Hallucination Detection and Correction in Medical VLMs via Counter-Evidence Verification

Nan Zhou[1], Ke Zou[2], Meng Liu[1], Linchao He[1], Jiaqi Zhu[4], Yi Zhang[3], Hu Chen[1✉], and Huazhu Fu[5✉]

[1] the College of Computer Science, Sichuan University, Chengdu, China.
[2] Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
[3] the School of Cyber Science and Engineering and the Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University, Chengdu, China.
[4] National Key Laboratory of Autonomous Intelligent Unmanned Systems, Beijing Institute of Technology, Beijing, China.
[5] the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore.
huchen@scu.edu.cn; hzfu@ieee.org
✉Corresponding authors.

Abstract. Vision-Language models (VLMs) reliability in medical diagnosis is challenged by trust-undermining hallucinations. Existing hallucination detection approaches mainly focus on identifying factual inconsistencies between generated text and reference data. While some studies analyze where models attend in images, they seldom verify whether such attention truly reflects the visual evidence supporting the generated text. To address this gap, we propose **Counter-Evidence Verification (CoEV)**, a training-free plug-and-play framework that detects and corrects hallucinations through evidence-based factual consistency verification. CoEV performs bidirectional verification between textual assertions and visual evidence, testing whether each statement is supported by its corresponding evidence region, and assigns each statement into a four-quadrant diagnostic map capturing combinations of text factuality and visual grounding. CoEV detects hallucinated content and serves as a post hoc refinement tool, correcting hallucinations without retraining. Extensive experiments on four medical datasets show that CoEV combats hallucinations in VLMs.For hallucination detection, CoEV consistently outperforms existing methods, improving average PR-AUC and ROC-AUC by 3.0% and 3.9% absolute points respectively, with notable gains of up to 18.5% in specific VQA scenarios. For hallucination correction, it improves Micro-F1 by up to 12.5%, reduces hallucination rates by over 11.9% on medical report generation, and also boosts medical VQA accuracy. These results show that CoEV enables reliable detection and correction of hallucinations, providing clinicians with dependable, evidence-based cues for diagnosis. Code is avaliable at https://github.com/Zhounan1222/CoEV.

Keywords: Hallucination Detection · Counterfactual Intervention · Medical Vision-Language Models.

1 Introduction

Medical Vision-Language Models (VLMs) [12,21,22,26] have demonstrated transformative potential in clinical workflows, enabling automated interpretation across diverse tasks such as Medical Visual Question Answering (Med-VQA), Radiology Report Generation (MRG), and precise phrase grounding [4, 15, 28, 29]. By bridging high-dimensional visual features with complex clinical vocabulary, these models facilitate rapid diagnostic assistance. However, the deployment of VLMs in high-stakes clinical environments remains constrained by persistent hallucinations, where models generate clinically inaccurate, contradictory, or fabricated descriptions that lack support from underlying visual evidence [16]. In the medical domain, such ungrounded outputs are not merely linguistic errors; they represent significant safety risks, causing erroneous treatment recommendations and undermining clinician trust in AI-assisted decision-making [16].

Existing methods for hallucination detection primarily leverage uncertainty estimation, cross-model consistency checks [7, 13, 17, 25], or traditional linguistic similarity metrics like BLEU [19] and ROUGE [14]. While these approaches provide a coarse-grained measure of output quality, they are often insensitive to the factual grounding required for medical accuracy. Specifically, textual similarity metrics fail to capture nuanced hallucinations arising from flawed visual reasoning, where a report may be fluently written but medically incorrect [7, 17]. On the other hand, interpretability tools such as attention visualization or gradient-based saliency maps provide spatial intuition [1, 20]; however, they merely highlight statistical correlations and do not verify the rigorous factual alignment between specific textual assertions and localized image regions. Consequently, there remains a need for a diagnostic framework that transcends simple detection to provide traceable evidence, identifying not only the presence of hallucinations but also localizing the underlying visual regions responsible for them.

To address this, we propose Counter-Evidence Verification (CoEV), a plug-and-play and training-free framework for detecting and correcting hallucinations in medical VLMs. Unlike methods requiring external models or fine-tuning [9,23], CoEV performs diagnosis-informed interventions by selectively masking visual regions as counterfactual probes. This tests whether assertions depend on visual evidence or linguistic priors. By analyzing behavioral shifts, CoEV introduces a Four-Quadrant Diagnostic Map to systematically categorize claims into states of factual consistency. As a post-hoc mechanism, CoEV enables effective mitigation during inference without retraining. Evaluations across four datasets and multiple VLMs demonstrate that CoEV significantly improves diagnostic accuracy and consistency, providing an interpretable pathway for clinical AI.

2 Method

2.1 Preliminaries

Hallucination Definition. Building on [16], we define hallucination $y \in \{0, 1\}$ as a binary phenomenon conditioned on both textual and visual factors. Given

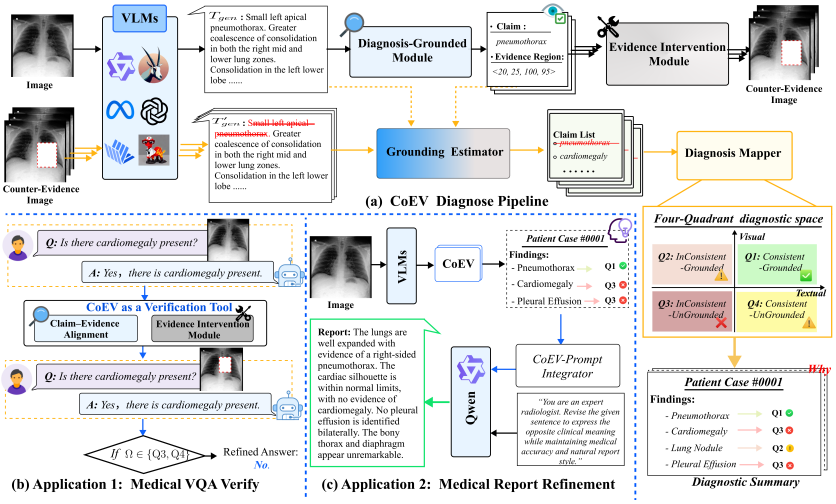

Fig. 1. (a) **CoEV Overview.** The CoEV diagnostic process performs counter-evidence verification and four-quadrant mapping to identify hallucinated claims. (b) Med-VQA Verification: Using diagnostic signals to validate answers and detect hallucinations. (c) Report Refinement: Rewriting hallucinated sentences using CoEV guidance.

a generated text T_{gen} and its corresponding image I, CoEV models the hallucination detection process as:

$$y = \mathbf{1}\big[(f(T_{\text{gen}}) = 0) \vee (z(T_{\text{gen}}, I) = 0)\big], \tag{1}$$

where $f(\cdot)$ measures factual correctness and $z(\cdot)$ evaluates grounding consistency with visual evidence. Specifically, $y = 1$ indicates a hallucination, which defined as a claim that is either factually incorrect ($f = 0$) or insensitive to the masking of referenced visual regions ($z = 0$). Conversely, a grounded claim ($y = 0$) must be both factually accurate and visually dependent. Fig. 1 illustrates how the CoEV integrate verification, reasoning, and visualization into an actionable diagnostic workflow for factual and visual correctness.

Diagnosis-Grounded Module (DGM). To operationalize the "What & Where" linkage, the DGM establishes grounding between claims and visual evidence. Let $\mathcal{P}(\cdot)$ be a semantic parser that extracts positive factual claims C_{pos} from generated text T_{gen}:

$$C_{\text{pos}} = \{c_i \mid \mathcal{P}(T_{\text{gen}})[i] = 1\}, \tag{2}$$

where label 1 indicates a positive clinical finding. For each $c_i \in C_{\text{pos}}$, a radiology expert $\mathcal{G}$ localizes a visual region $B_i \subset I$ with confidence $s_i \in [0, 1]$:

$$(B_i, s_i) = \mathcal{G}(I, c_i). \tag{3}$$

We define a grounding indicator $\delta_i \in \{0, 1\}$ as the basis for subsequent probing: $\delta_i = 1$ if $s_i > 0$, and $\delta_i = 0$ otherwise. This structured output from DGM serves as the prerequisite for our proposed counterfactual verification.

2.2 Counter-Evidence Consistency Evaluation (CECE).

The CECE module evaluates whether a generated claim is grounded in visual evidence or primarily driven by linguistic priors. CECE integrates two subprocesses: **Evidence Intervention Module (EIM) and Grounding Estimator (GE),** which jointly assess a claim's generation depends on visual support or linguistic bias. For each identified claim ($\delta_i = 1$), a counter-evidence image is constructed by masking its corresponding visual region B_i:

$$I_i' = \mathcal{M}(I, B_i), \tag{4}$$

where $\mathcal{M}$ is an intervention operator that masks B_i with black pixels while preserving the remaining image context. The VLM is then re-evaluated on the modified input to obtain $T_i' = g_\theta(I_i')$.

Let $\phi(c_i, T_i') \in \{0, 1\}$ indicate if claim c_i persists in the modified output T_i'. Specifically, T_i' is fed into the semantic parser ($\mathcal{P}(\cdot)$), and $\phi(c_i, T_i') = 1$ if c_i is still identified as a positive finding in T_i'; otherwise, $\phi(c_i, T_i') = 0$. The grounding attribution $z_i \in \{0, 1\}$ is defined as:

$$z_i = \begin{cases} 0, & \text{if } \phi(c_i, T_i') = 1, \\ 1, & \text{if } \phi(c_i, T_i') = 0, \end{cases} \tag{5}$$

Specifically, $z_i = 1$ confirms visual grounding via evidence reliance, while $z_i = 0$ reveals a prior-driven hallucination where the claim persists despite intervention. Unlike correctness-only methods, CECE explicitly verifies the *causal* dependence of claims on their supporting visual context.

2.3 Diagnosis Mapper (DM).

DM integrates factual correctness $f_i \in \{0, 1\}$ and grounding attribution $z_i \in \{0, 1\}$ into a unified diagnostic mapping $\mathcal{D}(f_i, z_i) \in \{Q_1, Q_2, Q_3, Q_4\}$:

$$\mathcal{D}(f_i, z_i) = \begin{cases} \text{Q1}, & f_i = 1, z_i = 1 \\ \text{Q2}, & f_i = 0, z_i = 1 \\ \text{Q3}, & f_i = 0, z_i = 0 \\ \text{Q4}, & f_i = 1, z_i = 0 \end{cases} \tag{6}$$

A claim is grounded ($y_i = 0$) if $(f_i, z_i) \in Q_1$, and hallucinated ($y_i = 1$) otherwise. These quadrants categorize assertions into factual alignment (Q_1), language-driven bias (Q_2), visual misinterpretation (Q_3), and total failure (Q_4). This taxonomy distinguishes medical inaccuracies from lack of visual dependence.

2.4 CoEV-Guided Correction Applications

CoEV enables correction across two downstream tasks: (1) verification for close-ended Med-VQA and (2) refinement for open-ended medical reports.

Med-VQA Verification. For a predicted answer $A = g_\theta(I, Q)$, CoEV evaluates its textual and visual integrity via $\Omega = \mathcal{F}(I, Q, A)$, ensuring the response is both accurate and visually grounded. where $\mathcal{F}$ denote the CoEV, and $\Omega \in \{Q_1, Q_2, Q_3, Q_4\}$. We define a verification operator that produces a corrected or validated answer:

$$\hat{A} = A \oplus \mathbb{I}[\Omega \in Q_2, Q_3] \tag{7}$$

This mechanism enables evidence-grounded verification of close-ended predictions and refines visually unsupported or hallucinated answers.

Report-level Refinement. For MRG, CoEV enables sentence-level correction of hallucinated content. Let the baseline report be $R_{\text{gen}} = \{s_1, s_2, \ldots, s_n\}$, and the diagnostic outcomes $\{(c_i, \Omega_i)\}$ be obtained by applying CoEV to each claim c_i in the report. Each sentence flagged as hallucinated ($\Omega_i \in \{Q_2, Q_3, Q_4\}$) is revised via a refinement operator $s_i' = \mathcal{R}(s_i, \mathtt{p})$, which prompts the LLM to rewrite the statement based on validated visual evidence, ensuring the final report is both factually accurate and grounded.

The refined report is reconstructed by replacing hallucinated sentences while leaving others unchanged:

$$\hat{R} = \begin{cases} s_i', & \text{if } \Omega_i \in \{Q_2, Q_3\}, \\ s_i, & \text{otherwise.} \end{cases} \tag{8}$$

This process leverages CoEV diagnostics to guide automatic correction, ensuring the final report aligns with factual and visual grounding consistency.

3 Experiments

3.1 Experimental Setup

Tasks and Datasets. We evaluate CoEV on two medical vision-language tasks: **(1) Med-VQA**: Conducted on VQA-RAD [11],and MIMIC-VQA [2], focusing on visually grounded clinical assertions.**(2) MRG:** Evaluated on MIMIC-CXR [10] and IU-Xray [6] datasets using official test splits. We utilized the GREEN model [18] to generate hallucination ground-truth labels by referencing gold-standard answers.

Evaluation Metrics. For hallucination detection, we employ PR-AUC and ROC-AUC. Report quality is assessed via diagnostic accuracy, following [4] (CheXpert-style F_1 scores: Mi-$F_{1\text{-}5}$ and Mi-$F_{1\text{-}14}$) and hallucination metrics: CHAIR-I/S [8] for instance/sentence-level errors, and MediHall [5] for clinically weighted penalties. VQA performance is measured by model accuracy. To assess generalizability, we benchmark CoEV across diverse VLMs, spanning general-purpose and medical-tuned architectures.

3.2 Results and Analyses

Hallucination Detection Performance. As shown in Table 1, CoEV consistently outperforms baselines across diverse model scales. Compared to the

Table 1. Hallucination detection performance (%) across MRG and Med-VQA tasks. **PR**: PR-AUC; **ROC**: ROC-AUC. Best results are **bolded**, second-best are underline.

Method	MIMIC-CXR		IU-Xray		VQA-RAD		MIMIC-VQA	
	PR ↑	ROC ↑	PR ↑	ROC ↑	PR ↑	ROC ↑	PR ↑	ROC ↑
General-purpose Model: InternVL3-2B [27]								
RadFlag [25]	70.61	54.02	77.02	57.97	65.32	54.09	70.54	62.71
SE [7]	76.96	52.54	78.62	59.37	66.51	52.58	60.93	64.10
VASE [13]	81.66	54.43	80.00	61.76	68.48	58.11	62.79	68.06
CoEV (Ours)	**82.35**	**56.75**	**80.21**	**61.89**	**72.35**	**67.45**	**81.31**	**76.95**
General-purpose Model: Qwen2-3B [3]								
RadFlag [25]	73.41	56.68	87.56	62.82	82.99	**74.37**	79.91	61.08
SE [7]	78.74	58.96	89.20	62.63	86.16	70.15	61.66	60.18
VASE [13]	75.18	52.19	**89.92**	**68.09**	89.71	61.81	71.01	64.34
CoEV (Ours)	**79.20**	**60.04**	89.18	64.61	**92.16**	69.25	**90.89**	**72.65**
Medical-tuned Model: Lingshu-7B [24]								
RadFlag [25]	83.29	55.31	72.67	40.21	87.71	52.39	84.10	61.37
SE [7]	84.73	59.48	77.53	44.68	81.85	52.47	66.69	61.79
VASE [13]	**87.59**	61.87	77.89	49.66	87.45	58.78	65.29	65.51
CoEV (Ours)	87.10	63.67	**79.90**	**49.69**	**89.73**	**65.63**	**85.10**	**70.60**
Medical-tuned Model: LLaVA-Med [12]								
RadFlag [25]	91.53	53.13	70.61	54.02	71.86	61.12	79.54	54.11
SE [7]	92.76	54.94	66.70	52.54	63.00	64.34	62.40	62.00
VASE [13]	**94.26**	59.34	75.94	56.50	63.02	67.01	62.57	71.08
CoEV (Ours)	93.94	**61.97**	**76.17**	**59.95**	**86.57**	**68.94**	**72.06**	**76.12**

second-best method, CoEV achieves an average absolute improvement of 3.0% in PR-AUC and 3.1% in ROC-AUC. While SE exhibits instability in larger models and VASE yields moderate results, CoEV provides robust detection. Notably, on InternVL3-2B for MIMIC-VQA, CoEV improves PR-AUC and ROC-AUC by 10.77% and 8.89% respectively. This highlights that our framework effectively isolates ungrounded clinical claims where purely statistical metrics fail. The performance gain across both open-set reporting and closed-set VQA tasks underscores CoEV's generalizability in safety-critical medical environments.

CoEV as a Post-hoc Refinement Tool. Beyond detection, CoEV serves as an effective pipeline for downstream clinical tasks. **(1) Medical Report Refinement:** As shown in Table 2, CoEV significantly enhances diagnostic accuracy. On average, $F_{1\text{-}5}$ improves by 7.01% on MIMIC-CXR and 8.16% on IU-Xray. Notably, for InternVL-8B, $F_{1\text{-}5}$ increases from 49.80% to 64.96% (+15.16%) on MIMIC-CXR. The reduction in hallucinations is evidenced by decreased CHAIR scores (e.g., C-I drops by 11.86% for IU-Xray on Qwen2-3B) and improved MediHall indices, confirming that CoEV successfully prunes unsupported clinical assertions while preserving factual findings. **(2) Med-VQA Verifica-**

Table 2. Hallucination detection performance (%). **F5/F14**: diagnostic F_1 scores; **C-I/C-S**: CHAIR-I/S ↓; **MH**: MediHall ↑. ✓ denotes our CoEV framework.

Model	CoEV	MIMIC-CXR					IU-Xray				
		Acc ↑		Hallu			Acc ↑		Hallu		
		F_5	F_{14}	C-I↓	C-S↓	MH↑	F_5	F_{14}	C-I↓	C-S↓	MH↑
Qwen2-3B	✗	16.95	16.46	69.68	16.25	42.80	12.50	17.41	90.80	18.82	70.73
	✓	19.69	18.64	65.92	14.66	43.80	32.78	57.63	86.93	14.32	72.74
InternVL-8B	✗	49.80	40.30	49.10	21.33	70.78	24.84	56.64	61.33	7.71	86.24
	✓	64.96	47.89	38.67	14.63	72.24	30.65	57.86	49.47	6.09	87.71
LLava-Med	✗	18.76	21.12	74.06	27.08	34.01	22.83	43.41	86.98	23.77	76.50
	✓	23.80	28.29	72.92	23.43	36.35	12.50	57.50	86.35	20.35	77.91
Lingshu-7B	✗	30.35	24.93	47.35	9.53	85.38	25.97	57.97	56.45	1.90	90.88
	✓	34.15	26.66	37.08	6.47	87.09	30.77	57.55	41.30	10.79	91.70

Table 3. Performance comparison on MIMIC-VQA and VQA-RAD. "Ours" indicates applying our refinement. Improvement (↑) is shown in absolute %.

Model	MIMIC-VQA			VQA-RAD		
	w/o Ours	w/ Ours	↑	w/o Ours	w/ Ours	↑
Qwen2-3B	51.83	78.99	**27.16**	71.31	78.09	**6.78**
InternVL-8B	60.38	74.13	**13.75**	76.89	82.47	**5.58**
LLaVA-Med	68.56	78.99	**10.43**	71.35	85.66	**14.31**
Lingshu-7B	64.42	78.22	**13.80**	78.09	91.24	**13.15**

tion: According to Table 3, CoEV yields a consistent mean accuracy gain of 13.04% across evaluated datasets. Smaller models exhibit the most pronounced gains, such as a +27.16% absolute increase for Qwen2-3B on MIMIC-VQA. Even for specialized models like Lingshu-7B, CoEV provides steady improvements (avg. +13.47%), underscoring its architecture-agnostic robustness. These results demonstrate that CoEV effectively mitigates hallucinations to improve the reliability of medical visual reasoning.

Table 4. Ablation of the contribution of CoEV's two diagnostic axes on MIMIC-CXR.

Textual Axis	Visual Axis	Precision	Recall	F1-Score
✗	✗	0.5233	0.4095	0.5064
✗	✓	0.6307	0.4532	0.5460
✓	✗	0.7474	0.5042	0.5986
✓	✓	0.7750	0.5046	0.6112

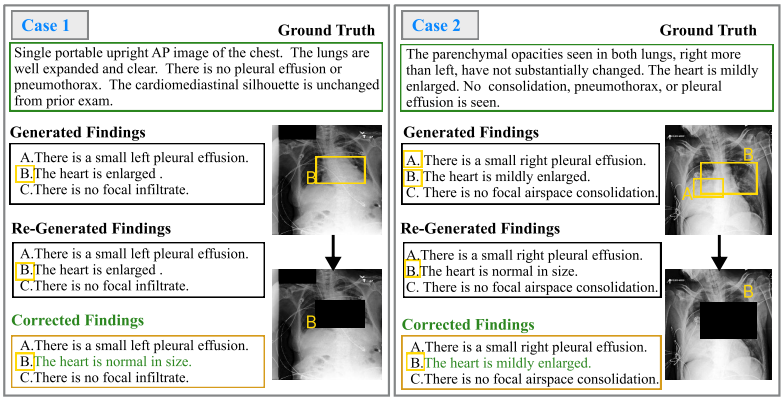

Fig. 2. Representative cases. **Case 1** (Q_2): description persists after masking, indicating ungrounded bias. **Case 2** (Q_1): description disappears post-masking, reflecting evidence-based reasoning. CoEV refines reports by aligning claims with visual evidence.

3.3 Ablation Study

We ablate CoEV's two core diagnostic dimensions: (1) the *textual axis*, which evaluates textual consistency, and (2) the *visual axis*, which evaluates whether a claim depends on the visual evidence.

As shown in Table 4, removing both axes yields the lowest performance with an F1-Score of 0.5064. Using only the visual axis increases the F1-Score to 0.5460, indicating that visual dependence carries strong discriminative value for identifying ungrounded assertions. Introducing only the textual axis improves precision by more than 20%, yet it remains limited because visually unsupported but plausible statements cannot be distinguished. Combining both axes achieves the highest overall F1-Score of 0.6112. This joint model consistently improves precision and recall simultaneously, demonstrating that factual correctness and visual dependence provide complementary signals. This confirms that CoEV's two-axis decomposition offers a meaningful and more discriminative structure for characterizing hallucination types. Fig. 2 illustrates CoEV's ability to distinguish and refine unsupported claims. In multi-disease scenarios, masking region A impacts only claim A, confirming that CoEV captures localized causal effects instead of global OOD noise, thus ensuring reliable refinement.

4 Conclusion

We present **CoEV**, a unified **Counter-Evidence Verification** framework for detecting and correcting hallucinations in medical VLMs. Unlike prior methods that constrain outputs during generation, CoEV enables a plug-and-play post-hoc evaluation that integrates seamlessly into existing clinical workflows. By intervening on visual evidence and mapping claims onto a four-quadrant diagnostic space, CoEV provides fine-grained, interpretable explanations that help

clinicians trace the origin of model errors. Extensive experiments on MRG and Med-VQA benchmarks demonstrate that CoEV consistently outperforms baselines, significantly enhancing factual accuracy while reducing hallucination rates across various architectures. Ultimately, CoEV establishes a robust foundation for evidence-based reliability in safety-critical AI. Looking forward, we aim to extend this framework to address multi-modal inconsistencies and explore its deployment in real-time clinical decision support systems.

Acknowledgements. This work was supported in part by the Sichuan Natural Science Foundation under Grant 2026NSFSC0426, the National Natural Science Foundation of China under Grant U25A20439, the National Natural Science Foundation of China under Grant 624B2027, and the Singapore National Medical Research Council under Grant MOH-001745-00.

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

References

1. An, J., Joe, I.: Attention map-guided visual explanations for deep neural networks. Applied Sciences **12**(8), 3846 (2022)
2. Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Chang, E., Kim, T., et al.: Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images (2024)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023)
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
5. 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)
6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2015)
7. Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. Nature **630**(8017), 625–630 (2024)
8. 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 p. 2500255 (2025)
9. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18135–18143 (2024)
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

11. 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), 1–10 (2018)
12. 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)
13. Liao, Z., Hu, S., Zou, K., Fu, H., Zhen, L., Xia, Y.: Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 669–679. Springer (2025)
14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
15. Liu, B., Zou, K., Zhan, L.M., Lu, Z., Dong, X., Chen, Y., Xie, C., Cao, J., Wu, X.M., Fu, H.: Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21310–21320 (2025)
16. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
17. Moslonka, C., Randrianarivo, H., Garnier, A., Malherbe, E.: Learned hallucination detection in black-box llms using token-level entropy production rate. arXiv preprint arXiv:2509.04492 (2025)
18. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. In: Findings of the association for computational linguistics: EMNLP 2024. pp. 374–390 (2024)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
20. Raghavan, K., B, S., v, K.: Attention guided grad-cam: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimedia Tools and Applications **83**(19), 57551–57578 (2024)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)
23. Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., Zhu, L.: Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 25543–25551 (2025)
24. 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)
25. Zhang, S., Sambara, S., Banerjee, O., Acosta, J., Fahrner, L.J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. arXiv preprint arXiv:2411.00299 (2024)

26. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv:2303.00915 (2023)
27. 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)
28. Zou, K., Bai, Y., Chen, Z., Zhou, Y., Chen, Y., Ren, K., Wang, M., Yuan, X., Shen, X., Fu, H.: Medrg: Medical report grounding with multi-modal large language model. arXiv e-prints pp. arXiv–2404 (2024)
29. Zou, K., Bai, Y., Liu, B., Chen, Y., Chen, Z., Zhou, Y., Yuan, X., Wang, M., Shen, X., Cao, X., et al.: Uncertainty-aware medical diagnostic phrase identification and grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)