on all seven datasets by 9.1 percentage points on average. Its mean accuracy exceeds that of the strongest prior medical agent by 3.5 percentage points on the three public benchmarks.

2 Related Work

Medical visual-tool agents and harnesses. Medical VLM agents increasingly call specialist tools before answering (Li et al., 2024; Fallahpour et al., 2025; Du et al., 2026; Jiang et al., 2025a), while foundation segmenters such as MedSAM and BiomedParse identify explicit image regions across modalities (Ma et al., 2024; Zhao et al., 2025). Recent systems improve this process in different ways. CARE proposes entities to segment and reviews evidence-grounded answers; Ophiuchus learns when and where to inspect (Du et al., 2026; Jiang et al., 2025a). MIRA searches over image operations, while Med-VRAgent coordinates region-level reasoning (Wang et al., 2026; Guo et al., 2025). These approaches largely focus on obtaining relevant evidence. We instead treat acquisition as the first stage of *evidence management*. Subsequent stages decide how to present the returned evidence and whether it should change the final answer.

Visual prompting and evidence presentation. Prior visual-prompting work mainly studies how to direct attention to a region. MedVP compares scribbles, rectangles, and ellipses (Zhu et al., 2025b), while radiology studies use arrows, boxes, and circles (Denner et al., 2024). Targeted Visual Prompting pairs a local crop with the full image (Tascon-Morales et al., 2025), and CARE selects

among crops, masks, and global indicators (Du et al., 2026). FAVP further considers masks and contours and finds that preferred styles vary across datasets (Yu et al., 2025). We take a different view: presentation should determine not only *where* the model looks, but also *what property* of a fixed region is made salient. Our interface library therefore spans localization, boundary and geometry, measurement, local appearance, and surrounding context. The Evidence Presenter learns from answer outcomes which view is useful for each question.

Evidence bias and answer review. Additional evidence can improve an answer while also introducing new biases. Language models may abandon correct predictions after receiving external context or tool outputs (Huang et al., 2025; Zhao et al., 2026; Tao & Agrawal, 2026). Multimodal models can also respond differently to equivalent content across modalities or visual markings (Lee et al., 2026; Shtedritski et al., 2023; Winkler et al., 2019). Existing medical agents typically review the evidence-conditioned answer itself (Du et al., 2026; Wang et al., 2026), which does not directly reveal whether the added evidence improved or merely swayed the model. TUNE instead preserves an independent no-tool answer and reviews the update by comparing it against the evidence-conditioned answer. Unlike self-correction methods that generate a revised answer (Jiang et al., 2025b), our Critic selects only between these two completed candidates.

3 Is Tool Use Alone Enough?

Prior work has made substantial progress in acquiring relevant visual evidence, but much less is known about how a VLM uses that evidence once it is returned. Studying this step requires multiple medical image types, masks introduced as additional evidence, and multiple renderings of each mask. Existing medical VQA benchmarks do not jointly provide these three properties (Table 1).

Benchmark setup. Mask-Evidence MedVQA is constructed from four medical segmentation datasets spanning breast ultrasound, chest X-ray, endoscopy, and CT (Al-Dhabyani et al., 2020; Tahir et al., 2021; Lan et al., 2021; Myronenko et al., 2023). For each image, we prompt an LLM with its ground-truth mask and diagnostic or category label to construct a question–answer pair supported by the labeled region. The question is written to remain answerable from the original image alone, so that the mask can later be introduced as additional evidence. We filter generated questions to ensure consistency with the source label and relevance of the masked region. The resulting benchmark contains 2,494 examples with ground-truth masks that support multiple renderings. These renderings let us hold localization fixed and isolate how the VLM uses the acquired evidence.

Human verification. We draw a stratified sample of 200 benchmark examples across the four image types. Three independent reviewers assess answerability from the original image, consistency with the source label, and relevance of the ground-truth mask to the question. Of the reviewed examples, 96% satisfy all three criteria according to at least two reviewers.

Findings and design implications. Two effects emerge from this controlled setting. First, changing only how a fixed mask is rendered can substantially change the answer, even though the underlying localized region remains identical. This shows that a mask is not merely a localization signal: how it is presented determines which properties of the region become salient to the VLM. Second, even a ground-truth mask can both correct an initially wrong answer and turn an initially correct one into a wrong answer. Thus, additional evidence carries an information–bias trade-off: it can reveal useful diagnostic information while also introducing visual cues that mislead the model.

4 TUNE: A Harness for Agentic Evidence Management

Overview. Given a medical image x and question q, TUNE manages segmentation evidence through three stages (Fig. 2). The **Target Planner** decides what visual evidence to segment, and the **Evidence Presenter** selects how to present the resulting mask to the VLM. The **Answer Critic** decides whether the evidence-conditioned answer should replace the model's original prediction. Together, these stages manage evidence acquisition, presentation, and final answer selection. All three roles are implemented by a shared VLM π_θ, while the segmentation tool and image renderers remain fixed external components.

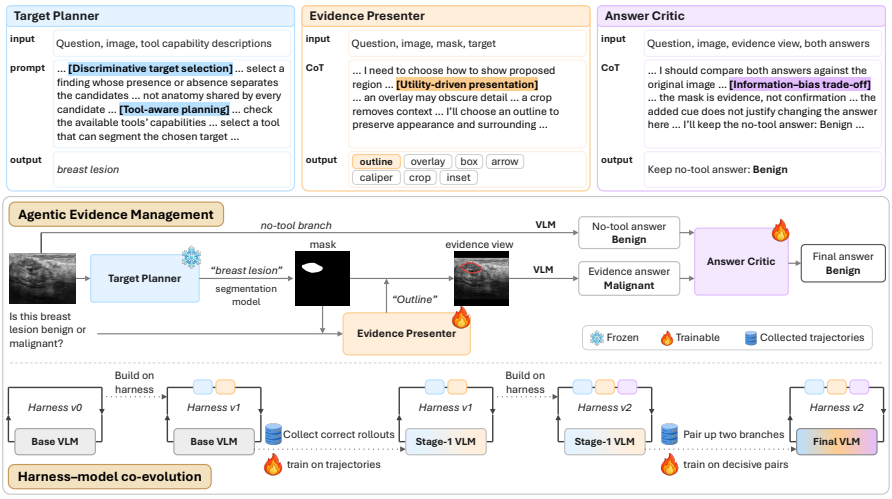

Figure 2: TUNE manages the path from a question to a warranted answer update. Top: the Planner requests a discriminative target; the external segmenter returns a mask; the Presenter selects a mask-derived interface; and the Critic chooses between the preserved no-tool answer and the evidence-conditioned answer. **Bottom:** the three TUNE roles on one example. Planning, presentation, answer generation, and criticism are role calls to a shared VLM π_θ; the segmenter and renderers remain fixed and external.

Pipeline. The Planner first proposes a semantic target t, which is passed to a segmentation tool S to obtain a mask m. The Evidence Presenter chooses an interface k from a fixed library $\mathcal{K}$, and a deterministic renderer produces the evidence view $v = R_k(x, m)$. The model independently generates a no-tool answer a_0 from (q, x) and an evidence-conditioned answer a_1 from (q, x, v). The Critic then compares the two candidates (a_0, a_1) and selects the final answer.

4.1 Target Planner: What Evidence to Acquire

The Target Planner determines what visual evidence should be acquired and which segmentation tool should be used to obtain it. Given the question q, image x, and descriptions of the available segmentation tools, the Planner outputs a semantic target t that is compatible with the selected tool. The selected tool then produces a segmentation mask m corresponding to the planned target.

Discriminative target selection. A target that is merely relevant to the question is often too broad to provide useful evidence for reasoning. For example, segmenting the *lung* identifies a relevant anatomical region but provides little information for distinguishing among candidate pulmonary diagnoses. Segmenting a *ground-glass opacity*, by contrast, exposes evidence that may directly support or rule out candidate answers. We therefore prompt the Planner to propose targets whose segmentation can provide *discriminative evidence*, rather than simply extracting entities that are related to the question.

Tool-aware planning. Segmentation models differ both in what they can segment and in the inputs they require. BiomedParse (Zhao et al., 2025) accepts text prompts for a predefined vocabulary of anatomical structures and lesions, but cannot segment targets outside that vocabulary. MedSAM (Ma et al., 2024), in contrast, takes a spatial prompt and can therefore handle targets such as *teeth* once they are localized. We describe these capabilities to the Planner and define a tool-specific workflow for each model. BiomedParse receives the planned semantic target directly; for MedSAM, a VLM first predicts a bounding box, which MedSAM uses to segment the target.

4.2 Evidence Presenter: How to Present the Evidence

The Evidence Presenter determines how the predicted segmentation mask should be presented to the VLM. Given the question q, image x, and planned target t, it selects an interface k from a fixed library $\mathcal{K}$. A deterministic renderer then produces the evidence view $v = R_k(x, m)$. The library includes overlays, outlines, contour fills, and mask-only views for region and boundary cues,

and bounding boxes, circles, crosshairs, and arrows for spatial cues. It also includes calipers for measurement, and crop and inset views for local appearance and surrounding context.

Utility-driven presentation. A mask specifies where the evidence lies, but how it is drawn shapes what the model can read from it. An outline can expose boundary shape, and a caliper can emphasize lesion extent; a crop or inset can reveal local appearance or preserve anatomical context. We therefore treat evidence presentation as an explicit reasoning decision rather than a fixed visualization choice. Because medical VQA datasets do not annotate useful presentations, we learn this choice from downstream answer utility. We hold the target and mask fixed, render every interface in $\mathcal{K}$, and supervise the Presenter using those that yield a correct evidence-conditioned answer.

4.3 Answer Critic: Whether to Trust the Evidence

The Answer Critic determines whether the evidence-conditioned answer should replace the model's original answer. Before observing the rendered evidence, the VLM produces a no-tool answer a_0; after observing v, it independently produces an evidence-conditioned answer a_1. The Critic receives q, x, t, v, both candidate answers, simple mask diagnostics, and relevant tool limitations, then selects the final answer.

Information–bias trade-off. Targeted evidence can expose useful diagnostic information, but the added visual cue can also induce misleading associations and turn a correct prediction into an incorrect one. We therefore preserve both the original answer a_0 and the evidence-conditioned answer a_1, and ask the Critic whether the new evidence justifies updating the model's initial judgment. Rather than generating another answer, the Critic selects only between these two candidates. We train it on *decisive pairs*, where a_0 and a_1 disagree and exactly one is correct. These pairs capture both cases where tool use helps and cases where it hurts. To prevent a positional shortcut, we randomly swap the order in which the two candidates appear in the Critic prompt during both training and inference.

4.4 Harness–Model Co-evolution

The Evidence Presenter and Answer Critic require supervision that standard medical VQA datasets do not provide. We derive this supervision from outcome-labeled trajectories as we develop the harness and model together.

Co-evolution recipe. Starting from a basic tool-use loop with the base VLM π_0, we first add the Target Planner and Evidence Presenter to improve the harness. Using a subset of $\mathcal{D}_{\text{train}}$, we enumerate the interfaces in $\mathcal{K}$ and retain rollouts with correct evidence-conditioned answers as $\mathcal{T}_0$. Fine-tuning π_0 on $\mathcal{T}_0$ yields an intermediate model π_1 that learns to select useful evidence presentations. We then add the Answer Critic to the harness and run π_1 over the full training set through both the no-tool and evidence-conditioned branches. Decisive answer pairs from these runs form answer-selection trajectories $\mathcal{T}_1$, which supervise when the evidence-conditioned answer should replace the original answer. Finally, we fine-tune the original base model π_0 on $\mathcal{T}_0 \cup \mathcal{T}_1$ to obtain π_2. With different role prompts, this model serves as the Target Planner, Evidence Presenter, answer generator, and Answer Critic.

The first harness revision thus supplies interface-selection supervision for π_1; the updated model then supplies the answer comparisons needed for the Critic and final model training.

5 Experiments

We evaluate TUNE from two perspectives. We first ask whether evidence management improves over direct tool use. Controlled analyses and ablations then examine how evidence presentation and answer review contribute to these gains.

5.1 Experimental Setup

Datasets. We evaluate on four Mask-Evidence MedVQA datasets (BreastUS, COVID-QU-Ex, NeoPolyp, and KiTS23) and three public medical VQA benchmarks (SLAKE, VQA-RAD, and OMVQA-3k) (Liu et al., 2021; Lau et al., 2018; Hu et al., 2024). Ground-truth masks are used only

Table 2: Accuracy (%) on public medical VQA benchmarks. Overall is the unweighted mean accuracy across the three datasets. Literature results may use different splits or scoring procedures, so rankings are contextual rather than controlled comparisons. Gold ●, silver ●, and bronze ● mark the top three distinct scores per column, with ties sharing a medal. Yellow rows denote our methods. TUNE ^{C} uses a fixed overlay; full TUNE includes model fine-tuning and learned evidence presentation.

System	Access	SLAKE	VQA-RAD	OMVQA-3k	Overall
Llama-3.2-11B-Vision (Meta, 2024)	Open	63.2	53.2	43.1	53.2
Qwen3.5-9B (Qwen Team, 2026)	Open	75.1	69.0	81.4	75.2
InternVL3-38B (Zhu et al., 2025a)	Open	68.7	63.0	78.6	70.1
DeepEyes-7B (Zheng et al., 2025)	Open	61.2	56.1	57.4	58.2
Gemma-4-E4B (Google DeepMind, 2026b)	Open	69.3	55.4	70.6	65.1
medgemma-27b (Sellergren et al., 2025)	Open	70.5	62.8	64.2	65.8
llava-med-v1.5-mistral-7b (Li et al., 2023)	Open	50.9	41.9	45.3	46.0
MedVLM-R1-2B (Pan et al., 2025)	Open	46.5	41.5	72.1	53.3
HuatuoGPT-Vision-34B (Chen et al., 2024)	Open	64.1	60.8	76.8	67.2
Lingshu-32B (LASA Team et al., 2025)	Open	82.3	64.8	84.0	77.0
Claude-Opus-5 (Anthropic, 2026)	Closed	80.7	69.6	83.0	77.8
GPT-5.6 (OpenAI, 2026)	Closed	83.1	74.3 ●	78.8	78.7
Gemini-3.7-Flash (Google DeepMind, 2026a)	Closed	83.0	74.3 ●	83.4	80.2
CARE-Flow-S (4B) (Du et al., 2026)	Open	78.4	56.3	94.5	76.4
CARE-Coord-S (Du et al., 2026)	Mixed	77.2	62.8	97.7 ●	79.2
CARE-Flow-B (10B) (Du et al., 2026)	Open	83.2	63.6	96.2	81.0 ●
CARE-Coord-B (Du et al., 2026)	Mixed	83.1	68.3	98.0 ●	83.1 ●
Ophiuchus-7B (Jiang et al., 2025a)	Open	83.9 ●	73.6	78.6	78.7
GPT-5.6 + TUNE ^{C}	Mixed	83.7	73.8 ●	75.3	77.6
Gemini-3.7-Flash + TUNE ^{C}	Mixed	85.0 ●	74.7 ●	80.7	80.1
Qwen3.5-9B + TUNE	Open	90.2 ●	71.6	98.2 ●	86.7 ●

in controlled analyses of evidence presentation and answer changes; TUNE itself always operates on predicted masks.

Models and baselines. Full TUNE is built on Qwen3.5-9B (Qwen Team, 2026), which we fine-tune using the trajectories described in Section 4.4. We additionally evaluate GPT-5.6 and Gemini-3.7-Flash (OpenAI, 2026; Google DeepMind, 2026a) with TUNE ^{C}, which retains evidence planning and answer review but uses a fixed overlay interface. For each backbone, we compare against direct tool use, which overlays the predicted mask and accepts the resulting answer directly. We also compare with tool-free VLMs and prior medical agents where applicable.

Metrics. We report answer accuracy, *corrections* (incorrect-to-correct changes), and *harmful flips* (correct-to-incorrect changes), with both types of change measured against the same no-tool answers. The latter two expose the information–bias trade-off that aggregate accuracy can hide.

5.2 Main Results

Does evidence management improve over direct tool use? Table 3 compares TUNE with direct tool use for each VLM backbone across all seven datasets. Our evidence-management variants improve every backbone on every dataset. For Qwen3.5-9B, full TUNE gains 9.1 percentage points on average over direct tool use. Even with presentation fixed to an overlay, TUNE ^{C} consistently improves GPT-5.6 and Gemini-3.7-Flash. Table 2 provides broader context against existing medical VLMs and agents. Qwen3.5-9B with TUNE reaches 86.7% mean accuracy across the three public benchmarks, the highest listed mean despite using an open 9B backbone. Because the literature results may use different evaluation protocols, this comparison is contextual rather than controlled.

How often does a good mask help versus hurt? Fig. 3 examines answer changes directly. Even a ground-truth mask produces both corrections and harmful flips. On COVID-QU-Ex, each occurs on 6.1% of questions, leaving aggregate accuracy unchanged despite substantial changes at the example level. Predicted masks exhibit similar behavior, showing that segmentation error alone does not

Table 3: Evidence management versus direct tool use. Accuracy (%) on seven medical VQA datasets, grouped by backbone. Tool overlays a predicted mask and accepts the resulting answer directly. Superscripts show percentage-point gains over the corresponding Tool row; yellow marks our methods. TUNE ^{C} uses a fixed overlay; full TUNE includes model fine-tuning and learned evidence presentation. COVID denotes COVID-QU-Ex.

System	Harness	Public Medical VQA			Mask-Evidence MedVQA			
		SLAKE	VQA-RAD	OMVQA-3k	BreastUS	COVID	NeoPolyp	KiTS23
GPT-5.6	Tool	78.6	71.6	64.3	82.0	63.5	75.5	72.9
	TUNE^{C}	83.7$^{+5.1}$	73.8$^{+2.2}$	75.3$^{+11.0}$	85.9$^{+3.9}$	66.6$^{+3.1}$	81.0$^{+5.5}$	73.7$^{+0.8}$
Gemini-3.7-Flash	Tool	80.0	71.8	75.0	83.6	74.8	72.5	71.2
	TUNE^{C}	85.0$^{+5.0}$	74.7$^{+2.9}$	80.7$^{+5.7}$	84.9$^{+1.3}$	76.1$^{+1.3}$	75.0$^{+2.5}$	75.2$^{+4.0}$
Qwen3.5-9B	Tool	74.8	67.4	81.2	83.6	62.7	77.5	73.0
	TUNE^{C}	77.1$^{+2.3}$	70.3$^{+2.9}$	83.3$^{+2.1}$	87.5$^{+3.9}$	65.3$^{+2.6}$	79.2$^{+1.7}$	76.8$^{+3.8}$
	TUNE	90.2$^{+15.4}$	71.6$^{+4.2}$	98.2$^{+17.0}$	89.8$^{+6.2}$	72.3$^{+9.6}$	83.0$^{+5.5}$	78.5$^{+5.5}$

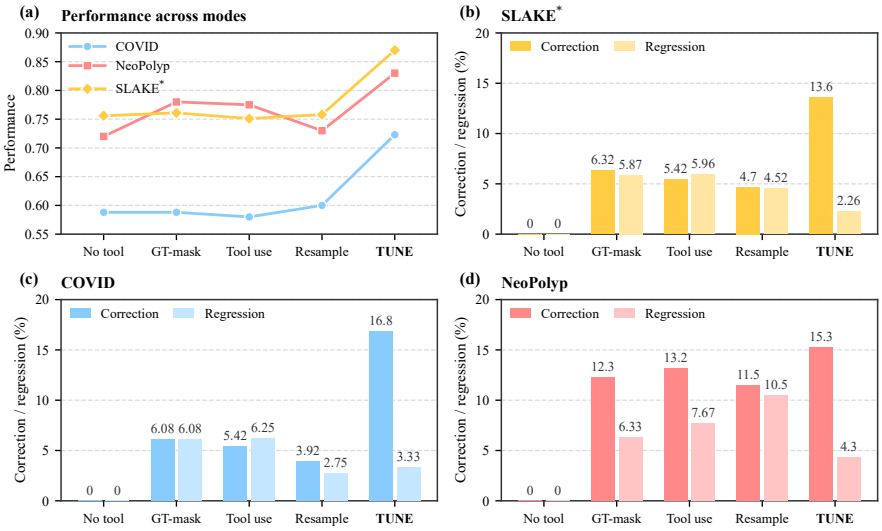

Figure 3: Mask evidence can correct or overturn no-tool answers. (a) Accuracy (fraction) on COVID-QU-Ex, NeoPolyp, and SLAKE*. **(b–d)** Corrections (wrong→right) and regressions (right→wrong), as percentages of all evaluated questions, relative to the same saved no-tool answer per question. GT-mask uses ground-truth masks; Tool use invokes tools directly; Resample uses resample-and-select; TUNE uses a learned critic. Predicted-mask Dice is 0.973, 0.931, and 0.808, respectively.

explain the trade-off. Masks also produce more corrections than resampling on all three controlled sets, indicating that they contribute genuine information. TUNE increases corrections across all three sets while reducing harmful flips from an average of 6.6% under direct tool use to 3.3%. Evidence management therefore improves not only final accuracy, but also how the model balances useful evidence against misleading updates.

5.3 Ablations and Design Analysis

Can rendering alone change outcomes? Table 4 holds the underlying mask fixed and changes only how it is presented. The effects are large and bidirectional. Relative to an overlay, calipers increase accuracy by 14.7 percentage points on COVID-QU-Ex but decrease it by 6.0 points on NeoPolyp. A crop changes accuracy on the same datasets by $+9.7$ and -7.5 percentage points, respectively. More broadly, no interface is consistently beneficial across image types. Even with localization unchanged, presentation can alter what the VLM extracts from the evidence. This supports treating interface selection as part of reasoning rather than a fixed visualization choice.

Table 4: Rendering effects vary across datasets. Changes in GPT-5.6 accuracy (percentage points) relative to the overlay interface, with masks held fixed across renderings. Dashes indicate the reference condition. Mask-Evidence MedVQA uses ground-truth masks; public pools use predicted masks.

	Mask-Evidence MedVQA				Public pools		
Interface	**BreastUS**	**COVID-QU-Ex**	**NeoPolyp**	**KiTS23-neut.**	**VQA-RAD**	**OMVQA-3k**	**MedXpert QA**
overlay (reference)	—	—	—	—	—	—	—
outline	+0.1	+6.0	-3.5	-0.3	+0.9	+2.0	0.0
outline (cyan)	+1.6	+4.3	-4.0	+1.7	-1.9	0.0	+0.8
contour fill	+1.6	+2.7	-3.0	-1.3	+1.4	-3.2	+1.2
bounding box	-0.8	+6.3	-3.5	-3.3	-2.4	-3.6	+2.0
circle	-1.6	+8.3	-6.5	+0.3	-1.4	-1.6	-0.4
crosshair	-3.9	+7.3	-2.0	-2.0	-2.4	+1.2	-0.8
arrow	-3.1	-2.0	-5.5	-2.3	-1.4	-2.4	-0.8
caliper	-0.8	+14.7	-6.0	0.0	0.0	-2.0	-0.8
crop	-4.7	+9.7	-7.5	-6.0	-0.9	-0.4	-1.2
inset	-0.8	+5.0	-3.5	+1.0	+1.4	-2.0	-0.4
mask only	+0.8	+3.3	-1.5	-3.7	-0.9	0.0	-0.8

Table 5: Component ablation of TUNE. Accuracy (%) as target planning, learned evidence presentation, and answer selection are progressively enabled. The Critic uses one-shot prompting or training on decisive answer pairs. Bold marks the highest accuracy per dataset; yellow highlights full TUNE.

Planner	**Presenter**	**Critic**	**BreastUS**	**COVID-QU-Ex**	**NeoPolyp**	**KiTS23**
—	—	—	82.8	58.8	72.0	71.3
✓	—	—	83.6	62.7	77.5	73.0
✓	✓	—	85.9	71.4	77.5	74.5
✓	✓	*One-shot*	85.2	70.3	77.0	71.3
✓	✓	*Trained*	**89.8**	**72.3**	**83.0**	**78.5**

Do learned presentation and answer review improve evidence use? Table 5 progressively enables the components of TUNE. Adding the Target Planner improves over no-tool answering on all four Mask-Evidence MedVQA datasets. The learned Presenter further improves BreastUS, COVID-QU-Ex, and KiTS23 and matches performance on NeoPolyp. Answer review becomes beneficial only after training: a one-shot Critic lowers accuracy on all four datasets, whereas training on decisive pairs yields the highest accuracy on each. These results show that both presentation and answer review benefit from supervision derived from downstream answer outcomes rather than being specified by prompting alone.

What information does the Critic need? We isolate the Critic on 27 decisive NeoPolyp pairs, each containing one correct and one incorrect candidate answer. With answer text alone, the Critic selects the correct answer in 55.6% of pairs. Adding the original image raises selection accuracy to 63.0%, while additionally providing the selected mask-derived rendering increases it to 74.1%. Thus, access to the evidence underlying an answer change improves selection accuracy by 18.5 percentage points over text alone, indicating that reliable answer review benefits from inspecting the visual evidence rather than comparing candidate answers in isolation.

6 Conclusion

We introduced **TUNE**, an evidence-management harness for medical VLM agents. Future work includes extending evidence management to richer medical tools and developing lighter, more scalable agent frameworks.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, February 2020. ISSN 2352-3409. doi: 10.1016/j.dib.2019.104863. URL http://dx.doi.org/10.1016/j.dib.2019.104863.

Anthropic. Introducing Claude Opus 5, 2026. URL https://www.anthropic.com/news/claude-opus-5.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *CoRR*, abs/2406.19280, 2024. doi: 10.48550/ARXIV.2406.19280. URL https://doi.org/10.48550/arXiv.2406.19280.

Stefan Denner, Markus Bujotzek, Dimitrios Bounias, David Zimmerer, Raphael Stock, and Klaus Maier-Hein. Visual prompt engineering for vision language models in radiology. *arXiv preprint arXiv:2408.15802*, 2024.

Yuexi Du, Jinglu Wang, Shujie Liu, Nicha C. Dvornek, and Yan Lu. CARE: Towards clinical accountability in multi-modal medical reasoning with an evidence-grounded agentic framework. *arXiv preprint arXiv:2603.01607*, 2026. URL https://arxiv.org/abs/2603.01607.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025.

Google DeepMind. Gemini 3.7 Flash model card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-7-flash/.

Google DeepMind. Gemma 4 model card, 2026b. URL https://ai.google.dev/gemma/docs/core/model_card_4.

Guangfu Guo, Xiaoqian Lu, and Yue Feng. Med-vragent: a framework for medical visual reasoning-enhanced agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 18613–18627, 2025.

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

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22170–22183. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02093. URL https://doi.org/10.1109/CVPR52733.2024.02093.

Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. To trust or not to trust? enhancing large language models' situated faithfulness to external contexts, 2025. URL https://arxiv.org/abs/2410.14675.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. Deep learning applied to chest x-rays: exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pp. 750–782. PMLR, 2020.

Helena Jambor, Alberto Antonietti, Bradly Alicea, Tracy L Audisio, Susann Auer, Vivek Bhardwaj, Steven J Burgess, Iuliia Ferling, Małgorzata Anna Gazda, Luke H Hoeppner, et al. Creating clear and informative image-based figures for scientific publications. *PLoS biology*, 19(3):e3001161, 2021.

Yankai Jiang, Yujie Zhang, Peng Zhang, Wenjie Li, Yichen Li, Jintai Chen, Xiaoming Shi, and Shihui Zhen. Ophiuchus: Incentivizing tool-augmented "think with images" for joint medical segmentation, understanding and reasoning. *arXiv preprint arXiv:2512.14157*, 2025a. URL https://arxiv.org/abs/2512.14157v2. Version 2, July 2026.

Yuhua Jiang, Yuwen Xiong, Yufeng Yuan, Chao Xin, Wenyuan Xu, Yu Yue, Qianchuan Zhao, and Lin Yan. Pag: Multi-turn reinforced llm self-correction with policy as generative verifier, 2025b. URL https://arxiv.org/abs/2506.10406.

Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet : Towards accurate colon polyp segmentation and neoplasm detection. In George Bebis, Vassilis Athitsos, Tong Yan, Manfred Lau, Frederick W. B. Li, Conglei Shi, Xiaoru Yuan, Christos Mousas, and Gerd Bruder (eds.), *Advances in Visual Computing - 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II*, volume 13018 of *Lecture Notes in Computer Science*, pp. 15–28. Springer, 2021. doi: 10.1007/978-3-030-90436-4_2. URL https://doi.org/10.1007/978-3-030-90436-4_2.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *CoRR*, abs/2506.07044, 2025. doi: 10.48550/ARXIV.2506.07044. URL https://doi.org/10.48550/arXiv.2506.07044.

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(1), November 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.251. URL http://dx.doi.org/10.1038/sdata.2018.251.

Jungyeon Lee, Yejin Yoon, and Taeuk Kim. Same semantics, different outcome: On the modality robustness of multimodal llms under knowledge conflict, 2026. URL https://arxiv.org/abs/2609.00550.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. Mmedagent: Learning to use medical tools with multimodal agent. *arXiv preprint arXiv:2407.02483*, 2024.

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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

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. *arXiv preprint arXiv:2102.09542*, 2021.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), January 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-44824-z. URL http://dx.doi.org/10.1038/s41467-024-44824-z.

Meta. Llama 3.2-Vision model card, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD_VISION.md.

Andriy Myronenko, Dong Yang, Yufan He, and Daguang Xu. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. In *International Challenge on Kidney and Kidney Tumor Segmentation*, pp. 1–7. Springer, 2023.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, Minh-Son To, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. Localizing before answering: A benchmark for grounded medical visual question answering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*, pp. 7670–7678, 2025. doi: 10.24963/ijcai.2025/853. URL https://www.ijcai.org/proceedings/2025/853.

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

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2025 - 28th International Conference, Daejeon, South Korea, September 23-27, 2025, Proceedings, Part VII*, volume 15966 of *Lecture Notes in Computer Science*, pp. 337–347. Springer, 2025. doi: 10.1007/978-3-032-04981-0_32. URL https://doi.org/10.1007/978-3-032-04981-0_32.

Qwen Team. Qwen3.5-9B model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023.

Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, December 2021. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.105002. URL http://dx.doi.org/10.1016/j.compbiomed.2021.105002.

Yufei Tao and Ameeta Agrawal. No-worse context-aware decoding: Preventing neutral regression in context-conditioned generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9736–9751, 2026.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Targeted visual prompting for medical visual question answering. In *Applications of Medical Artificial Intelligence*, volume 15384 of *Lecture Notes in Computer Science*, pp. 64–73. Springer, 2025. doi: 10.1007/978-3-031-82007-6_7. URL https://doi.org/10.1007/978-3-031-82007-6_7.

Shengzhi Wang, Jun Yang, Kai Wu, Xiaozhong Ji, Yiwen Ye, Ziyang Chen, Mingliang Xiong, Wen Fang, Mingqing Liu, Mengyuan Xu, Miaoxuan Shan, Caiyan Liu, Bin He, and Qingwen Liu. MIRA: Medical image reflection for agentic diagnosis, 2026. URL https://arxiv.org/abs/2608.10827.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155 (10):1135–1141, 2019.

Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*, 2024.

Ting Yu, Zixuan Tong, Jun Yu, and Ke Zhang. Fine-grained adaptive visual prompt for generative medical visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9662–9670, 2025.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: visual instruction tuning for medical visual question answering. *CoRR*, abs/2305.10415, 2023. doi: 10.48550/ARXIV.2305.10415. URL https://doi.org/10.48550/arXiv.2305.10415.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, Brian Piening, Carlo Bifulco, Mu Wei, Hoifung Poon, and Sheng Wang. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22(1):166–176, January 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02499-w. URL http://dx.doi.org/10.1038/s41592-024-02499-w.

Yufeng Zhao, Junnan Liu, Hongwei Liu, Dongsheng Zhu, Yuan Shen, Songyang Zhang, and Kai Chen. When tools hurt llm reasoning: State-dependent belief revision under external evidence, 2026. URL https://arxiv.org/abs/2508.15754.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. URL https://arxiv.org/abs/2505.14362.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a. URL https://arxiv.org/abs/2504.10479.

Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. Guiding medical vision-language models with diverse visual prompts: Framework design and comprehensive exploration of prompt variations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11726–11739. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.587. URL https://aclanthology.org/2025.naacl-long.587/.