MEDLESIONVQA: A MULTIMODAL BENCHMARK EMULATING CLINICAL VISUAL DIAGNOSIS FOR BODY SURFACE HEALTH

Anonymous authors
Paper under double-blind review

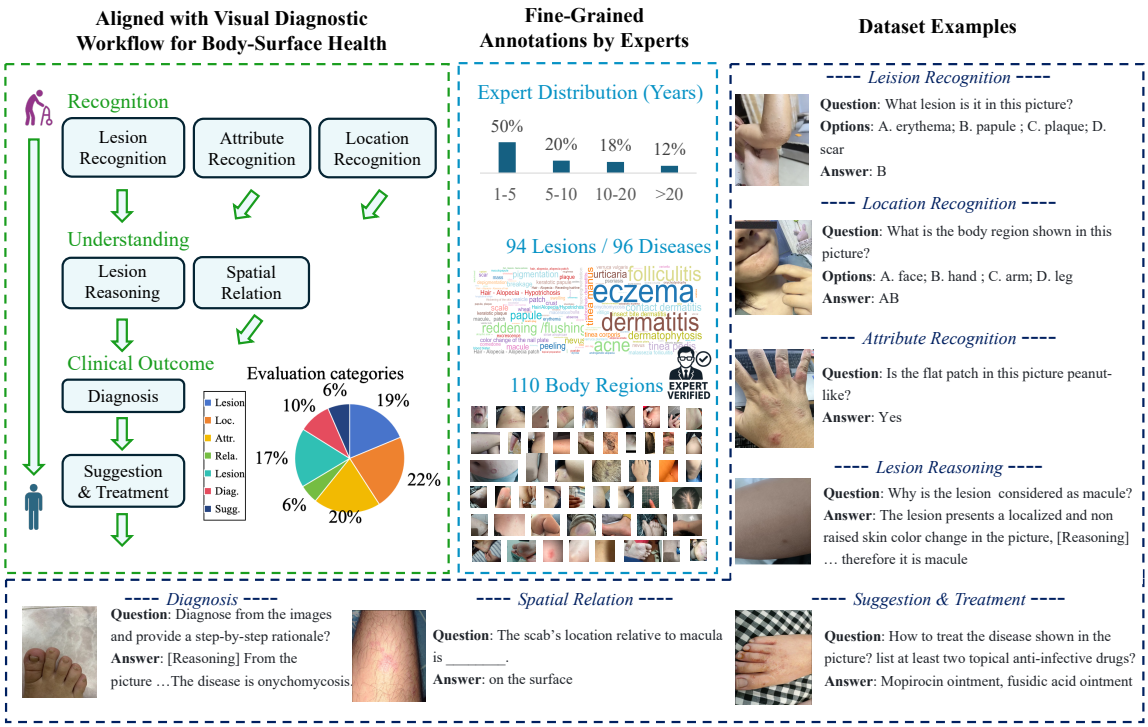


Figure 1: Overview of MedLesionVQA. The benchmark is designed to emulate the visual diagnostic workflow of physicians (top-left), covering seven core abilities with fine-grained annotations. Expert physicians with over 20 years of experience validated annotations (middle), which include detailed identification of 94 lesion types, 96 diseases, and 110 body regions (bottom).

ABSTRACT

Body-surface health conditions, spanning diverse clinical departments, represent some of the most frequent diagnostic scenarios and a primary target for medical multimodal large language models (MLLMs). Yet existing medical benchmarks are either built from publicly available sources with limited expert curation or focus narrowly on disease classification, failing to reflect the stepwise recognition and reasoning processes physicians follow in

real practice. To address this gap, we introduce MedLesionVQA, the first benchmark explicitly designed to evaluate MLLMs on the visual diagnostic workflow for body-surface conditions in large scale. All questions are derived from authentic clinical visual diagnosis scenarios and verified by medical experts with over 20 years of experience, while the data are drawn from 10k+ real patient visits, ensuring authenticity, clinical reality and diversity. MedLesionVQA consists of 12K in-house volunteer images (*never publicly leaked*) and 19K expert-verified question–answer pairs, with fine-grained annotations of 94 lesion types, 110 body regions, and 96 diseases. We evaluate 20+ state-of-the-art MLLMs against human physicians: the best model reaches 56.2% accuracy, far below primary physicians (61.4%) and senior specialists (73.2%). These results expose the persistent gap between MLLMs and clinical expertise, underscoring the need for the multimodal benchmarks to drive trustworthy medical AI.

1 INTRODUCTION

Taking a photo and consulting multimodal large language models (MLLMs) has become a main approach for addressing body surface health concerns, including the skin, nails, hair, oral cavity, genitals, and other visible areas. It requires MLLMs Saab et al. (2024); Moor et al. (2023); Li et al. (2023a); Chen et al. (2024a); Lin et al. (2025); Nath et al. (2024) and medical MLLMs Tian et al. (2023); Chen et al. (2023); Wei Zhu & Wang (2023); Wang et al. (2025) to give visual diagnosis results according to body lesion images photographed by users via smartphone or other device. Although current MLLMs have shown the ability for medical assistance Esteva et al. (2017); Coustasse et al. (2019); Tschandl et al. (2020), they still struggle to replicate the visual diagnostic workflow Weller et al. (2014) that physicians rely on for body-surface health—spanning fine-grained recognition, reasoning, diagnosis, and treatment suggestions across departments such as dermatology, dentistry, and general surgery. The critical challenge is how to evaluate whether MLLMs can truly align with this workflow and perform like physicians in authentic clinical settings.

Existing medical benchmarks are either assembled from publicly available sources with limited expert curation or focus narrowly on disease classification, failing to capture the visual diagnostic workflow for body-surface health that physicians follow in practice. General-purpose benchmarks, such as GMAI-MMBench Ye et al. (2024) and OmniMedVQA Hu et al. (2024), extend to up to 38 modalities by aggregating data from open-source websites. Although these datasets are extensive, publicly sourced information often includes outdated or basic-level data and lacks expert annotations critical for lesion interpretation and treatment recommendations. Conversely, specialized datasets such as SkinCon Daneshjou et al. (2022b) and DDI Daneshjou et al. (2022a) integrate expert annotations but focus narrowly on singular tasks, such as disease classification, not adequately reflecting real-world clinical practice. For instance, SkinCon Daneshjou et al. (2022b) introduces lesion concepts, which are visual symptoms of disease, without open-ended diagnostic queries. DDI employs binary labeling (e.g., malignant vs. benign), which oversimplifies the real-world clinical complexities. Additionally, SkinCon contains only 3,700 images, and DDI encompasses merely 656 cases Daneshjou et al. (2022a), which are insufficient for robust evaluation.

To address these issues, we introduce MedLesionVQA, the first benchmark explicitly designed to evaluate the visual diagnostic workflow for body-surface health. To ensure authenticity and close alignment with physician practice, we collaborated with senior medical directors with over 20 years of experience and defined seven core diagnostic abilities by referring to authoritative textbooks and clinical literature Weller et al. (2014). These abilities span lesion recognition, reasoning, diagnosis, and treatment across dermatology & STD, dentistry, and surgery. Our dataset comprises 12K images collected directly from real patient volunteers, guaranteeing that *none originate from internet sources or leaked repositories*. With 12K images and 19K question–answer pairs, MedLesionVQA is substantially larger than prior expert-curated benchmarks for body-surface health, enabling more robust and diverse evaluation. Beyond its authenticity and scale, MedLesionVQA implements

a fine-grained annotation system, covering 94 lesion types, 96 diseases, and 110 anatomical regions. For example, a human hand is subdivided into nine distinct regions, from the purlicue to the fingertip, enabling highly detailed evaluation of model performance.

Furthermore, our QA generation pipeline is grounded in real clinical questions, which serve as templates for automatic generation and are then refined through rigorous expert review. This yields over 19K diverse, high-quality QA pairs with expert-level accuracy and statistical reliability, addressing gaps left by prior benchmarks. After extensive prompt tuning and iterative refinement, we establish an LLM-based scoring system developed with physicians, ensuring strong consistency between automated assessments and human judgments. Our key contributions are summarized as follows:

- **The first body-surface benchmark aligned with visual diagnostic workflow.** We introduce the first multimodal benchmark explicitly designed to evaluate the visual diagnostic workflow for body-surface health, moving beyond narrow disease classification. MedLesionVQA evaluates the stepwise diagnostic abilities of state-of-the-art MLLMs, providing a foundation for their advancement toward real-world clinical use.
- **Expert-level and fine-grained annotation system.** Our benchmark benefits from valuable expert annotations, covering over 96 prevalent diseases, 110 body regions and sub-regions, and 94 distinct lesion types. All annotations are conducted and rigorously verified by clinical experts following a systematic clinical lexicon tree.
- **Comprehensive evaluation.** We conducted an extensive evaluation involving more than 20 widely-used MLLMs. Additionally, we established human baselines by engaging general practitioners and senior physicians, enabling a thorough and systematic comparison between MLLMs and medical experts.

Table 1: Difference between MedLesionVQA and other existing benchmarks/datasets.OmniMedVQA* Hu et al. (2024) and GMAI-MMBench*Ye et al. (2024) contains a subset of lesion images for dermatology-related evaluation.

Benchmark	Images/QA	VQA	Data source	**Anno./Eval. dimension**
OmniMedVQA* Hu et al. (2024)	119K / 128K	✓	public	lesion (*unknown*) body region (25)
GMAI-MMBench*Ye et al. (2024)	26K / 26K	✓	public	disease (*unknown*)
Fitzpatrick17K Groh et al. (2021)	17K / *null*	✗	public	disease (114)
DermNet der (2023)	19K / *null*	✗	public	disease (23)
SkinCon Daneshjou et al. (2022b)	3230 / *null*	✗	public	lesion concepts (48)
DDI Daneshjou et al. (2022a)	656 / *null*	✗	volunteer	disease (2)
SNU-134 Han (2019)	2101 / *null*	✗	volunteer	disease (134)
MedLesionVQA	12K / 19K	✓	volunteer	lesion (94) and attribute (7) body region (110) disease (96) suggestion & treatment

2 Related Works

2.1 Multimodal Large Language Models

Numerous Multimodal Large Language Models have been developed, focusing primarily on improving image captioning, visual question answering, and cross-modal retrieval Achiam et al. (2023); Anthropic (2025a); Bai et al. (2023); Chen et al. (2024d;e); Liu et al. (2023c); Chen et al. (2024e;b). Representative models include the GPT-4V Achiam et al. (2023), DeepSeek series Guo et al. (2025), LLAVA series Li et al. (2024); Liu et al. (2023c), InternVL series Chen et al. (2024e;c), Qwen series Bai et al. (2025); Wang et al. (2024b),

and CogVLM series Wang et al. (2024c); Hong et al. (2024), among others Laurençon et al. (2023); Ding et al. (2021). These works have significantly contributed to the development of the community. To address specific medical tasks, researchers have trained and fine-tuned MLLMs using specialized medical data, leading to the development of medical vision-language models Li et al. (2023a); He et al. (2024); Wu et al. (2023); Liu et al. (2023d), which integrate medical images (such as X-rays, MRIs, and CT scans, *etc.*) with clinical data (including patient records, diagnosis, and treatment plans, *etc.*) Ye et al. (2024); Antonelli et al. (2022); Irvin et al. (2019). However, achieving precise medical question answering and fine-grained multimodal diagnostics remains a significant challenge.

2.2 Benchmarks

The field of MLLMs has experienced rapid advancements, both in terms of models Achiam et al. (2023); Bai et al. (2023); Anthropic (2025a) and benchmarks Bitton et al. (2023); Zhu et al. (2024); Li et al. (2025); Ray et al. (2024); Lim et al. (2024); Yu et al. (2023; 2024); Xu et al. (2023); Lee et al. (2024); Roberts et al. (2024). Evaluating the medical capabilities of MLLMs requires specific benchmarks, and the representative medical benchmarks include VQA-RAD Lau et al. (2018), SkinCon Daneshjou et al. (2022b), SkinCAP Zhou et al. (2024), DDI Daneshjou et al. (2022a), SCIN Ward et al. (2024), SLAKE Liu et al. (2021), RadBench Wright & Reeves (2016), MMMU Yue et al. (2024), GMAI-MMBench Ye et al. (2024), OmniMedVQA Hu et al. (2024) and MediConfusion Sepehri et al. (2024), *etc.*. Among which, OmniMedVQA Hu et al. (2024) introduces the largest medical VQA dataset to date, covering 12 data modalities and 20 anatomical regions, with over 100k images. GMAI-MMBench Ye et al. (2024) includes various medical imaging data, such as X-rays, CT scans, MRIs, and ultrasounds, along with corresponding clinical information. RadBench Wright & Reeves (2016) focuses on radiology, involving tasks such as modality recognition and disease diagnosis. In this work, we introduce MedLesionVQA, which consists of 12K+ in-house volunteer body lesion images and 19K expert-verified QA pairs. It uniquely targets the stepwise visual diagnostic multimodal abilities that are central to real visual diagnosis workflows.

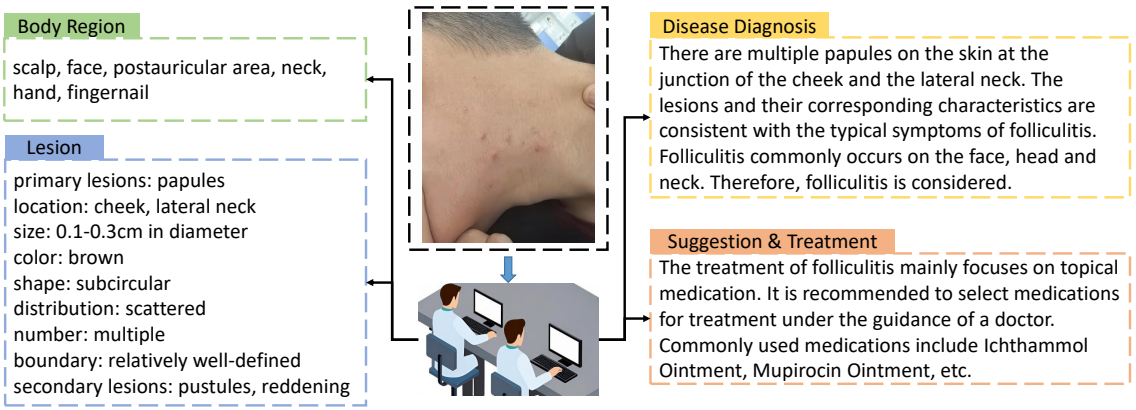


Figure 2: The Annotation procedure. The physicians sequentially annotate the body regions, lesions, attributes, disease diagnosis, and finally suggestion & treatments.

3 Establishment of MedLesionVQA

3.1 Overview of benchmark

MedLesionVQA contains 12K inhouse images collected from volunteers under ethical approvals in data collection process. We cooperate with senior physicians to design and implement an annotation protocol,

referencing authoritative materials Weller et al. (2014); James et al. (2011). The protocol covers 96 prevalent diseases, 94 lesion types, and 119 body regions. Then, inspired by diagnosis and treatment pipeline in clinical practice, we construct 19K diverse question-answer samples involved with 7 stepwise visual diagnostic abilities, and some examples are shown in Fig. 1. These 7 abilities include lesion recognition, attribute recognition, region recognition, spatial relation, lesion reasoning, disease diagnosis and suggestion & treatment, and detailed explanation can be found in supplement materials. Finally, we propose an automated scoring pipeline to calculate the metric of MLLMs' benchmark results, and the scoring pipeline is tuned to align physician judgment metric with negligible difference.

3.2 Data collection

We recruite more than 10K+ volunteers aging from 15 to 75 years old to take photos on their body lesion regions. Each person is instructed to take at least 5 photos at near, medium, and far camera focus, respectively. Finally, these images are preprocessed through image quality filtering, content inspecting, personal information desensitizing, and distribution balancing.

3.3 Annotation protocol

More than tens of physicians are invited into the image annotation process, which contains image filtering, annotation labeling, and annotation reviewing. First, a group of annotators check the quality of each image, such as its clarity, and discard the unqualified images as well as those that do not show the exposed human skin or the oral cavity. Second, body region type, lesion type, lesion attribute type, disease type, and suggestion & treatment annotations are labeled under annotation rules, which are developed by an expert panel of senior experts. Finally, other senior experts review the annotation results and correct any errors, ensuring the annotation quality with entity-level precision and recall of over 95%.

Body region. The physicians are asked to annotate all visible parts of the human body and the internal parts of the oral cavity. We have respectively constructed the corresponding lexical trees for part division, and the annotation is carried out according to the secondary nodes of the lexical trees. More information of the lexical trees is detailed in Appendix A.2.

Lesion. Our dataset has annotations for 94 types of lesions. For each lesion, we describe its key attributes. These attributes are: size, color, shape, quantity, distribution, and boundary. We also pinpoint the exact location of each lesion. To do this, we use a very detailed body map, much like the fine branches of a tree. All our labels have multiple options, not just "yes or no," and most come with at least 7 different text descriptions. Finally, we identify primary and secondary lesions. We also describe their relationship and how often they appear together.

Disease. Each image is provided with up to 3 differential disease diagnosis by two independent physicians, which are sorted in the order they consider the most reasonable. Then, the inverse of the rank is used as the weight to combine the annotation results of the two physicians, to obtain the final sorting result. For the list of total disease labels in the annotation data, please refer to Table 4 of the supplementary material. The logic of diagnostic reasoning is also provided during annotation.

Suggestion & **Treatment.** For each image, physicians are required to provide corresponding treatment suggestions based on the unique disease diagnosis or differential disease diagnosis, including advice on seeking medical treatment, medication, matters needing attention in daily life, and so on.

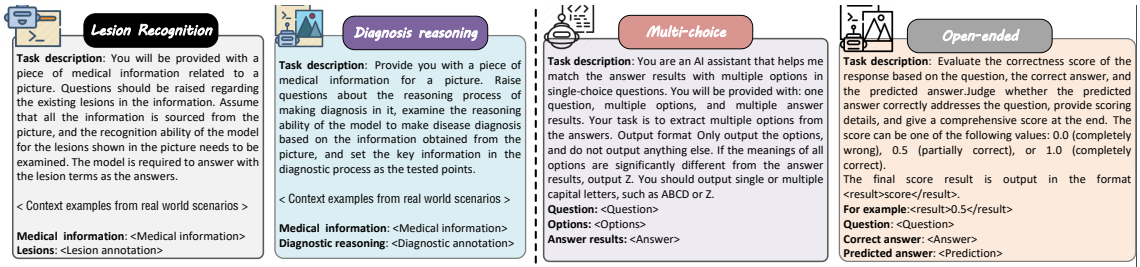
3.4 Question-answer construction

This section introduces the process of question generation, including category balance, prompt design tailored for assessing different cognitive abilities, and the development of various question types.

Evaluation category balance. We balanced the distribution of questions across seven abilities to closely reflect their real-world distribution in clinical practice, as illustrated in Fig. 1. Lesion, attribute, and location recognition questions comprise 61% of the MedLesionVQA dataset, as accurate fine-grained recognition is fundamental for subsequent diagnostic tasks. Specifically, the evaluation assigned equal weighting to each lesion type according to the real-world distribution, ensuring comprehensive coverage for accurate skin lesion identification and analysis.

QA construction prompts. In the context of real-world question examples, we design different QA generation templates for different evaluated abilities in order to test the corresponding capabilities. Two typical prompts are displayed in Fig. 3(a), and the rest will be included in the supplementary materials.

Diverse question types. The generated questions are categorized into two types: multi-choice and open-ended questions, while open-ended questions include judgment, fill-in-the-blank, and short-answer questions. For multi-choice questions, we create similar distracted options based on the correct answer and then randomize the order of all options, ensuring that the correct answer has an equal likelihood of appearing in any position. To prevent answers from being overly diverse and difficult to assess, the answers to open-ended questions are kept relatively concise. This approach enables the judging model to provide more consistent scores in the subsequent evaluation.



(a) Prompts of automatic QA construction for evaluation abilities.

(b) Prompts for extracting answers and scoring predictions.

Figure 3: The prompt template used on MedLesionVQ. Medical information includes body region, lesion, attribute, disease diagnosis, and suggestion & treatment information annotated above.

Manual review and improvement. To enhance the medical accuracy and ensure appropriate difficulty in QA sets, physicians manually review all auto-generated QA pairs. This review focuses primarily on verifying the correctness of critical medical information within both the questions and the answers. Ambiguous questions are clarified, and non-standard answers are revised accordingly. Additionally, distractors in multi-choice questions are assessed regarding their accuracy and difficulty. A few open-ended questions, particularly those concerning suggestion & treatment and lesion reasoning, are converted into multi-choice format due to the inherent complexity of determining definitive answers. The final benchmark comprises 19,843 question-answer pairs (QAs), which are partitioned into a validation subset containing 1,499 QAs (7.55% of total samples) and a test subset consisting of 18,344 QAs (92.45% of total samples).

3.5 Automatic scoring pipeline

For *multiple-choice questions*, since MLLMs occasionally fail to output exact option answer, we need to extract the option answer from the answer set and the raw prediction output using extracting-answer prompt and then compare it with the correct answer. To calculate score, we have set the following rules: 1) If the predicted answer contains options that are not in the correct answer set, it is considered completely wrong and receives a score of 0; 2) If the predicted answer fails to identify all correct answers, the score is calculated based on the ratio of the number of correctly answered options to the total number of correct answers.

For *open-ended questions*, the prompt for the judge model is designed as indicated in Fig. 3(b). With this prompt, the judge model will analyze the predicted answer, compare its similarity to the correct answer, and most importantly, determine whether the question has been answered.

Evaluation consistency test. We use GPT-4 as judger to score the model's predicted answers for open-ended QAs. Moreover, we invite physicians to score the answers, also using the three scoring levels of $0 - 0.5 - 1.0$. Through the analysis of inconsistent cases, we find that the model is *too strict* in scoring for attributes such as color and size. For example, or color descriptions like "pink" and "skin tone", and size descriptions like "pinpoint" and "millimeter", due to the lack of specialized medical knowledge, the judge model tends to be overly strict according to general criteria. When we supplement the evaluation details for color and size in the prompt, therefore the high consistency rate between the judge model's scores and manual scores can be ensured. The details can be found in Appendix.

4 EXPERIMENTS

4.1 EVALUATION

MLLMs baseline. For closed-source models, we evaluate several well-known models, including GPT series models Achiam et al. (2023), Gemini series models Google (2025); DeepMind (2024), and Claude4-opusAnthropic (2025a). For open-source models, we comprehensively evaluate model parameters ranging from 0.256 billion to 72 billion, including the famous LLaVA seriesLiu et al. (2023b); Li et al. (2023b), Qwen2.5 series Wang et al. (2024a), InternVL seriesChen et al. (2024e) and DeepSeek-VL series Wu et al. (2024).

Physician baseline. We invite two groups of 15 primary and 15 senior physicians to answer the 1499 questions in the validation set, respectively. Primary physicians are general practitioner, while senior physicians are specialized expert from dermatology or dentistry departments. Questions are randomly distributed, and each question is completed by at least 2 different physicians. The physicians are not allowed to consult textbooks or search the Internet during the question completion task.

Evaluation Implementation. The evaluation is conducted using the VLMEvalKit Duan et al. (2024) framework. We evaluate all models using a zero-shot setting. All tests are conducted on 8 NVIDIA H20 GPUs (96GB). We additionally add a text-only baseline input to isolate the contribution of the visual modality, helping to evaluate the model's reliance on visual versus textual information.

4.2 MAIN RESULTS

The evaluation results presented in Tab. 2 compare the performance of 22 vision-language models on MedLesionVQA which includes 7 medical tasks aligned closely with real clinical setting, assessed through both multiple-choice and open-ended question formats. Fig 4 presents the performance of 10 representative MLLMs across the 7 ability dimensions defined in MedLesionVQA. In general, Gemini-2.5-proGoogle (2025) shows the best performance across nearly all capabilities with 56.24% average accuracy. Senior physicians achieve averaged score of 73.21%, far beyond the best MLLMs. Key findings from this comprehensive comparison include:

Insight 1: MLLMs Cannot Function as Body Surface Health Doctors. MedLesionVQA presents significant challenges for multimodal large language models (MLLMs). The overall accuracy of representative MLLMs on our MedLesionVQA benchmark is below 57%, emphasizing the need for implementing real-world visual diagnostic tests. Although many MLLMs claim to perform at a physician's level, Tab. 2 indicates that even the best MLLM performs notably worse than primary care physicians (by 5%) and significantly worse than expert clinicians (by 17%). The primary reason of incorrect diagnosis are errors in recognizing

Table 2: The overall accuracy of open-source and closed-source models on the test set and validation set. *:Some closed-source commercial models are evaluated only on the valid set due to API access limitations. The table is sorted in descending order based on the AVG_test score.

Model	AVG_val (1499)	AVG_test (18344)	Recognition			Understanding			
			Lesion Recognition (3340)	Location Recognition (3986)	Attribute Recognition (3508)	Spatial Relation (1133)	Lesion Reasoning (3071)	Disease Diagnosis (1693)	Suggestion Treatment (1613)
Text + Image as Input									
Senior physicians*	0.7321	-	0.6826	0.7583	0.7046	0.7102	0.6533	0.7313	0.8574
Primary physicians*	0.6144	-	0.5932	0.6218	0.5203	0.6336	0.5412	0.6258	0.8162
Gemini-2.5-pro*Google (2025)	0.5624	-	0.4902	0.5166	0.4300	0.6223	0.5754	0.6048	0.8482
GPT-5*OpenAI (2025)	0.5252	-	0.4741	0.5109	0.4039	0.6932	0.4550	0.4444	0.5684
Claude4-opus*Anthropic (2025b)	0.5139	-	0.3906	0.4513	0.4488	0.7412	0.4458	0.5744	0.6076
GPT-O3*OpenAI (2024)	0.5092	-	0.4379	0.4881	0.4718	0.6288	0.4302	0.3826	0.4229
GPT-4V OpenAI (2024)	0.4938	0.4915	0.4071	0.4780	0.4050	0.6308	0.3393	0.5132	0.8216
Gemini-2.0-flashDeepMind (2024)	0.4954	0.4801	0.4062	0.4453	0.3923	0.6112	0.3443	0.5219	0.8136
Qwen2.5-VL-72B Wang et al. (2024a)	0.4904	0.4904	0.3735	0.4636	0.417	0.6618	0.3608	0.5272	0.8246
InternVL2.5-78B Chen et al. (2024e)	0.4790	0.4757	0.3352	0.4981	0.4259	0.6601	0.3084	0.4800	0.7963
GLM-4V-9B GLM et al. (2024)	0.4654	0.4474	0.3472	0.4528	0.3584	0.5596	0.3283	0.4929	0.7281
Qwen2.5-VL-7B Wang et al. (2024a)	0.4243	0.4243	0.3256	0.4005	0.3547	0.5482	0.3356	0.4248	0.7474
Deepseek-vl2-smallWu et al. (2024)	0.4142	0.4164	0.3226	0.4107	0.3627	0.5297	0.2534	0.4822	0.7192
Deepseek-vl2 Wu et al. (2024)	0.3882	0.3928	0.3293	0.3383	0.3514	0.5563	0.2468	0.4309	0.7147
Qwen2-VL-2B Wang et al. (2024a)	0.3536	0.3533	0.2876	0.3319	0.3059	0.4448	0.2057	0.4171	0.6675
LLaVA-InternLM-7B Contributors (2023)	0.3467	0.3316	0.2700	0.3135	0.2967	0.3887	0.1947	0.3981	0.5959
Deepseek-vl2-tiny Wu et al. (2024)	0.3168	0.3293	0.2660	0.2869	0.3079	0.4529	0.1817	0.3953	0.6109
LLaVA-v1.5-13B Liu et al. (2023b)	0.2980	0.3008	0.2437	0.3270	0.2742	0.3177	0.1798	0.3082	0.4966
InternVL2.5-38B Chen et al. (2024e)	0.3096	0.2994	0.3035	0.3247	0.2796	0.3109	0.1474	0.2772	0.4082
ShareGPT4V-7B Chen et al. (2024b)	0.2897	0.2831	0.2232	0.2914	0.2656	0.4158	0.1476	0.3256	0.4235
LLaVA-mistral-7B Liu et al. (2023a)	0.2911	0.2731	0.2205	0.2714	0.2640	0.3740	0.1585	0.2399	0.4913
LLaVA-v1.5-7B Liu et al. (2023b)	0.2648	0.2595	0.2254	0.2456	0.2288	0.3169	0.1605	0.3042	0.423
InternVL2.5-4B Chen et al. (2024e)	0.2632	0.254	0.1895	0.3151	0.2428	0.2172	0.1336	0.3121	0.2965
SmolVLM-500M Marafioti et al. (2025)	0.1898	0.1761	0.1711	0.1602	0.1897	0.2656	0.0992	0.1417	0.2190
SmolVLM-256M Marafioti et al. (2025)	0.1564	0.156	0.1397	0.1418	0.1507	0.2172	0.0912	0.1691	0.2274
LLaVA-med-v1.5-7B Li et al. (2023b)	0.0885	0.0791	0.0372	0.0715	0.1104	0.1258	0.0466	0.0535	0.1426
Only Text as Input									
InternVL2.5-78B Wang et al. (2024a)	0.3636	0.3839	0.3378	0.3089	0.3763	0.6606	0.2967	0.3946	0.8014
Qwen2.5vl-72B Wang et al. (2024a)	0.3478	0.3537	0.2640	0.2784	0.2987	0.5818	0.3194	0.3016	0.8124
InternVL2.5-4B Chen et al. (2024e)	0.3403	0.3406	0.2071	0.3023	0.3190	0.5266	0.2981	0.2645	0.7446
GPT-4V Achiam et al. (2023)	0.3089	0.3185	0.2201	0.1687	0.3200	0.6076	0.2441	0.2844	0.8140
Qwen2.5VL-7BWang et al. (2024a)	0.3153	0.3097	0.2217	0.2376	0.2646	0.4900	0.2939	0.2945	0.7404
Deepseek-vl2 Wu et al. (2024)	0.2981	0.2851	0.2452	0.1685	0.2916	0.5455	0.1996	0.3032	0.7227
Qwen2-VL-2B Wang et al. (2024a)	0.2693	0.2814	0.2146	0.2384	0.2636	0.4195	0.1873	0.2232	0.6389
ShareGPT4V-7B Chen et al. (2024b)	0.2193	0.2477	0.1940	0.1171	0.2293	0.3374	0.1439	0.2668	0.4247
LLaVA-med-v1.5-7B Li et al. (2023b)	0.0842	0.0763	0.0349	0.0535	0.1096	0.1533	0.0398	0.0739	0.1899

lesion types, locations, attributes, or relationships-tasks that human doctors perform reliably while the best lesion recognition accuracy for MLLMs is only 49%. Our results from MedLesionVQA show that MLLMs frequently fail in diagnostic tasks and often struggle to align with physicians in real clinical settings. These findings underscore the need for caution when employing MLLMs as medical practitioners and highlight the necessity to develop more advanced medical-specific MLLMs.

Insight 2: Textual Capabilities Can Cause MLLMs to Appear More Competent Than They Are

People often perceive MLLMs as highly knowledgeable experts and report positive experiences during question-and-answer interactions. However, our MedLesionVQA benchmark suggests that MLLMs seem more competent than they are due to their impressive text generation abilities, even when subjective questions are minimized in MedLesionVQA. A comparison between text-only and vision-text evaluations indicates that "suggestion" scores remain high regardless of the modality (82.4% vs. 81.2% with and without images). The

high accuracy of treatment recommendations demonstrates that large language models can generate effective general advice, even without specialized expertise in body health images. In contrast, MLLMs perform poorly on more visually demanding tasks, such as lesion and location recognition. These findings underscore the necessity of comprehensive clinical pipeline evaluations when applying MLLMs in medical contexts.

Insights 3: Performance Improves as Model Size Increases. The results demonstrate a generally positive correlation between model size and performance, but with diminishing returns and notable exceptions. Models under 1B parameters (e.g., SmoMLLM-256M/500M) show limited capabilities across all tasks (scores below 0.2), while mid-scale models (1B-10B) like Qwen2-VL-2B and Deepseek-vl2-tiny (3.4B) exhibit significant performance jumps, particularly in recognition and diagnostic tasks. The GLM-4V-9B model achieves near-state-of-the-art results, rivaling much larger models with average of 0.465 compared to the 0.309 socre of InternVL2.5-38B. However, scaling beyond 10B parameters shows inconsistent returns – while Qwen2-VL-72B dominates in most metrics, the InternVL2.5-78B underperforms smaller models in key areas like disease diagnosis, suggesting current architectural or training limitations in MLLMs. Generally, closed-source models consist of hundreds of billions of parameters and provide the relatively high performance.

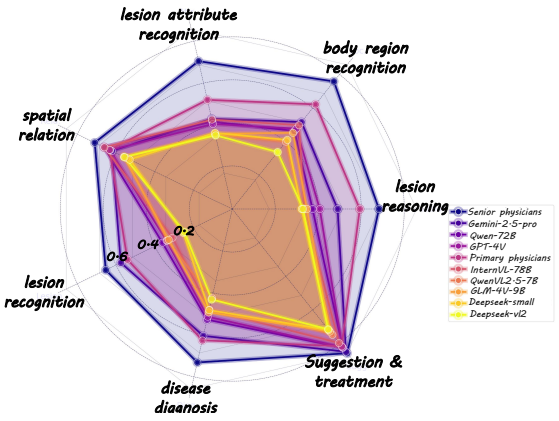


Figure 4: Results of 10 representative MLLMs across the 7 ability dimensions defined in MedLesionVQA.

Insight 4: The Need to Rethink Domain-Specific Models. The comparison between LLaVA1.5-7B and LLaVA-Med-7B highlights the trade-off between specialization and generalization. LLaVA-Med-7B performs 18% worse than LLaVA1.5-7B on the MedLesionVQA dataset, yet demonstrates superior performance on VQA-RAD. Simply applying instruction tuning to general-purpose foundation models may diminish model performance in other domains, even within the same medical concept.

To show more evaluation results, we also analyze the error instances sampled from the model's predictions and give the distribution of these errors, including lack of knowledge, text misunderstanding, and judgment error, etc, in Appendix B.2 and B.3.

5 Conclusion

In this paper, we propose MedLesionVQA, a large-scale and body surface oriented benchmark evaluating the lesion, region, diagnosis, and treatment-related recognition and reasoning ability for medical MLLMs. MedLesionVQA contains 12K body lesion images with expert-level fine-grained annotations of 96 prevalent dermatological diseases, 94 distinct lesion types and 110 body regions. The evaluation dimension of MedLesionVQA is built on basis of 7 multimodal stepwise visual diagnostic abilities, including lesion recognition, lesion attribute recognition, body region recognition, lesion spatial relation recognition, lesion reasoning, disease diagnosis and suggestion & treatment, which ensure the alignment with the authentic clinic senary. Mainstream MLLMs are evaluated on the benchmarks, and Gemini-2.5-pro has the best score of 56.24. Furthermore, senior and primary physicians are invited to answer the questions of benchmark and obtain score of 61.44 and 73.21, respectively. The results show that there is large improvement for MLLMs on the benchmark and indicates significant challenges and medical specialization of the MedLesionVQA.

REFERENCES

Dermnet, 2023. https://dermnet.com/ [2024].

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Anthropic. Claude 3 model family. https://www.anthropic.com, 2025a.

Anthropic. Introducing claude 4. Anthropic News, 2025b. URL https://www.anthropic.com/news/claude-4?_bhlid=aeb6fd9f68ee0feec09df9256d36a1ef7371ca56.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. Nature communications, 13(1):4128, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. Text Reading, and Beyond, 2, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. arXiv preprint arXiv:2308.06595, 2023.

Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, et al. Huatuogpt-ii, one-stage training for medical adaption of llms. arXiv preprint arXiv:2311.09774, 2023.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In European Conference on Computer Vision, pp. 370–387. Springer, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024c.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences, 67(12):220101, 2024d.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24185–24198, 2024e.

XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023.

Alberto Coustasse, Raghav Sarkar, Bukola Abodunde, Brandon J Metzger, and Chelsea M Slater. Use of teledermatology to improve dermatological access in rural areas. Telemedicine and e-Health, 25(11): 1022–1032, 2019.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. Science advances, 8(31):eabq6147, 2022a.

Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. Advances in Neural Information Processing Systems, 35:18157–18167, 2022b.

Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens, 2024. URL https://arxiv.org/abs/2403.05530. Accessed: 2025-04-30.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. Advances in neural information processing systems, 34:19822–19835, 2021.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 11198–11201, 2024.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639): 115–118, 2017.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Google. Gemini 2.5 pro, 2025. URL https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro. Large language model; Capable of handling various modalities such as text, audio, image, and video; Supports a context window of 1 million tokens.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1820–1828, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

Seung Seog Han. SNU dataset + Quiz. 3 2019. doi: 10.6084/m9.figshare.6454973.v12. URL https://figshare.com/articles/dataset/SNU_SNU_MELANOMA_and_Reddit_dataset_Quiz/6454973.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv e-prints, pp. arXiv–2404, 2024.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500, 2024.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22170–22183, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pp. 590–597, 2019.

William D James, Dirk Elston, and Timothy Berger. Andrew's diseases of the skin E-book: clinical dermatology. Elsevier Health Sciences, 2011.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. Scientific data, 5(1):1–10, 2018.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. Advances in Neural Information Processing Systems, 36: 71683–71702, 2023.

Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. Vhelm: A holistic evaluation of vision language models. Advances in Neural Information Processing Systems, 37:140632–140666, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36:28541–28564, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890, 2023b.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024.

Haodong Li, Xiaofeng Zhang, and Haicheng Qu. Ddfav: Remote sensing large vision language models dataset and evaluation benchmark. Remote Sensing, 17(4):719, 2025.

Hyeonseok Lim, Dongjae Shin, Seohyun Song, Inho Won, Minjun Kim, Junghun Yuk, Haneol Jang, and KyungTae Lim. Vlr-bench: Multilingual benchmark dataset for vision-language retrieval augmented generation. arXiv preprint arXiv:2412.10151, 2024.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. arXiv preprint arXiv:2502.09838, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In 2021 IEEE 18th international symposium on biomedical imaging (ISBI), pp. 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36:34892–34916, 2023c.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. arXiv preprint arXiv:2310.17956, 2023d.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299, 2025.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Machine Learning for Health (ML4H), pp. 353–367. PMLR, 2023.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. arXiv preprint arXiv:2411.12915, 2024.

OpenAI. Gpt-4o, 2024. URL `https://chat.openai.com`. Large language model; Prompt: "".

OpenAI. Chatgpt (gpt-5 version), 2025. URL `https://chat.openai.com/chat`.

Sourjyadip Ray, Kushal Gupta, Soumi Kundu, Payal Arvind Kasat, Somak Aditya, and Pawan Goyal. Ervqa: A dataset to benchmark the readiness of large vision language models in hospital environments. arXiv preprint arXiv:2410.06420, 2024.

Josselin S Roberts, Tony Lee, Chi H Wong, Michihiro Yasunaga, Yifan Mai, and Percy Liang. Image2struct: Benchmarking structure extraction for vision-language models. Advances in Neural Information Processing Systems, 37:115058–115097, 2024.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416, 2024.

Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models. arXiv preprint arXiv:2409.15477, 2024.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. arXiv preprint arXiv:2311.06025, 2023.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. Nature medicine, 26(8):1229–1234, 2020.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, et al. Baichuan-m1: Pushing the medical capability of large language models. arXiv preprint arXiv:2502.12671, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024b.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems, 37:121475–121499, 2024c.

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. JAMA Network Open, 7(11):e2446615–e2446615, 11 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2024.46615. URL https://doi.org/10.1001/jamanetworkopen.2024.46615.

Wenjing Yue Wei Zhu and Xiaoling Wang. Shennong-tcm: A traditional chinese medicine large language model. https://github.com/michael-wzhu/ShenNong-TCM-LLM, 2023.

Richard B Weller, Hamish JA Hunter, and Margaret W Mann. Clinical dermatology. John Wiley & Sons, 2014.

Chris Wright and Pauline Reeves. Radbench: benchmarking image interpretation skills. Radiography, 22(2): e131–e136, 2016.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. arXiv preprint arXiv:2308.02463, 2023.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Cheng Xu, Xiaofeng Hou, Jiacheng Liu, Chao Li, Tianhao Huang, Xiaozhi Zhu, Mo Niu, Lingyu Sun, Peng Tang, Tongqiao Xu, et al. Mmbench: Benchmarking end-to-end multi-modal dnns and understanding their hardware-software implications. In 2023 IEEE International Symposium on Workload Characterization (IISWC), pp. 154–166. IEEE, 2023.

Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. Advances in Neural Information Processing Systems, 37:94327–94427, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023.

Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. arXiv preprint arXiv:2408.00765, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9556–9567, 2024.

Juexiao Zhou, Liyuan Sun, Yan Xu, Wenbin Liu, Shawn Afvari, Zhongyi Han, Jiaoyan Song, Yongzhi Ji, Xiaonan He, and Xin Gao. Skincap: A multi-modal dermatology dataset annotated with rich medical captions. arXiv preprint arXiv:2405.18004, 2024.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. arXiv preprint arXiv:2410.21311, 2024.