Beyond Visual Forensics: Auditing Multimodal Robustness for Synthetic Medical Image Detection

Ching-Hao Chiu[1], Hao-Wei Chung[1], Gelei Xu[1], Xueyang Li[1], Pin-Yu Chen[2], John Kheir[3,4], Meysam Ghaffari[5], Carlos Morato[5], Ahmed Abbasi[1], and Yiyu Shi[1⋆]

[1] University of Notre Dame, Notre Dame, IN, USA
{cchiu3, hchung6, gxu4, xli34, aabbasi, yshi4}@nd.edu
[2] IBM Research, Yorktown Heights, NY, USA
Pin-Yu.Chen@ibm.com
[3] Department of Cardiology, Boston Children's Hospital, Boston, MA, USA
John.Kheir@cardio.chboston.org
[4] Department of Pediatrics, Harvard Medical School, Boston, MA, USA
[5] Optum AI, UnitedHealth Group, Minneapolis, MN, USA
{seyed_ghaffaridehkordi, carlos.morato}@optum.com

Abstract. With the rapid adoption of generative AI, synthetic medical images pose growing risks, including diagnostic deception and insurance fraud. Although prior work has explored vision–language model (VLM)–based synthetic image detection, these evaluations typically consider images in isolation. In clinical practice, however, images are interpreted alongside structured records and metadata, and VLMs are increasingly deployed under joint image–record inputs. We uncover a previously underexamined multimodal vulnerability: when given both modalities, VLMs may overweight record context in authenticity judgments, such that the same image receives different predictions solely due to changes in its accompanying text. This raises concerns about robustness in real-world deployment. To systematically characterize this effect, we reformulate synthetic medical image detection as an audit of multimodal robustness at the image–record interface and introduce a paired benchmark that holds the image fixed while swapping controlled metadata variants. Across multiple imaging modalities, we evaluate diverse open-weight and frontier API VLMs and quantify how metadata alone shifts authenticity predictions. Our benchmark provides a standardized tool for assessing and improving multimodal robustness beyond image-only settings. The code is available at https://github.com/chiuhaohao/Beyond-Visual-Forensics.

Keywords: Synthetic Medical Image Detection · Vision-Language Models · Robustness

⋆ Corresponding author: yshi4@nd.edu

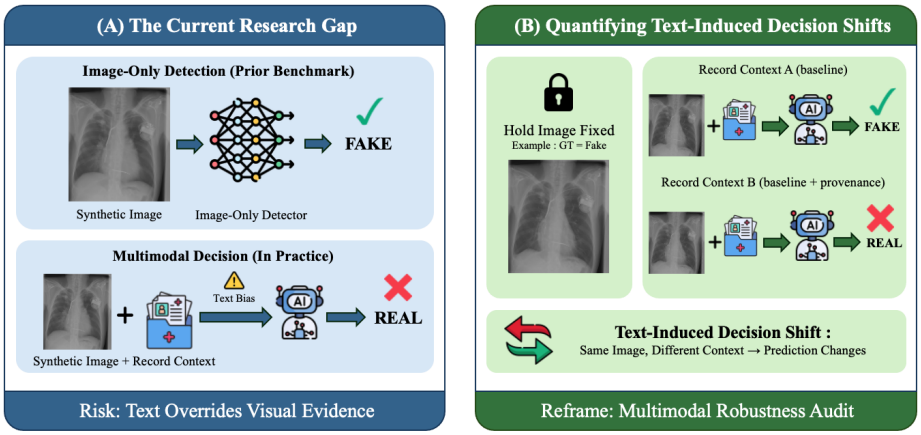

Fig. 1. Research concept: auditing multimodal robustness for synthetic medical image detection. (A) Prior benchmarks often evaluate authenticity from images alone, while real-world decisions use multimodal inputs where accompanying records can influence the verdict. **(B)** We measure **Text-Induced Decision Shift** by holding the image fixed and changing only the accompanying record (represented as structured metadata) with a controlled provenance field.

1 Introduction

Every day, medical images support high-stakes clinical and administrative decisions, and these workflows are increasingly mediated by AI systems [18]. In this setting, vision-language models (VLMs) are increasingly integrated into workflows that combine visual inputs with textual records [15, 19, 27, 3], such as report generation [10, 16] and insurance claims pipelines [5]. However, rapid advances in generative imaging [17, 13] have reshaped the risk landscape. The increasing quality and scale of synthetic medical images pose practical threats to healthcare systems [11, 12], as highly realistic forgeries can mislead even medical experts [14] and enable diagnostic deception or medical insurance fraud. In insurance claims pipelines, for example, a forged image that is mistakenly trusted may enable substantial improper payouts, while a legitimate image that is wrongly flagged may trigger prolonged investigations and delayed treatment. This raises a deployment-level challenge: as VLMs increasingly mediate high-stakes workflows, how robust are their decisions to AI-generated medical images that may enter the pipeline as inputs?

Although synthetic medical image detection has been widely studied, it is still predominantly evaluated with image-only inputs [14, 8, 1], treating visual evidence as the primary signal of authenticity. As a result, existing evaluations cannot fully characterize multimodal decision behavior in healthcare deployment, where medical images are often judged alongside clinical records and authenticity decisions are formed under joint image–record inputs.

This gap matters because a key deployment risk is not merely detecting whether an image is synthetic, but whether accompanying records can bias or steer the authenticity judgment for the same image. We refer to changes in authenticity judgment induced by altering only the accompanying record context as a **Text-Induced Decision Shift**. Recent work on VLMs shows that holding the image fixed while perturbing the accompanying text can induce substantial decision changes, and models may overly trust textual cues even when they contradict clear visual evidence [7, 29]. One might argue that an image-only detector could serve as an initial filter. However, even with such filtering, errors are unavoidable and many real workflows still require joint image–record decisions; therefore, multimodal behavior at the image–record interface must be evaluated directly. This raises our central research question: for the same image, how sensitive is a VLM's authenticity judgment to the accompanying record context, and to what extent can record context steer the final verdict when the image is unchanged? Fig. 1 summarizes this gap and previews our paired framework for quantifying text-induced decision shift.

To quantify deployment risk under joint image–record inputs, we introduce a benchmark to audit multimodal robustness in synthetic medical image detection across NIH Chest X-ray14 [25] (NIH-CXR14), ISIC2019 [23], and a private in-house pediatric chest X-ray dataset (PediCXR), spanning open-weight VLMs and frontier API models (e.g., GPT-5 [20]). Motivated by record annotations that may indicate whether documentation is manually written or AI-assisted [22, 21], we instantiate record context as structured metadata variants that differ only in a single provenance field. In real deployments, such provenance cues may be recorded in different ways, and their explicitness can vary across systems and workflows. We therefore use a controlled `Source` field with intentionally strong, explicitly stated values as a task-aligned, upper-bound stress test of how much metadata context can influence authenticity judgments. Concretely, we use `Source: Hospital` and `Source: AI-edited` (implemented as `Source: Edited by Nano Banana (AI Editing Technique)`; referred to as `AI-edited`) and keep the image fixed to test whether the provenance field can override visual grounding. We find that changing only this field can flip authenticity judgments; for example, on authentic images, adding `Source: AI-edited` reduces accuracy by **61.1%** on average across all models and datasets. Beyond quantifying this robustness risk, the benchmark also reveals models' relative reliance on image evidence versus metadata. This insight supports targeted subsequent improvements (e.g., alignment or retraining when metadata consistently dominates the decision) to improve robustness in multimodal decision-making. This work makes three contributions:

- **Benchmark method.** We introduce a paired, controllable benchmark that audits multimodal robustness in synthetic medical image detection by holding the image fixed while changing only a single provenance field in the accompanying metadata. This design serves as a task-aligned stress test to quantify text-induced decision shift and isolate how metadata context can override visual evidence in VLM-based authenticity judgments.

- **Cross-model finding.** Across multiple datasets and model families, text-induced decision shift is pervasive for both authentic and synthetic images, including in frontier API models (e.g., GPT-5).
- **Evaluation tool.** Our benchmark provides a reproducible evaluation tool to assess and compare the robustness of VLMs under multimodal inputs, supporting future model development.

2 Method

2.1 Synthetic Data Generation Pipeline

We generate paired synthetic images using an LLM-guided edit–verify–refine loop [4] with quality control, converting each authentic image x with label y to a sampled target label $y' \neq y$ while preserving medical plausibility and consistency with the original acquisition context. For each dataset, we construct a target pool $\mathcal{Y}$ using label frequencies calculated from the official evaluation split. For ISIC2019 (multi-class), $\mathcal{Y}$ contains all diagnostic classes. For multi-label datasets (NIH-CXR14 and PediCXR), $\mathcal{Y}$ includes individual labels and frequent co-occurring label sets. Given an original label y, we sample $y' \neq y$ from $\mathcal{Y}$ in proportion to its label frequency.

For each (x, y, y') triplet, we run an editing pipeline. We use Gemini-2.5-Pro [6] to produce a structured, reusable instruction template and to revise it in a consistent format, and use Gemini-2.5-Flash-Image [6] for image editing. Each edited image is then verified by Gemini-2.5-Pro via an LLM-as-judge procedure. We accept an edit only if the judge confirms (i) the target diagnosis is present, (ii) anatomy is plausible, and (iii) the image appears realistic; otherwise, we revise the instruction using structured failure feedback and retry for up to five rounds. The pipeline discards on average 1.2, 0.1, and 0.4 failed edit attempts per retained image for NIH-CXR14, ISIC2019, and PediCXR, respectively. Finally, all accepted edits undergo final review by two clinicians for CXR and dermoscopy, respectively; implausible cases are excluded through conservative quality control for plausibility, realism, and consistency with the target diagnosis.

We form **Base Metadata** by retaining essential demographics and diagnostic finding labels from the original dataset metadata, and update the label field to y' for synthetic samples. Starting from this Base Metadata, we create two **Source-Augmented Metadata** variants (**Source-H** and **Source-AI Metadata**) by appending a controlled `Source` field. This single-field `Source` intervention yields a controlled, task-aligned upper-bound stress test for multimodal authenticity judgments, isolating the impact of provenance cues.

2.2 Evaluation Framework

Task Formulation and Input Conditions. We formulate the task as synthetic medical image detection, where VLMs must determine whether a medical

image is **Real** (authentic) or **Fake** (synthetic). We evaluate each model under four input conditions: **I-Only** (image only), **I+Base** (image + Base Metadata), **I+Source-H** (image + Source-H Metadata), and **I+Source-AI** (image + Source-AI Metadata). To quantify text-induced decision shift, we perform paired judgments that hold the image fixed while swapping only the metadata context between matched variants of the same image.

Prompt and Output Format. We use a standardized forensic prompting template to ensure comparability across different architectures. Each model completes a checklist of five static visual criteria, including texture, noise patterns, edges, anatomical plausibility, and color, followed by a check for potential contradictions (within the image for image-only inputs; between visual evidence and the accompanying metadata when multimodal inputs are provided). We do not convert checklist responses into a decision; instead, the model outputs the final decision by weighing all available evidence. After the checklist, we enforce a dual verdict output format: the model outputs a `FINAL ANSWER` (a holistic judgment using all provided inputs) followed by a `VISUAL VERDICT` (a judgment based on the image alone). We use `FINAL ANSWER` as the primary output for evaluation and treat `VISUAL VERDICT` as an image-based consistency probe produced in the same generation pass.

3 Results

3.1 Experimental Setup and Datasets

To ensure that our robustness evaluation generalizes across medical domains, we construct a multimodal benchmark encompassing distinct visual semantics. We use NIH-CXR14 [25], where synthetic findings can blend into grayscale anatomy and appear highly realistic, and ISIC2019 [23], where edits in images with rich texture often leave visible artifacts. We choose these established datasets because they provide official splits and structured metadata, enabling reproducible experiments and controlled counterfactual metadata contexts. We also include PediCXR, a private in-house dataset used under institutional IRB approval, to reduce concerns that certain models may have had prior exposure to public benchmark test data.

In terms of sample selection, we sequentially iterate through the source test sets and retain cases that successfully pass the synthetic generation pipeline. For the public benchmarks (NIH-CXR14 and ISIC2019), we retain the first 500 valid samples, and for PediCXR, we retain the first 200. Following Sec. 2.1, we construct paired **Real**–**Fake** images, where **Real** is the original image and **Fake** is its AI-edited counterpart. For each paired image, we evaluate the image alone (I-Only) and pair it with three matched metadata variants (Base, Source-AI, Source-H) as described in Sec. 2.2, using each resulting input to obtain the model's authenticity judgment.

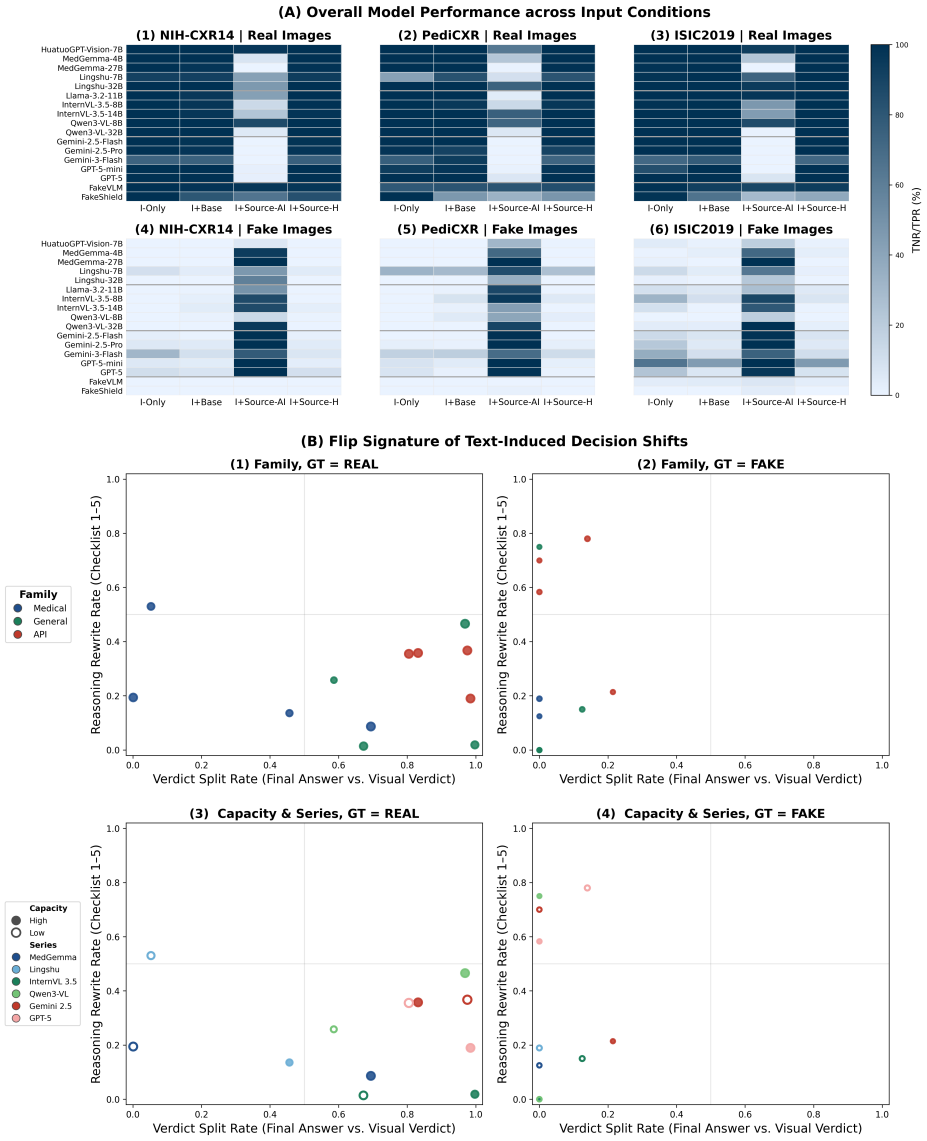

Fig. 2. Main results: text-induced decision shift and flip signatures. (A) TNR and TPR on NIH-CXR14, PediCXR, and ISIC2019 under four input conditions: I-Only, I+Base, I+Source-AI, and I+Source-H. **(B)** Flip signature scatter over flip cases (aggregated across datasets), with point size proportional to the number of flip cases. The rates are computed over flip cases and reflect failure modes rather than overall performance.

To analyze robustness across model types, we group evaluated systems into four families: open-weight medical VLMs, open-weight general VLMs, frontier API models, and DetectFake VLMs, i.e., VLMs explicitly designed for synthetic image detection (we include general domain DetectFake VLMs as baselines, since medical synthetic image detectors are often image-only and do not accept record text, making them incompatible with our multimodal setting). Medical VLMs include MedGemma [19], Lingshu [27], and HuatuoGPT-Vision [3], which are trained on biomedical corpora covering CXR and dermoscopy; general VLMs include InternVL-3.5 [24], Qwen3-VL [2], and Llama-3.2-Vision [9]. To investigate capacity effects, we compare different model capacities within the same architecture: MedGemma (4B vs. 27B), Lingshu (7B vs. 32B), Qwen3-VL (8B vs. 32B), and InternVL-3.5 (8B vs. 14B). All open-weight models are run on a compute node equipped with four NVIDIA A10 GPUs; due to resource constraints, we evaluate HuatuoGPT-Vision-7B and Llama-3.2-Vision-11B. For frontier API models, we evaluate GPT-5 [20], GPT-5-mini, Gemini-2.5-Flash [6], Gemini-2.5-Pro, and Gemini-3-Flash. For DetectFake VLMs, we evaluate FakeVLM [26] and FakeShield [28]. All models use default decoding settings from each official implementation or API.

3.2 Main Results

Text-Induced Decision Shift and Within-Family Differences. Fig. 2(A) summarizes text-induced decision shift under four input conditions. We report TNR (True Negative Rate) for Real images and TPR (True Positive Rate) for Fake images across NIH-CXR14, PediCXR, and ISIC2019 (invalid outputs counted as incorrect), with models grouped by family.

On Real images (Fig. 2(A)(1–3)), most models perform well under I-Only and remain strong under I+Base. However, appending an AI-origin provenance field to the metadata (I+Source-AI) can sharply reduce accuracy by pushing models to label authentic images as Fake. For example, on NIH-CXR14 (Real), switching from I+Base to I+Source-AI reduces MedGemma-27B from 97.6% to 0.0% and MedGemma-4B from 100.0% to 7.6%. Similar drops also appear in general VLMs, with severity varying across model series and capacity; capacity does not consistently predict robustness. For instance, Qwen3-VL-32B collapses across all three Real-image settings, whereas Qwen3-VL-8B remains substantially more robust. On the other hand, for InternVL-3.5, the 14B variant is more robust than the 8B variant on NIH-CXR14 and PediCXR, while their performance is comparable on ISIC2019. The same failure mode appears in frontier API models and DetectFake VLMs (e.g., Gemini models degrade to nearly zero under I+Source-AI), whereas others (e.g., FakeVLM) are comparatively more resistant.

On Fake images (Fig. 2(A)(4–6)), many models have low I-Only accuracy, indicating that detecting highly realistic AI-edited medical images from the image alone is challenging. Adding Base Metadata does not help: I+Base often fails to improve over I-Only and may degrade performance, suggesting that metadata without provenance field can distract the model when it is visually uncertain.

For example, on ISIC2019 (Fake), InternVL-3.5 drops from I-Only to I+Base (8B: 32.4%→10.0%; 14B: 9.6%→0.6%). In contrast, Source-AI Metadata often inflate accuracy on Fake images; given the weak I-Only and I+Base baselines, these apparent gains are best explained by reliance on the provenance field rather than improved visual detection of synthetic artifacts. Source-H Metadata often reduces or maintains performance relative to Base Metadata (e.g., PediCXR (Fake): Gemini-3-Flash, 19.0%→12.0%), suggesting that provenance field indicating a hospital source may increase false negatives when the model is uncertain about the visual evidence. DetectFake VLMs perform poorly across medical domains, likely due to domain mismatch from general domain training. Overall, on Fake images, Base Metadata and Source-H Metadata can increase false negatives, whereas Source-AI Metadata can inflate TPR via provenance cues rather than improved visual detection.

Flip Signature Analysis of Text-Induced Decision Shift. Although Fig. 2(A) shows that the provenance field can flip correctness under a fixed image, it cannot explain the mechanism, i.e., do flips arise mainly from changes in the integrated decision while the image-based judgment remains stable, or do they also involve rewriting visual reasoning? We therefore analyze **flip case**s, where the same image's `FINAL ANSWER` switches from correct to incorrect between (i) image + Base Metadata and (ii) image + Source-Augmented Metadata, and summarize each model with two rates in Fig. 2(B). Using `VISUAL VERDICT` as an image-based same-pass consistency probe, the **Verdict Split Rate** is the fraction of flip cases where `FINAL ANSWER` $\neq$ `VISUAL VERDICT` under the Source-Augmented condition. This probe distinguishes cases where metadata changes only the integrated decision from cases where metadata also shifts the model's image-based verdict. The **Reasoning Rewrite Rate** is the fraction of flip cases where the checklist changes between I+Base and the corresponding Source-Augmented input; after labeling each checklist item (1–5) as {Positive, Negative, Neutral} (supporting `REAL`, supporting `FAKE`, or uncertain), we count a rewrite if any item flips polarity across the two inputs.

In Fig. 2(B)(1), for flip cases on Real images, medical VLMs tend to show lower Verdict Split Rates, meaning `FINAL ANSWER` and `VISUAL VERDICT` more often move together when provenance changes. In contrast, general and frontier API models more often flip `FINAL ANSWER` while keeping `VISUAL VERDICT` unchanged, resulting in higher Verdict Split Rates. In the series & capacity view (Fig. 2(B)(3)), compared to smaller models, larger models often exhibit higher Verdict Split but lower Reasoning Rewrite Rates. This suggests that although the integrated decision (`FINAL ANSWER`) is shifted by the provenance field, larger models can maintain their image-based reasoning and keep their `VISUAL VERDICT` relatively stable. For Fake images in Fig. 2(B)(2) and (B)(4), Verdict Split Rates are generally low compared to Real images, indicating that `FINAL ANSWER` and `VISUAL VERDICT` usually agree. The main cross-model variation occurs in Reasoning Rewrite, with API models often higher. Capacity effects are less consistent on Fake images. Notably, points for Fake image flip cases tend to

be smaller, indicating fewer eligible flips. This is expected under our flip definition, because many models start with low baseline accuracy on Fake images, leaving fewer initially correct predictions that can be flipped by provenance field.

4 Conclusion

We propose a benchmark for auditing multimodal robustness in synthetic medical image detection by keeping the image fixed while swapping the paired metadata context. Across multiple datasets and model families, we find that changing the metadata context alone can dominate the integrated decision and flip authenticity judgments despite unchanged visual evidence. These results suggest that image-only evaluations may underestimate deployment risk under multimodal inputs. Because our provenance cues are intentionally strong, these effects represent an upper-bound stress test. This benchmark motivates future work on robust multimodal authenticity judgments, including robustness to more natural record variations and better handling of image–record conflicts.

Disclosure of Interests. M.G. and C.M. are Optum AI employees; Optum AI provided no funding. No other competing interests are declared.

References

1. Ali, A., Basha, H.A., Thanuja, K., Puneet, Gupta, S.K., Kim, S.: Enhancing tumor deepfake detection in mri scans using adversarial feature fusion ensembles. Scientific Reports (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
4. Chen, Z., Feng, M.: Med-banana-50k: A cross-modality large-scale dataset for text-guided medical image editing. arXiv preprint arXiv:2511.00801 (2025)
5. Cheng, L., Lu, J., Chan, Y.X., Nguyen, Q.K., Bi, J., Ho, S.: A hybrid architecture for multi-stage claim document understanding: Combining vision-language models and machine learning for real-time processing. arXiv preprint arXiv:2601.01897 (2026)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
7. Deng, A., Cao, T., Chen, Z., Hooi, B.: Words or vision: Do vision-language models have blind faith in text? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3867–3876 (2025)
8. Grabovski, F.M., Yasur, L., Amit, G., Mirsky, Y.: Back-in-time diffusion: Unsupervised detection of medical deepfakes. ACM Transactions on Intelligent Systems and Technology **16**(6), 1–26 (2025)

9. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
10. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. Frontiers in artificial intelligence **7**, 1430984 (2024)
11. Khosravi, B., Purkayastha, S., Erickson, B.J., Trivedi, H.M., Gichoya, J.W.: Exploring the potential of generative artificial intelligence in medical image synthesis: opportunities, challenges, and future directions. The Lancet Digital Health (2025)
12. Kondylakis, H., Osuala, R., Puig-Bosch, X., Lazrak, N., Diaz, O., Kushibar, K., Chouvarda, I., Charalambous, S., Starmans, M.P., Colantonio, S., et al.: A review of methods for trustworthy ai in medical imaging: The future-ai guidelines. IEEE Journal of Biomedical and Health Informatics (2025)
13. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2024)
14. Li, S., Xing, Z., Wang, H., Hao, P., Li, X., Liu, Z., Zhu, L.: Toward medical deepfake detection: A comprehensive dataset and novel method. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 626–637. Springer (2025)
15. Nath, V., Li, W., Yang, D., Myronenko, A., Zheng, M., Lu, Y., Liu, Z., Yin, H., Law, Y.M., Tang, Y., et al.: Vila-m3: Enhancing vision-language models with medical expert knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14788–14798 (2025)
16. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
17. Reddy, S.: Generative ai in healthcare: an implementation science informed translational path on application, integration and governance. Implementation Science **19**(1), 27 (2024)
18. Ryu, J.S., Kang, H., Chu, Y., Yang, S.: Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomedical Engineering Letters **15**(5), 809–830 (2025)
19. 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)
20. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
21. Stults, C.D., Deng, S., Martinez, M.C., Wilcox, J., Szwerinski, N., Chen, K.H., Driscoll, S., Washburn, J., Jones, V.G.: Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA network open **8**(5), e258614 (2025)
22. Tai-Seale, M., Baxter, S.L., Vaida, F., Walker, A., Sitapati, A.M., Osborne, C., Diaz, J., Desai, N., Webb, S., Polston, G., et al.: Ai-generated draft replies integrated into health records and physicians' electronic communication. JAMA Network Open **7**(4), e246565 (2024)

23. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 180161 (2018)
24. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
26. Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., Li, W.: Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv preprint arXiv:2503.14905 (2025)
27. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
28. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., Zhang, J.: Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. arXiv preprint arXiv:2410.02761 (2024)
29. Zhang, C., Ding, W., Liu, J., Wu, M., Wu, Q., Mooney, R.: Do images speak louder than words? investigating the effect of textual misinformation in vlms. arXiv preprint arXiv:2601.19202 (2026)