Right Answer, Wrong Evidence: Revealing Visual Confabulation in Multi-Image Medical VLMs with the Evidence Attribution Score

Chidwipak Kuppani[1] and Bheemappa Halavar[1]

Department of Computer Science and Engineering, Indian Institute of Information Technology Sri City, India
chidwipak@gmail.com bheemappa.h@iiits.in

Abstract. Vision-language models achieve increasing accuracy in medical imaging tasks, but accuracy alone cannot reveal whether models rely on the correct visual evidence. In clinical settings, decisions must be justified by evidence, not merely correct outcomes. Correct answers built on wrong evidence undermine clinical trust. We introduce the Evidence Attribution Score (EAS), a behavior-based metric for evaluating visual grounding in model decisions. It compares the images a model actually depends on, measured through leave-one-out ablation, with the images it claims to use, measured through repeated prompting. Unlike accuracy, EAS reveals whether a model genuinely relied on the relevant images when arriving at a correct answer, a distinction that is critical for clinical trust. Evaluating eight vision-language models across multi-image medical questions, we find that 40 to 79 percent of correct answers are confabulated, meaning the model identified the right answer without using the right visual evidence. Three independent causal validations support that EAS captures genuine visual grounding. Chain-of-thought prompting dramatically increases EAS in open-source models while leaving closed-source models unchanged, revealing a systematic behavioral divide between model families that is invisible to accuracy alone. Our results show that accuracy-only evaluation can systematically overestimate reliability, and that EAS highlights scenarios where deployment may require additional safeguards.

Keywords: Evidence attribution · visual question answering · medical imaging · vision-language models · trustworthy AI

1 Introduction

Consider a longitudinal query comparing a current chest X-ray to a prior one. If the model correctly identifies "worsening consolidation" but behaviorally attends only to the prior image, ignoring the current one entirely, it has confabulated the progression. It got the right answer by guessing from the patient's history in the text prompt, not by comparing the pixels. In a clinical workflow where decisions depend on comparing multiple scans, a model that achieves high accuracy by

always favoring the most recent scan will fail when the progression pattern reverses. We call this failure mode *visual confabulation*: producing a correct answer without using the relevant visual evidence.

Vision-language models are increasingly deployed across radiology, pathology, and dermatology. Accuracy is the standard way to evaluate them, but it tells us nothing about whether these models actually used the images. In clinical settings, decisions must be justified by evidence, not merely correct outcomes. We introduce a score that tells clinicians whether a model looked at the right scans when it got the answer right.

Existing hallucination benchmarks measure whether models produce wrong outputs. We focus on a subtler and more dangerous problem, namely correct answers with wrong evidence. We define hallucination as producing an incorrect output, and confabulation as producing a correct output while relying on incorrect or irrelevant visual evidence. A model that hallucinates is visibly wrong. A model that confabulates appears right but will fail unpredictably when the visual evidence changes. We introduce the Evidence Attribution Score (EAS), which compares the images a model behaviorally depends on (via leave-one-out ablation) with those it claims to use (via prompting). Our contributions are fourfold. We define EAS and formalize visual dependency types. We show that 40–79% of correct answers across eight VLMs are confabulated. We validate EAS using three independent causal interventions. Finally, we demonstrate that chain-of-thought prompting substantially reduces confabulation in open-source models.

2 Related Work

Recent medical vision-language models, including Med-Flamingo, LLaVA-Med, BiomedCLIP, and RadFM [23, 26, 28, 42, 44], have demonstrated strong performance on medical visual question answering, evaluated on benchmarks such as VQA-RAD, SLAKE, and PathVQA [11, 21, 25]. More recent generalist systems extend these capabilities to broader multimodal clinical reasoning [34, 39], and MedFrameQA evaluates multi-image reasoning across radiology, pathology, and clinical photography [43]. Most evaluations in this area report accuracy as the primary metric. At the same time, surveys document the increasing deployment of VLMs in clinical workflows [37, 38]. However, accuracy only indicates whether the answer is correct; it does not reveal whether the model relied on the appropriate visual evidence. We address this limitation directly. To our knowledge, no prior work has explicitly quantified alignment between behavioral evidence and self-reported evidence in multi-image medical VLMs.

Prior work on hallucination focuses on detecting incorrect outputs. CHAIR and POPE measure object hallucination in image captioning [24, 32], while FaithScore evaluates faithfulness in long-form generation [17]. Recent surveys catalog the causes and mitigation of hallucination in multimodal models [4, 16], and decoding-time methods such as visual contrastive decoding and over-trust penalties reduce ungrounded generation [13, 22], while annotation-based detectors flag

unfaithful descriptions in long-form outputs [10]. In the medical domain, MedHallMark and MedVH benchmark hallucination in VLMs [5,9]. These approaches identify when models produce wrong answers. We instead study a different phenomenon: correct answers supported by incorrect or irrelevant evidence. As defined in the introduction, this confabulation is harder to detect because the output itself is correct, allowing it to pass standard evaluation metrics.

Explainability research seeks to interpret model reasoning. GradCAM and attention visualization highlight regions that models attend to [1,36], but attention does not imply behavioral necessity. SHAP and LIME estimate input contributions to predictions [27,31], yet they are correlational and do not measure whether models accurately report the evidence they use. In contrast, we apply input-level interventions through leave-one-out ablation to estimate behavioral dependence [29]. Our framework addresses a gap in multi-image reasoning by measuring whether models depend on, and correctly identify, the images essential for their predictions. Because our approach operates at the input-output level, it is model-agnostic and applicable to both open-source and closed-source systems.

Our work builds on a broad line of research asking whether models are right for the right reasons. Studies of shortcut learning and Clever Hans behavior show that high accuracy can mask reliance on spurious cues rather than task-relevant evidence [8, 20, 33]. In response, the interpretability community has formalized faithfulness criteria and rationale benchmarks that test whether a stated explanation reflects the computation behind a prediction [7, 14], a related debate questions whether attention weights themselves constitute explanations [15], and saliency studies caution that attributions may not survive sanity checks or input-ablation tests [3, 12]. Closest to our setting, work on visual grounding in single-image VQA uses human rationales to test whether a model both identifies and relies on relevant regions [30, 35]. We adapt and operationalize these ideas for multi-image medical VLMs, where gradients and region annotations are typically unavailable. Rather than comparing explanations against human rationales, EAS compares behavioral image dependence against the model's own claimed evidence at the image level, which yields an annotation-free and model-agnostic audit.

3 Method

Given a multi-image medical question with N images $(I_1, \ldots, I_n)$, answer options, and a VLM M, we ask whether M uses the correct visual evidence to arrive at its answer. We decompose this into two complementary measurements: *behavioral dependencies*, which identify images that cause answer changes when removed, and *claimed evidence*, which captures the images the model reports as important. We then measure their alignment.

Behavioral Evidence Set (BES). BES identifies images the model actually depends on. Operationally, we remove each image one at a time and re-query the model; if the answer changes, that image is essential. Formally,

$\text{BES}(q) = \{I_k : M(I \setminus I_k, q) \neq M(I, q)\}$. Leave-one-out ablation estimates behavioral dependence through input-level intervention without requiring gradient access, enabling comparison between open- and closed-source models. Importantly, EAS measures observable behavioral dependence rather than internal representation, aligning with deployment-time auditing goals where only input-output behavior is accessible.

We classify BES into three types. *Individual dependency*: a single image is essential. *Collective-redundant dependency*: multiple images each independently cause the answer to change, indicating information spread across images. *Text-prior dependency*: no image removal changes the answer, indicating usage of text only. Under our definition, correct answers with no measurable visual dependence are treated as confabulated. We note that this captures lack of observable visual dependence under our ablation protocol, rather than claiming absence of any internal visual representation.

Claimed Evidence Set (CES). The Claimed Evidence Set captures which images the model says are important. Single prompts make models list almost all images indiscriminately, so we ask three times with different phrasings and keep only images that are consistently mentioned. An image enters CES if it is identified in at least two of the three responses. This 2/3 majority threshold balances sensitivity against noise, and we verified that thresholds from 1/3 to 3/3 produce qualitatively similar results.

Evidence Attribution Score (EAS). EAS measures the overlap between what the model does (BES) and what it says (CES), as illustrated in Fig. 1:

$$\text{EAS}(q) = \frac{|\text{BES}(q) \cap \text{CES}(q)|}{|\text{BES}(q) \cup \text{CES}(q)|} \tag{1}$$

EAS is high when the model uses and claims the same images, low when it confabulates. We report mean EAS over all correctly answered questions; for text-prior cases ($\text{BES} = \emptyset$), $\text{EAS} = 0$ regardless of CES since the model shows no visual dependency. EAS therefore quantifies alignment between functional dependence and user-facing explanation. If a model cannot correctly identify the evidence it relies on, it behaves as a black box from the user's perspective, which is problematic for clinical trust.

We classify a model as a "reliable expert" when $\text{EAS} \geq 0.3$. $\tau = 0.3$ corresponds to minimal non-trivial agreement beyond chance-level overlap, implying at least one-third alignment between behavioral and claimed evidence. This threshold is analogous to slight-to-fair agreement in the Landis and Koch interpretation of the kappa statistic, representing the lowest level of meaningful alignment considered acceptable in behavioral evaluation [19]. We choose a modest threshold that credits partial overlap; our findings are robust across $\tau \in [0.1, 0.5]$ (confabulation always >36%).

4 Experimental Setup

We evaluate eight VLMs: five open-source (InternVL3-1B/2B/8B [6], Qwen2.5-VL-3B/7B [40]) and three closed-source (Gemini 2.5 Flash, GPT-4o [2], Claude 3

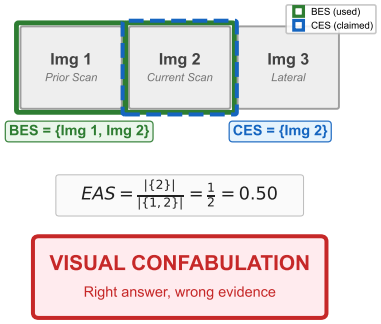


Fig. 1: EAS quantifies overlap between images a model depends on (BES) and images it claims are important (CES). The partial mismatch yields EAS = 0.50, indicating visual confabulation. Example adapted for illustration.

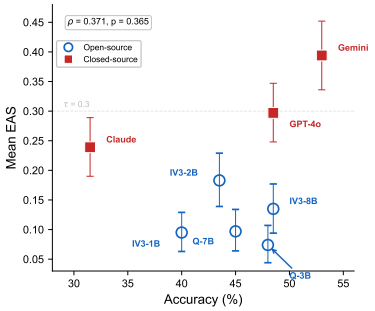


Fig. 2: Accuracy does not predict EAS across eight models (Spearman $\rho = 0.371$, $p = 0.365$). Qwen2.5-VL-3B achieves the highest open-source accuracy but the lowest EAS.

Haiku). Open-source models run on an NVIDIA L4 GPU. All models were evaluated with deterministic decoding (temperature = 0).

We use 200 multi-image questions from MedFrameQA [43], stratified by image count (50 each of 2–5 images). This sample size yields confidence intervals (width 0.06–0.12) sufficient for significance ($p < 0.001$). While modest in size, this sample matches prior causal evaluation studies and requires over 2,000 forward passes per model, making larger-scale evaluation computationally prohibitive.

For each model, we run the full EAS pipeline: 200 questions with leave-one-out ablation across all images (over 2,000 forward passes per model) and three CES prompts per question (6,000+ prompts total). For causal validation, we run three independent experiments (evidence swap, perturbation fragility, and Shapley analysis). All significance tests use paired permutation tests. All hyperparameters, prompts, and random seeds are documented to ensure reproducibility, and our evaluation code, prompts, and question identifiers are publicly available.

5 Results

We define the confabulation rate as the proportion of correctly answered questions with EAS $< \tau$ (where $\tau = 0.3$). Table 1 reveals three patterns: high confabulation rates across all models, higher mean EAS in closed-source models despite modest accuracy gains, and consistent agreement across validation signals.

Confabulation is pervasive (40–79% of correct answers). Even Gemini 2.5 Flash, the highest-accuracy model, confabulates on >50% of correct answers. Closed-source models show higher mean EAS than open-source (0.310 vs 0.117, $p < 0.001$), as summarized in Table 1.

Table 1: EAS results and causal validation across eight models. Acc = accuracy (%), Confab = confabulation rate (%), Swap Δ = evidence swap break rate difference (BES − non-BES), Frag Δ = perturbation fragility difference (high-EAS − low-EAS), Shap = Shapley-BES agreement. All Swap Δ and Frag Δ significant at $p < 0.001$.

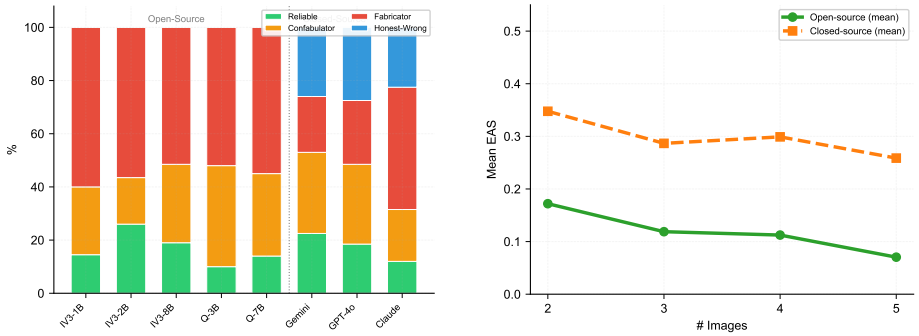
Model	Acc	EAS (95% CI)	Confab%	Swap Δ	Frag Δ	Shap
Open-source						
InternVL3-1B	40.0	.095 [.063,.129]	63.7	+22.1	+.098	.154
InternVL3-2B	43.5	.183 [.139,.229]	40.2	+24.1	+.166	.461
InternVL3-8B	48.5	.135 [.094,.177]	60.8	+22.0	+.150	.375
Qwen2.5-VL-3B	48.0	.074 [.044,.107]	79.2	+32.1	+.301	.208
Qwen2.5-VL-7B	45.0	.097 [.064,.134]	68.9	+44.6	+.257	.424
Closed-source						
Gemini 2.5 Flash	53.0	.394 [.336,.452]	57.5	+37.0	+.221	.479
GPT-4o	48.5	.297 [.248,.347]	61.9	+37.6	+.235	.531
Claude 3 Haiku	31.5	.239 [.190,.289]	61.9	+40.0	+.202	.513

Three independent validations support EAS. Evidence swap breaks BES images 22–45 percentage points more often ($p < 0.001$), fragility is two to fifteen times higher for high-EAS, and Shapley agrees with BES (0.393).

As shown in Fig. 2, accuracy rankings do not predict EAS rankings. Qwen2.5-VL-3B achieves the highest open-source accuracy but the lowest mean EAS, with nearly four out of five correct answers confabulated. Conversely, InternVL3-2B has lower accuracy but the highest open-source mean EAS. Across all eight models, the Spearman correlation between accuracy and EAS is $\rho = 0.371$ ($p = 0.365$), indicating no significant relationship. This reversal suggests that selecting models based solely on accuracy may overlook differences in visual grounding reliability. While we interpret low EAS as reliance on text priors, an alternative explanation is that Qwen uses visual signals not captured by leave-one-out ablation. However, our three causal validations consistently identify low-EAS answers as more brittle, supporting the confabulation interpretation.

EAS decreases as input images increase (Fig. 3). For open-source models, two-image questions yield mean EAS 0.14–0.24, dropping to 0.03–0.07 for five images. Closed-source models show a similar but less dramatic decline. This suggests multi-image reasoning becomes progressively harder, increasing reliance on text priors.

The chain-of-thought discovery. Chain-of-thought prompting dramatically increases visual evidence use in open-source models but leaves closed-source models almost unchanged. As shown in Table 2 and Fig. 4, open-source mean EAS increases by 0.174 to 0.364 (all $p < 10^{-6}$), while closed-source mean EAS changes by less than 0.04 (all $p > 0.11$). Notably, CoT does not improve accuracy. We hypothesize that CoT forces engagement with visual evidence that contradicts text priors, trading shortcuts for genuine grounding.



(a) Quadrant distribution per model.

(b) EAS decreases with image count.

Fig. 3: EAS characterization. (a) Open-source models dominated by fabricators (red) and confabulators (orange); closed-source models distribute more evenly. (b) EAS declines with image count, effect stronger in open-source models.

Model	Type	Base	CoT	ΔEAS	p	ΔCnfb
InternVL3-1B	Open	.095	.459	+.364	1.8e-17	−21.7 pp
InternVL3-2B	Open	.183	.504	+.321	5.5e-14	−1.1 pp
InternVL3-8B	Open	.135	.405	+.270	7.1e-13	−24.3 pp
Qwen2.5-VL-3B	Open	.074	.249	+.175	8.9e-08	−17.3 pp
Qwen2.5-VL-7B	Open	.097	.271	+.174	5.4e-08	−7.9 pp
Mean Open		**.117**	**.378**	**+.261**	$<10^{-6}$	**−14.5 pp**
Gemini 2.5 Fl	Closed	.394	.406	+.012	.633	−5.5 pp
GPT-4o	Closed	.297	.325	+.028	.129	−7.1 pp
Claude 3 Haiku	Closed	.239	.275	+.036	.117	−6.0 pp
Mean Closed		**.310**	**.335**	**+.025**	>0.11	**−6.2 pp**

Table 2: Effect of CoT prompting on EAS. Open-source models show dramatic improvement ($p < 10^{-6}$), closed-source models show no significant change.

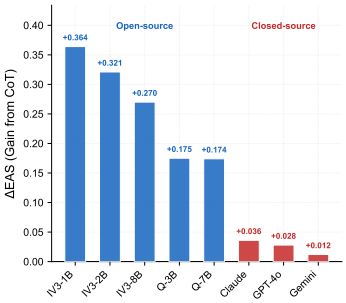


Fig. 4: CoT increases EAS in open-source models (+0.261) but not closed-source (+0.025).

InternVL3-2B jumps from mean EAS = 0.183 to 0.504 under CoT (accuracy unchanged), while Gemini 2.5 Flash remains at 0.394. CoT cuts confabulation in open-source models by double-digit points without retraining.

We interpret this as evidence that open-source models have latent visual reasoning abilities that direct answering fails to activate. This pattern holds consistently across five open-source models (1B–8B), suggesting a pattern consistent across the models and benchmark evaluated. Closed-source models show higher baseline EAS and no CoT improvement, aligning with findings that CoT forces explicit reasoning [18, 41]. This suggests that, under our operational grounding definition, these findings are consistent with confabulation reflecting inference-time routing behavior rather than purely representational limitations.

The BES type distribution helps explain the mechanism behind this divide. In open-source models, text priors dominate: 41 to 72 of 200 questions are answered

entirely from text with no visual dependency. Closed-source models have more balanced distributions (for example, Gemini: 78 text-prior, 54 collective, 68 individual), indicating greater visual engagement. After CoT prompting, open-source text-prior counts drop substantially, consistent with the EAS improvement.

6 Discussion

Our results indicate that visual confabulation is not an isolated failure mode but a consistent behavioral pattern across current medical VLMs. Even high-accuracy models routinely answer correctly without appropriate visual evidence, creating an illusion of competence and the risk of unpredictable failure when the visual evidence is critical. In medical imaging, correct answers built on wrong evidence may introduce reliability concerns in high-stakes workflows.

The ranking reversal between Qwen2.5-VL-3B (highest open-source accuracy, lowest EAS) and InternVL3-2B illustrates why accuracy is insufficient. Just as clinicians require both sensitivity and specificity, VLM deployment requires EAS to describe *how* a prediction was made. EAS is not intended to replace accuracy but to complement it by measuring grounding reliability, and future benchmarks should report the two together.

The quadrant distribution in Fig. 3(a) makes this divide concrete. Open-source models are predominantly populated by confabulators and fabricators, while closed-source models distribute more evenly with a larger share of reliable answers. As Fig. 3(b) shows, this gap widens with more images, suggesting closed-source models degrade more gracefully under increased visual complexity.

The CoT divide reveals a behavioral difference, since open-source models encode visual information that direct answering does not engage. In this setting CoT serves as an actionable safeguard for clinical deployment, separating latent capability from default usage.

Unlike interpretive methods, EAS rests on three independent causal validations, namely evidence swap, perturbation fragility, and Shapley agreement, all significant at $p < 0.001$. Because these validations are behavioral and require no human annotation, they enable scalable auditing of closed-source APIs.

Our study has several boundary conditions. We evaluate 200 questions from MedFrameQA, yielding bootstrap confidence intervals of 0.06–0.12 that support $p < 0.001$ detection of the reported differences. Leave-one-out BES assumes independence and does not capture distributed features, though Shapley analysis confirms its reliability. CES extraction relies on the model's articulation, and closed-source APIs prevent internal inspection.

Several directions merit exploration. Using EAS as a training signal during fine-tuning could directly optimize for evidence-grounded reasoning, real-time approximations without leave-one-out computation would enable deployment-time monitoring, and extending EAS to free-text generation and to alignment with clinician judgment would connect model behavior to domain expertise.

7 Conclusion

We introduced the Evidence Attribution Score (EAS), a behavior-based metric that reveals visual confabulation in multi-image medical VLMs. Across eight models, our results show that 40 to 79 percent of correct answers are confabulated, meaning the model arrived at the right answer without using the right visual evidence. Three independent validations (evidence swap, perturbation fragility, Shapley analysis) confirm that EAS captures genuine visual grounding ($p < 0.001$). Our chain-of-thought analysis reveals a systematic behavioral divide: CoT dramatically increases EAS in open-source models (mean +0.261, all $p < 10^{-6}$) but leaves closed-source models unchanged, suggesting confabulation reflects inference-time routing rather than purely representational limitations. These findings suggest the path to reliable clinical AI may depend less on scale and more on inference-time mechanisms. Our results show that visual reliability cannot be inferred from answer accuracy alone. Trust requires not just the right answer, but the right evidence.

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

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: ACL (2020)
2. Achiam, J., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: NeurIPS (2018)
4. Bai, Z., et al.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
5. Chen, Q., et al.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
6. Chen, Z., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR (2024)
7. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized NLP models. In: ACL. pp. 4443–4458 (2020)
8. Geirhos, R., et al.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
9. Gu, Z., Chen, J., Liu, F., Yin, C., Zhang, P.: MedVH: Toward systematic evaluation of hallucination for large vision language models in the medical context. Advanced Intelligent Systems **8**(1), 2500255 (2026)
10. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: AAAI (2024)
11. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
12. Hooker, S., Erhan, D., Kindermans, P.J., Kim, B.: A benchmark for interpretability methods in deep neural networks. In: NeurIPS (2019)

13. Huang, Q., et al.: OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR (2024)
14. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: ACL. pp. 4198–4205 (2020)
15. Jain, S., Wallace, B.C.: Attention is not explanation. In: NAACL-HLT. pp. 3543–3556 (2019)
16. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
17. Jing, L., et al.: FaithScore: Evaluating hallucinations in large vision-language models. arXiv preprint arXiv:2311.01477 (2023)
18. Kojima, T., et al.: Large language models are zero-shot reasoners. In: NeurIPS (2022)
19. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
20. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications **10**(1), 1096 (2019)
21. 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**, 180251 (2018)
22. Leng, S., et al.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR (2024)
23. Li, C., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS (2023)
24. Li, Y., et al.: Evaluating object hallucination in large vision-language models. In: EMNLP (2023)
25. 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: ISBI (2021)
26. Liu, H., et al.: Visual instruction tuning. NeurIPS (2023)
27. Lundberg, S., Lee, S.I.: A unified approach to interpreting model predictions. In: NeurIPS (2017)
28. Moor, M., et al.: Med-Flamingo: A multimodal medical few-shot learner. Machine Learning for Health (ML4H) (2023)
29. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press, 2nd edn. (2009)
30. Reich, D., Putze, F., Schultz, T.: Measuring faithful and plausible visual grounding in VQA. In: Findings of EMNLP. pp. 3129–3144 (2023)
31. Ribeiro, M.T., et al.: "why should i trust you?": Explaining the predictions of any classifier. In: KDD (2016)
32. Rohrbach, A., et al.: Object hallucination in image captioning. In: EMNLP (2018)
33. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: IJCAI. pp. 2662–2670 (2017)
34. Saab, K., et al.: Capabilities of Gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
35. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a HINT: Leveraging explanations to make vision and language models more grounded. In: ICCV. pp. 2591–2600 (2019)
36. Selvaraju, R.R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)

37. Singhal, K., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
38. Thirunavukarasu, A.J., et al.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023)
39. Tu, T., et al.: Towards generalist biomedical AI. arXiv preprint arXiv:2307.14334 (2023)
40. Wang, P., et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
41. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022)
42. Wu, C., et al.: Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463 (2023)
43. Yu, S., et al.: MedFrameQA: A multi-image medical VQA benchmark for clinical reasoning. arXiv preprint arXiv:2505.16964 (2025)
44. Zhang, S., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)