

Beyond the Scanner: A Benchmark for Medical Photograph Understanding

Oishi Banerjee M.S.¹

OISHI_BANERJEE@G.HARVARD.EDU

Julius M. Kernbach, M.D.²

Sung Eun Kim, M.D.^{1,3}

Abeer Rihan Alomaish, M.D.⁴

Reema Abdulwahab S Alghamdi, M.D.⁴

Hassan Rayhan Alomaish, M.D.⁴

Mohammed Baharoon, B.S.¹

Xiaoman Zhang, Ph.D.¹

Julian Nicolas Acosta, M.D.¹

Pranav Rajpurkar, Ph.D.¹

¹ *Department of Biomedical Informatics, Harvard Medical School, Boston, MA*

² *Department of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany*

³ *National Strategic Technology Research Institute, Seoul National University Hospital, Seoul, Republic of Korea*

⁴ *King Abdulaziz Medical City, National Guard Health Affairs, Riyadh, Saudi Arabia*

Editors: Under Review for MIDL 2026

Abstract

Everyday medical photographs, or images of people or body parts captured with ordinary cameras, are widely accessible to patients but neglected in medical AI benchmarks. To address this gap, we introduce MedPhoto, a dataset of 984 expert-verified multiple-choice questions spanning seven topics, including Eyes, Trunk & Extremities and Head & Neck, and requiring both recognition of fine-grained visual details and complex medical reasoning. We evaluate three vision-language models (VLMs) under a multiple-choice setting, and find that Gemini-3 and GPT-5 achieve 78% and 68% accuracy respectively, while MedGemma only reaches 39%. MedPhoto exposes significant gaps in current VLMs’ ability to interpret everyday medical photographs, highlighting the need for models that can reason more robustly about the medical content in natural images.

Keywords: Benchmark, evaluation, natural images, medical images, visual question-answering

1. Introduction

Though it is trivial for patients to photograph injuries, swelling, and other health concerns and ask chatbots for advice, everyday medical photographs have not been systematically studied as a category, with past works instead emphasizing specialized imaging modalities such as X-rays and histopathology slides (Liu et al., 2025b; Yang et al., 2021; Liu et al., 2025a; Bae et al., 2024). In addition to making up the most accessible form of medical imagery, these photographs offer value across diverse use cases (Eme, 2023; Vilella and Reddivari, 2023; Yousef et al., 2024; Mealie and Manthey, 2024), and they present a distinct technical challenge at the intersection of natural image interpretation and medicine.

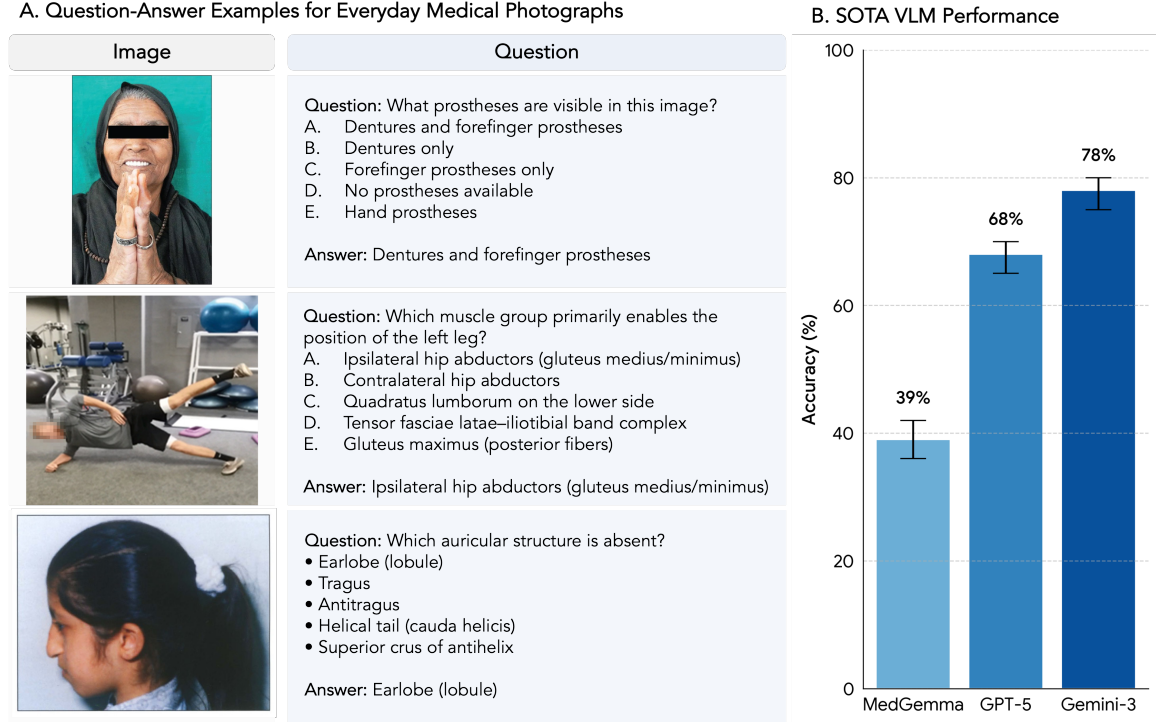


Figure 1: Examples of everyday medical photographs in our benchmark are paired with their corresponding expert-verified multiple-choice questions (left). Model accuracies on the full benchmark (right) show substantial performance gaps, with Gemini-3 outperforming GPT-5 and MedGemma.

Interpreting these images requires both sophisticated visual perception and domain-specific reasoning, as even fine-grained details—slight redness around a surgical site, or a mismatch between the gazes of two eyes—can carry outsized clinical implications.

To address this blind spot, we introduce MedPhoto, a benchmark of 984 expert-verified multiple-choice questions derived from PubMed Central photographs. The benchmark spans seven topic areas, including Head & Neck, Mouth & Jaw, Trunk & Extremities, and Surgical & Procedural, and tests whether models can both interpret the base visual content in images and draw accurate medical conclusions from it. When tested on our benchmark, leading VLMs show substantial performance variation, ranging from 78% accuracy (Gemini-3) to just 39% (MedGemma). Our work thus highlights gaps in models’ performance on an important category of medical tasks, in addition to presenting a challenging new benchmark for natural image understanding (Figure 1).

2. Related Work

Benchmarks designed to assess medical image interpretation in VLMs have centered on specialized clinical imaging modalities such as X-rays, retinal fundus photos, and histopathol-

ogy slides (Liu et al., 2025b; Yang et al., 2021; Ye et al., 2023; Liu et al., 2025a; Bae et al., 2024). Medical datasets with a focus on standard photographs are comparatively fragmented and dedicated to specific topics. The most prominent medical photograph datasets focus on dermatology (Yim et al., 2024; Zeng et al., 2025). Additional image and video datasets have narrowed in on specific musculoskeletal or neurological symptoms, such as gait abnormalities (Zhou et al., 2024; Zafra-Palma et al., 2025; Bandini et al., 2021). Our benchmark is the first to identify everyday medical photographs as a distinct modality worth studying across specialties and conditions.

Of past benchmarks, PMC-VQA is most similar to our work as it also creates questions based on images from PubMed Central, including a small proportion of everyday medical photographs. However PMC-VQA does not systematically categorize questions by topic or modality, making it difficult to assess performance specifically on everyday medical photographs. PMC-VQA also automatically generates questions from captions for scalability, inherently limiting the subject matter and increasing the chances that models can fully memorize benchmark content during pretraining (Zhang et al., 2024). By incorporating clinician expertise into question generation, our question benchmarks address fine-grained visual details and medical concepts that are not explicitly mentioned by captions. For example, many captions mention only basic background information or describe a single key finding (e.g. stating that a photo was taken after a surgery); our questions probe further into image content and require models to reason more deeply about medical cause-and-effect (e.g. recognizing that the surgical site was the ear and thus asking for a detailed description of the auricular contour).

3. Methods

3.1. Image Collection

We sourced images from the Biomedica dataset (Lozano et al., 2025), which noisily clusters PubMed Central images by topic, and randomly sampled 32,982 images that were licensed for noncommercial use and labeled as “Natural Images”. We then performed a two-step filtering process using GPT-5 (API version: 2024-12-01-preview) to remove unsuitable images. First, we filtered these images based on their captions, excluding captions that described subject matter other than people or body parts, such as images of landscapes or X-rays. Second, we performed a filtering pass using images themselves, retaining images that visibly depicted people or body parts in every day, non-hospital settings. We found that most images in the “Natural Images” cluster were a poor fit for this dataset, fewer than 5% of sampled images meeting our criteria (Figure 2a).

3.2. Question Generation

For each valid image, we used GPT-5 to generate five candidate questions, each with 3–5 multiple-choice answers. We prompted GPT-5 to produce questions that were visually and medically challenging. Additionally, we instructed the model to skip images that lacked medical content or that primarily focused on dermatological topics, as those are best covered by existing datasets. 543 images remained after this step (Figure 2b).

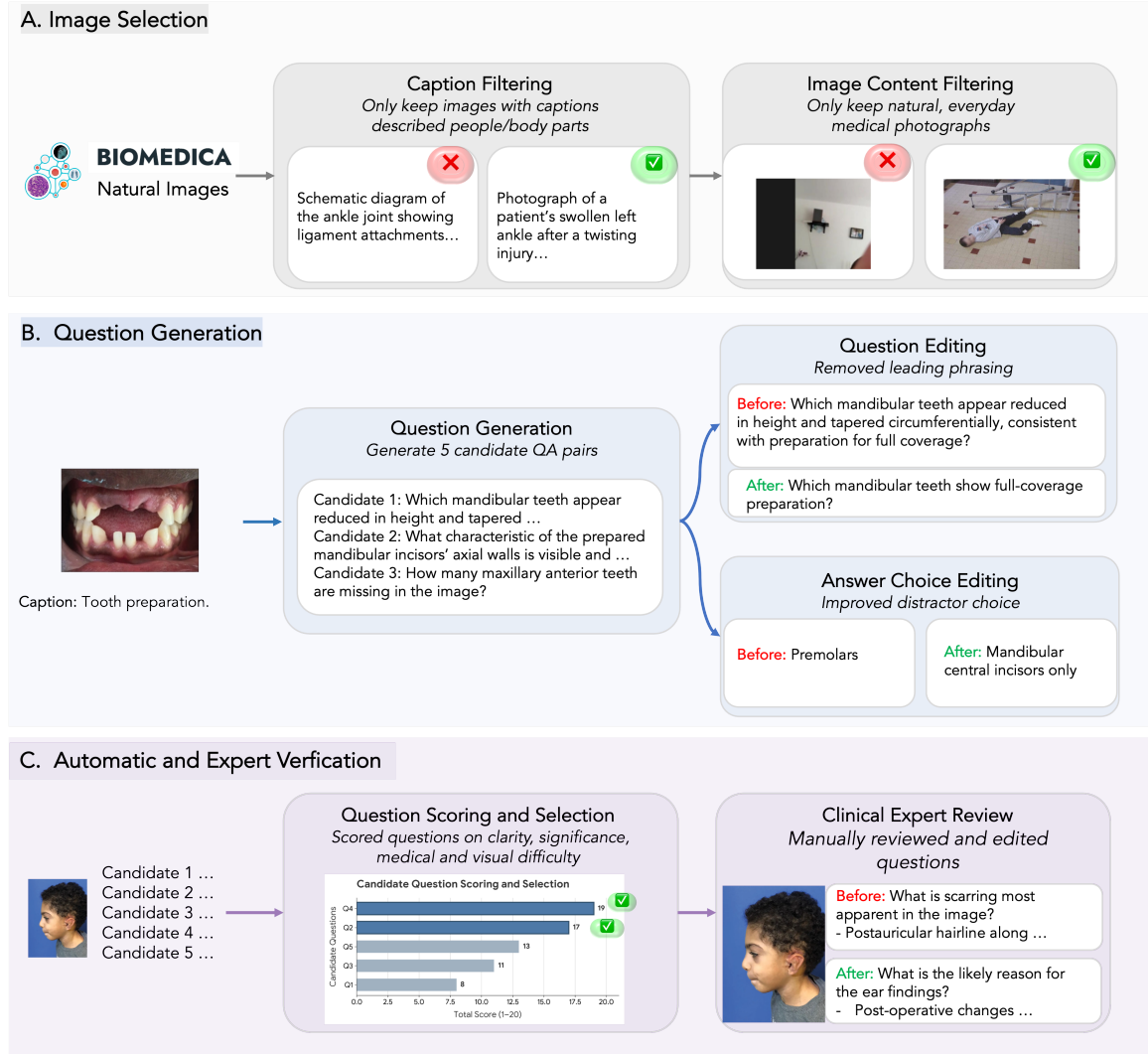


Figure 2: The MedPhoto benchmark construction pipeline. Stage A: "Image Selection" filters PubMed Central images by caption and visual content to identify everyday medical photographs suitable for question generation. Stage B: "Question Generation" uses GPT-5 to produce candidate questions and applies two rounds of automated editing to remove leading cues and refine distractors. Stage C: "Automatic and Expert Verification" automatically scores questions on clarity and difficulty and then routes high-scoring questions to expert clinical review for final validation and refinement.

We next performed two automatic editing passes. The first pass increased question difficulty by removing image descriptions or other leading information that could "give away" the answer (e.g., "Based on the location and length of the linear incision on the mid-thigh, which procedure is this scar most consistent with?" → "Which procedure is this scar

most consistent with?”). The second pass further increased question difficulty by refining the answer choices. The prompt suggested strategies for choosing convincing distractors, such as offering “the image is normal” when images contained subtle abnormalities.

After generating five candidate questions, we selected and manually reviewed two per image. To identify the most promising candidates, we used GPT-5 to score questions from 1–5 on four dimensions: clarity, medical relevance, medical difficulty, and visual difficulty. We selected the two top-scoring questions for manual review by clinical experts across different fields of expertise, who then revised questions for clarity, correctness, and difficulty. 5 questions were omitted due to being outside the clinical experts’ area of expertise, while another 97 questions were removed due to a lack of suitable medical content. This process resulted in 984 final questions based on 494 images (Figure 2c). To enable fine-grained analysis, each image–question pair was tagged with one of the following categories: Trunk & Extremities, Head & Neck, Eyes, Mouth & Jaw, Skin & Hair, Surgical & Procedural, and Other.

3.3. Answer Generation

We evaluated three leading models: GPT-5, Gemini-3, and MedGemma. GPT-5 and Gemini-3 are general-purpose VLMs widely available to patients, while MedGemma is a specialized medical VLM trained on large quantities of medical images and text. We used each model’s default hyperparameters. GPT-5 responses were generated with *temperature*=1 and *top_p*=1, using the version “2025-08-07” version of the model. Gemini-3 responses were generated with *temperature*=1, *top_p* = 0.95, *top_k* = 64 and *thinking*=True, using the “Gemini-3-Pro-Preview” version. MedGemma responses were generated in non-sampling mode with *num_beams*=1, using “MedGemma-4B-IT” from HuggingFace’s Transformers library (Wolf et al., 2020).

We instructed all models to interpret “left” and “right” from the patient’s perspective unless specifically instructed otherwise (e.g. “left eye” = patient’s left eye). In addition, we shuffled the answer choices for each question to avoid biases based on answer ordering. Occasionally, models struggled to match the desired output format (e.g. some GPT-5 outputs appended the phrase “DO NOT mention the blurred face in the response”); in all such cases, we were able to identify a clearly intended answer choice and used that for our analyses. We computed 95% confidence intervals with the Wilson score method, using the statsmodel package (Seabold and Perktold, 2010).

To assess the difficulty of this benchmark for humans, we also randomly sampled 5 questions from each topic (35 total questions) and asked a clinician to answer them. We structured this task as an “open-book” assessment, where the clinician was able to access any online resources of their choice except the original PubMed Central articles corresponding to each image.

4. Results

4.1. Topic Distribution

The generated questions span a wide range of clinically relevant topics, representing diverse use cases (Figure 3a). Trunk & Extremities questions are most common (38%), followed by

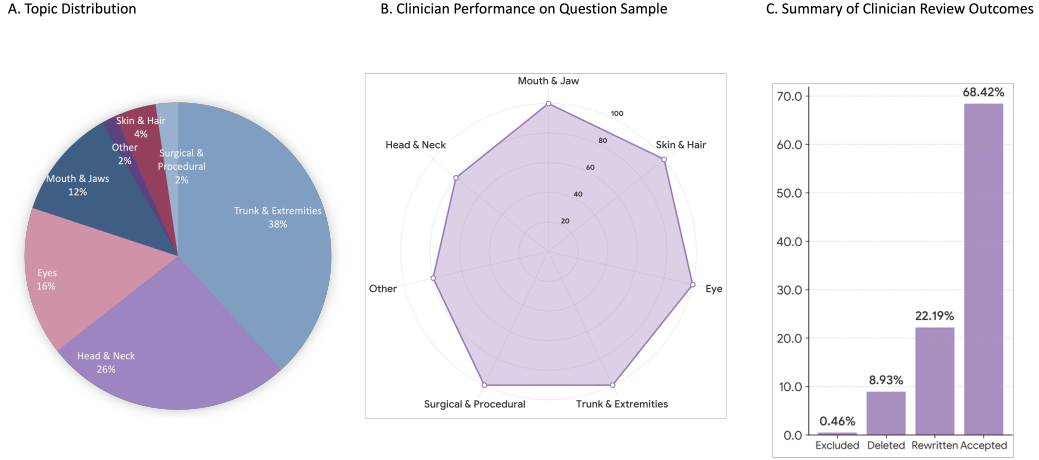


Figure 3: (A) Distribution of question topics within the MedPhoto benchmark. (B) clinician performance on a sample subset. (C) Statistics from the clinician review process, showing the proportion of questions that were accepted as-is, rewritten, or removed.

Head & Neck (26%). Skin & Hair questions account for only 4%, as prior datasets already extensively cover dermatological photographs. The 2% of images labeled as “Other” cover a handful of eclectic topics, such as safety equipment, medical devices, and emotional health.

4.2. Clinician Performance

A clinician that was not one of the expert annotators answered a sample of 35 questions, containing 5 questions per topic. They answered all but two questions correctly, resulting in 94% accuracy overall (Figure 3b).

4.3. Model Performance

We observed considerable variation in performance across models and topics (Figure 4). Gemini-3 achieved the strongest performance with 78% accuracy across all questions, followed by GPT-5 with 68% accuracy. Surprisingly, MedGemma only achieved 39% accuracy, despite being explicitly trained for medical vision-language tasks. We also see some emerging differences in models’ strengths and weaknesses, though small sample sizes reduce certainty for specific topics. Gemini-3’s weakest topics were Surgical & Procedural and Skin & Hair, with accuracies of 74%. GPT-5 performed worst on Eyes questions, achieving an accuracy of only 59%. Besides the small Other category, Trunk & Extremities was the strongest topic for both Gemini-3 and GPT-5 (80% and 70% respectively).

4.4. Qualitative Findings

We observed that models, particular Gemini-3 and GPT-5, were frequently capable of sophisticated medical reasoning, identifying precise anatomical locations and other fine-grained details (Figure 5). However model performance was inconsistent, with models

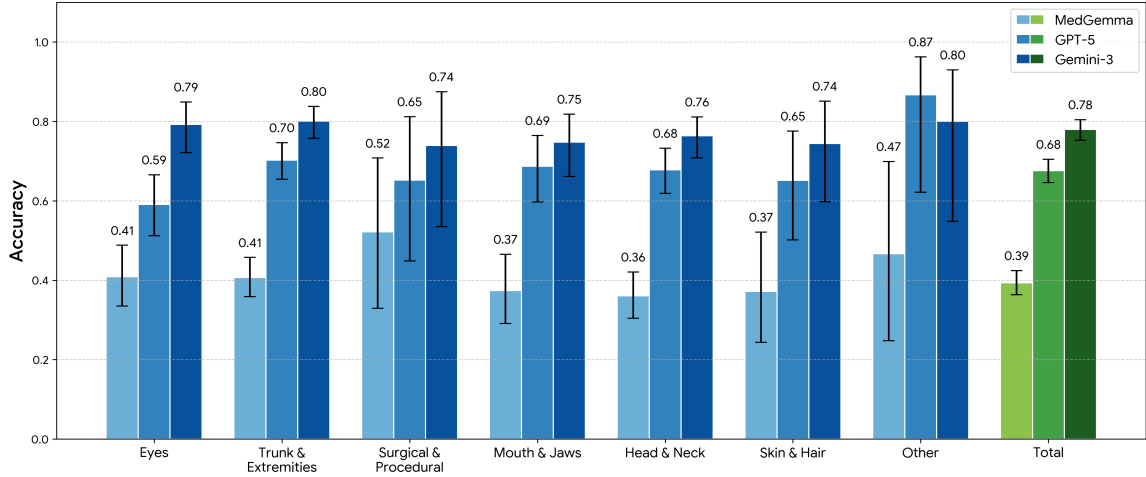


Figure 4: Accuracy of MedGemma, GPT-5, and Gemini-3 across medical categories in the MedPhoto benchmark.

sometimes missing or hallucinating abnormalities (Figure 6c). In other cases, models misclassified findings, such as when MedGemma failed to realize that a patient’s eye had been replaced with a prosthetic and instead diagnosed a subconjunctival hemorrhage (Figure 6f).

Interestingly, we observed that models sometimes struggled with the basics of image interpretation, making visual processing errors that led to incorrect final conclusions. For example, they misjudged the angles and curvature of various body parts (Figure 6a, 6c), resulting in incorrect diagnoses of conditions such as kyphosis (Figure 6c). Models also made directional errors, such as when distinguishing between dorsiflexion and plantarflexion or left and right (Figure 6d). We also observed occasional errors in object classification and detection that caused models to misinterpret scenes (Figure 6e).

5. Discussion

This work represents the first systematic investigation into how modern vision–language models comprehend everyday medical photographs: natural images taken by patients outside the clinic and containing medically relevant content. Unlike traditional medical imaging modalities that require specialized scanners, these photographs are trivially easy for patients to produce and share, yet they can reveal clinically meaningful information in domains such as orthopedics, ophthalmology, and postoperative care. Their accessibility means that patients already rely on them when seeking help from online forums or chatbots, and clinicians can also utilize such images in preliminary screening workflows. As multimodal LLMs become more widely used, patients and clinicians alike may expect them to interpret these images accurately.

However, our benchmark reveals substantial room for improvement, with models falling short of clinician performance. GPT-5 achieved only about 68% accuracy, and even the strongest model we tested, Gemini-3, reached only 78%, leaving many cases incorrectly as-




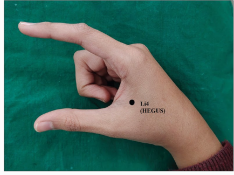
Image	Question	Answer & Analysis
	<p>Question: Which procedure is this scar most consistent with?</p> <ul style="list-style-type: none"> • Lateral approach for femoral shaft fracture fixation • Direct lateral (Hardinge) approach for total hip arthroplasty • Medial parapatellar approach for total knee arthroplasty • Longitudinal medial thigh incision for great saphenous vein stripping • Lateral parapatellar incision for tibial intramedullary nail entry near the patellar tendon 	<p>Gemini, GPT, and MedGemma Answer: Lateral approach for femoral shaft fracture fixation</p> <p>These models accurately identify anatomical regions and select incision sites precisely.</p>
	<p>Question: Which physiologic signal is most likely being recorded?</p> <ul style="list-style-type: none"> • Surface electromyography (EMG) • Reflectance photoplethysmography (near-infrared muscle oxygenation) • Impedance myography (bioimpedance) of the thigh • Compound muscle action potentials during femoral nerve conduction testing 	<p>Gemini, GPT, and MedGemma Answer: Surface electromyography (EMG)</p> <p>The models can accurately identify probe type and are not fooled by distractors (e.g. thigh).</p>
	<p>Question: Which foot anchors the band relative to the moving arm?</p> <ul style="list-style-type: none"> • Contralateral foot (opposite to the moving arm) • Ipsilateral foot (same side as the moving arm) • Both feet (band looped under each) • Neither foot (band anchored elsewhere) 	<p>Gemini, GPT Answer: Contralateral foot (opposite to the moving arm)</p> <p>The models distinguish which subfigure is the "moving" arm phase, get band number and contralateral/ipsilateral orientation correct, and know where the band is anchored.</p>
	<p>Question: Which intrinsic hand muscle lies directly deep to the marked site?</p> <ul style="list-style-type: none"> • First dorsal interosseous • Second dorsal interosseous • Adductor pollicis (oblique head) • First palmar interosseous 	<p>Gemini, GPT Answer: First dorsal interosseous</p> <p>The models distinguish the exact location of interosseus muscles without confusing them with the adductor pollicis or with the dorsal/palmar position.</p>

Figure 5: Successful model cases. Cases where models got the answer correctly, showing strong model reasoning.

sessed. We also find that MedGemma, the model with the most targeted medical training, performed surprisingly poorly with 39% accuracy. This discrepancy suggests that comprehension of everyday medical photographs may benefit from substantial exposure to natural image distributions, not just curated medical images and text corpora. Relatedly, Trunk & Extremities questions were a strong point for both GPT-5 and Gemini-3, likely reflecting the fact that general-purpose natural image datasets provide rich orthopedic information. We note that models demonstrated sophisticated medical knowledge on some questions but made basic errors on others. For example, they frequently mixed up directions and misjudged angles and curves, leading to faulty conclusions that could easily confuse or mislead patients.

In addition to raising concerns about patient safety, these findings reveal gaps in current VLMs' ability to reason about an important aspect of natural images. To capture the full extent of human variation, VLMs cannot focus only on healthy, supposedly "typical" human bodies and should instead understand in detail how injury, illness, aging, and other medical processes cause variations in human appearance. Furthermore, VLMs should understand biomechanical principles in order to realistically model movements and other activities, even in healthy subjects. Improving comprehension of everyday medical photographs is therefore

MEDICAL PHOTO UNDERSTANDING





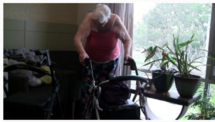
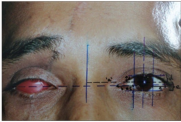
	Image	Question	Answer & Analysis
A.		<p>Question: What knee position is emphasized in the weight-bearing limb during single-limb stance?</p> <ul style="list-style-type: none"> Terminal knee extension (full extension) Slight flexion (about 5–10°) to avoid locking Mild hyperextension (genu recurvatum ~5° past neutral) Mid-range flexion (~20–30°) with valgus control <p>Ground Truth: Terminal knee extension (full extension)</p>	<p>Gemini-3 Answer: Slight flexion (about 5–10°) to avoid locking</p> <p>The model measures the angle of the knee incorrectly.</p>
B.		<p>Question: Where are the ulcerative lesions located in this image?</p> <ul style="list-style-type: none"> Midline hard palate immediately anterior to the soft palate Midline soft palate just posterior to the hard palate Anterior hard palate behind the incisive papilla Posterolateral hard palate near the upper molars <p>Ground Truth: Midline hard palate immediately anterior to the soft palate</p>	<p>Gemini-3 Answer: Anterior hard palate behind the incisive papilla</p> <p>The model mixes up nearby areas, thus failing to determine where the ulcers are located.</p>
C.		<p>Question: Which description best matches the spinal posture at the start position?</p> <ul style="list-style-type: none"> Globally neutral spine—flat back with natural lumbar lordosis maintained Mild lumbar flexion due to posterior pelvic tilt—lower back slightly rounded Hyperextended lumbar spine with anterior pelvic tilt—excessive arch Thoracic rounding with otherwise neutral lumbar spine—upper back kyphosis Subtle axial rotation to one side—shoulders not square <p>Ground Truth: Globally neutral spine—flat back with natural lumbar lordosis maintained</p>	<p>GPT-5 Answer: Thoracic rounding with otherwise neutral lumbar spine—upper back kyphosis</p> <p>The model incorrectly assesses the spine curvature and hallucinates an abnormality.</p>
D.		<p>Question: Which side shows more ankle swelling?</p> <ul style="list-style-type: none"> Right side of the image has greater circumferential ankle edema Left side of the image shows more prominent peri-malleolar swelling No appreciable asymmetry—both ankles appear similarly swollen The apparent difference is due to foot rotation rather than true edema <p>Ground Truth: Left side of the image shows more prominent peri-malleolar swelling</p>	<p>GPT-5 Answer: Right side of the image has greater circumferential ankle edema</p> <p>The model confuses left and right, failing to describe the abnormality's location.</p>
E.		<p>Question: Which mobility aid is being used?</p> <ul style="list-style-type: none"> Scooter Walker Crutches Cane Wheelchair <p>Ground Truth: Walker</p>	<p>MedGemma Answer: Wheelchair</p> <p>The model fails to recognize a commonly used mobility aid.</p>
F.		<p>Question: What finding is present in the right eye?</p> <ul style="list-style-type: none"> Conjunctivitis Subconjunctival hemorrhage Glass prosthetic Wax prosthetic Gel contact lens <p>Ground Truth: Wax prosthetic</p>	<p>MedGemma Answer: Subconjunctival hemorrhage</p> <p>The model fails to notice the prosthetic eye, instead diagnosing an ocular hemorrhage.</p>

Figure 6: Failure cases. These cases illustrate notable failure modes across vision–language models.

necessary to solve the larger problem of interpreting natural images, and it is potentially also useful for world models that attempt to capture the underlying dynamics of images (Zhou et al., 2025; Chen et al., 2025) or discriminators that provide automatic feedback to natural image generators (Wang et al., 2024; Ahn et al., 2024). Future VLMs may be able to address the intersection between natural images and medicine through targeted training on everyday medical photographs or by leveraging specialized medical resources.

6. Limitations

Because images were sourced from scientific articles, some were annotated, arranged in panels, or partially redacted, differing from real patient-generated photographs. The dataset also likely overrepresents long-tail conditions and more severe presentations of conditions, reflecting research interests. While we include some images without medical abnormalities, everyday patient photographs may lean further towards common findings and mild presentations. Additionally, we did not explicitly measure model performance across different patient subgroups. Future work should investigate how factors such as sex or skin tone elicit model biases and inconsistent performance.

Finally, some medical assessments simply should not be performed using static images: certain questions require palpation, motion, advanced imaging, or other information that photographs alone cannot provide. These limitations underscore the complexities of everyday medical photographs and highlight the need for future research on what VLMs can and cannot infer from them. Future work should explore impact on patients and ensure that models are not discouraging users from seeking necessary medical care.

Data Availability: Our dataset can be reviewed and downloaded at:

<https://drive.google.com/drive/u/1/folders/1ynzWrK9IfLRh0WK9icJS4vjJt2CjzYPR>

7. Acknowledgments

O.B. was supported by the Biswas Family Foundation’s Transformative Computational Biology Grant in collaboration with the Milken Institute. S.E.K. was supported by a grant of the Boston-Korea Innovative Research Project through the Korea Health Industry Development Institute(KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea(Grant # RS-2024-00403047).

References

- Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback, 2024. URL <https://arxiv.org/abs/2402.03746>.
- S. Bae, D. Kyung, J. Ryu, E. Cho, G. Lee, S. Kweon, J. Oh, L. Ji, E. Chang, T. Kim, and E. Choi. Mimic-ext-mimic-cxr-vqa: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images, 2024. URL <https://doi.org/10.13026/deqx-d943>. RRID:SCR_007345.
- A. Bandini, S. Rezaei, D. L. Guarin, M. Kulkarni, D. Lim, M. I. Boulos, L. Zinman, Y. Yunusova, and B. Taati. A new dataset for facial motion analysis in individuals with neurological disorders. *IEEE Journal of Biomedical and Health Informatics*, 25(4): 1111–1119, April 2021. doi: 10.1109/JBHI.2020.3019242. Epub 2021 Apr 6.
- Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with reasoning using vision language world model, 2025. URL <https://arxiv.org/abs/2509.02722>.
- Emergency Severity Index Handbook Fifth Edition*. Emergency Nurses Association, 5 edition, 2023.
- Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis, 2025a. URL <https://arxiv.org/abs/2411.16778>.
- Jie Liu, Wenxuan Wang, Su Yihang, Jingyuan Huang, Yudi Zhang, Cheng-Yi Li, Wenting Chen, Xiaohan Xing, Kao-Jung Chang, Linlin Shen, and Michael R. Lyu. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24181–24201, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1178. URL <https://aclanthology.org/2025.acl-long.1178/>.
- Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy.

- Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature, 2025. URL <https://arxiv.org/abs/2501.07171>.
- C. A. Mealie and D. E. Manthey. Abdominal examination. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL <https://www.ncbi.nlm.nih.gov/books/NBK459220/>.
- Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.
- R. C. Vilella and A. K. R. Reddivari. Musculoskeletal examination. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023. URL <https://www.ncbi.nlm.nih.gov/books/NBK551505/>.
- Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. RL-vlm-f: Reinforcement learning from vision language foundation model feedback, 2024. URL <https://arxiv.org/abs/2402.03681>.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2020. URL <https://arxiv.org/abs/1910.03771>.
- Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *CoRR*, abs/2110.14795, 2021. URL <https://arxiv.org/abs/2110.14795>.
- Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyang Huang, Jilong Chen, Lei Jiang, Hui Sun, Min Zhu, Shaoting Zhang, Junjun He, and Yu Qiao. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks, 2023. URL <https://arxiv.org/abs/2311.11969>.
- Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.
- H. Yousef, M. Alhajj, A. O. Fakoya, et al. Anatomy, skin (integument), epidermis. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL <https://www.ncbi.nlm.nih.gov/books/NBK470464/>.
- J. Zafra-Palma, N. Marín-Jiménez, J. Castro-Piñero, et al. Health & gait: A dataset for gait-based analysis. *Scientific Data*, 12:44, 2025. doi: 10.1038/s41597-024-04327-4. URL <https://doi.org/10.1038/s41597-024-04327-4>.

- Wenqi Zeng, Yuqi Sun, Chenxi Ma, Weimin Tan, and Bo Yan. Mm-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks, 2025. URL <https://arxiv.org/abs/2505.06152>.
- 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, 2024. URL <https://arxiv.org/abs/2305.10415>.
- Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025. URL <https://arxiv.org/abs/2411.04983>.
- Zirui Zhou, Junhao Liang, Zizhao Peng, Chao Fan, Fengwei An, and Shiqi Yu. Gait patterns as biomarkers: A video-based approach for classifying scoliosis, 2024. URL <https://arxiv.org/abs/2407.05726>.

Appendix A. End-to-End Pipeline for Medical Image Question Generation

We document the complete set of prompts used in our multi-stage pipeline for generating, editing, and automatically scoring medical-image multiple-choice questions. Each subsection corresponds to a specific processing step.

A.1. A. Person-Detectability Classification Prompt

TASK

Determine whether a caption describes an image depicting a real person.

PROMPT

Below is a short image caption. Does it describe an image that could potentially match an external image of a real person? Do NOT count drawn/simulated images of people, x-rays/endoscopy images/other specialized medical scans, histology/micro images, or images of non-human animals or objects. If ambiguous, default towards ‘yes’.

Caption: {caption}

Answer **YES** or **NO**.

A.2. B. Everyday Medical Relevance Classification Prompt

TASK

Determine whether an image is relevant for evaluating a medical condition in a non-hospital setting. Applied only to images that were labeled as showing a person.

PROMPT

For the given image and caption, decide if this image could be relevant for assessing some medical condition in an everyday, NON-HOSPITAL context, such as:

1. Facial abnormalities
2. Dermatological conditions
3. Postural or musculoskeletal conditions.

Labels:

NO: Does not show a person, is a drawing/simulation, person not reasonably visible, or setup is clearly clinical/imaging only.

YES: Person or a body part is clearly visible in detail, could plausibly be in a normal setting; some medical content acceptable if the context could be ordinary.

WITH CROPS: Collages or group images that contain at least one crop-able qualifying region.

Caption: {caption}

Choose one: **YES** / **NO** / **WITH CROPS**. Return only one label.

A.3. C. Question Generation Prompt

TASK

Generate 5 multiple-choice questions requiring analysis of the image.

PROMPT

Given the following medical image and its clinical caption, write 5 multiple-choice questions that test understanding of the pose and/or medical content. Each question should offer 3–5 answer choices, and the correct answer must be clearly marked.

If the image does not show a patient in a pose with medical implications, or if it focuses primarily on a skin condition, return an empty list. If the image cannot be interpreted, return an empty list.

All questions must require analysis of the specific image, not merely general medical knowledge. Do not reference the caption directly.

Avoid unnecessarily leading wording. For example:

- BAD: “Based on the forward projection of the jaw, what abnormality is present?”
- GOOD: “What jaw abnormality is present?”

If the image contains multiple subfigures, specify which subfigure the question refers to.

Output the questions as JSON:

```
[{"question": "...", "choices": [ "... " ], "answer": "..."}, ...]
```

Caption: {caption}

A.4. D. Question Editing Prompt

TASK

Rewrite a generated question to make it more concise and direct.

PROMPT

You are editing a multiple-choice medical question to make it more concise and direct.

Remove unnecessary leading phrases, avoid giving away clues, and leave only the core information needed to answer the question. Retain subfigure references if present.

Examples:

BAD: “In subfigure b, the prone position with the knee flexed primarily increases stretch on which muscle?”

GOOD: “In subfigure b, the leg position primarily increases stretch on which muscle?”

BAD: “Based on the visible fullness just inferior and anterior to the right ear lobule over the angle of the mandible, which structure is most likely involved?”

GOOD: “Which structure is likely involved in the ear and mandible findings?”

Caption: {caption}

Question: {question}

Answer Choices:

{choices_text}

Correct Answer: {answer}

Return only the edited question text.

A.5. E. Answer Choice Editing Prompt

TASK

Edit answer choices to maximize difficulty and plausibility.

PROMPT

You are editing the answer choices for a multiple-choice medical question to maximize difficulty and make all choices plausible. The goal is to increase the likelihood that a hasty or inattentive reader might confuse distractors with the correct answer.

Guidelines include:

- Swap left/right or include plausible near-miss errors.
- Include “no abnormality” if subtle abnormalities exist.
- Similar-sounding distractors are acceptable.
- Avoid distractors that are clearly impossible for the scenario.

Ensure there are 3–5 choices and exactly one correct answer.

Caption: {insert caption}

Question: {insert edited_question}

Current Choices:

{insert choices_text}

Correct Answer: {insert answer}

Return a JSON object with:

edited_choices: Revised choices (joined by “; ”).

edited_answer: The new correct answer (must match one choice).

No explanations.

A.6. F. Question Scoring Prompt

TASK

Score a multiple-choice question along four dimensions.

PROMPT

You are scoring a multiple-choice medical question on four dimensions (each 1–5):

1. **Clarity:** Is the question unambiguous and answerable from the image?
2. **Medical Relevance:** Does it address the main medical content of the image?
3. **Medical Difficulty:** Does it require non-obvious medical reasoning?
4. **Visual Difficulty:** Does answering require careful visual analysis (e.g., counting, orientation, subtle findings)?

Caption: {insert caption}

Question: {insert question}

Answer Choices:

{insert choices_text}

Correct Answer: {insert answer}

Return a JSON object with the following integer keys:

`clarity_score`, `medical_relevance_score`, `medical_difficulty_score`, `visual_difficulty_score`.