MedLesionVQA: A Multimodal Benchmark Emulating Clinical Visual Diagnosis for Body Surface Health

Deli Yu*, Shengzhi Wang*, Kai Wu*, Haihua Yang+, et al
dely-yu@qq.com, wukaiwork@gmail.com, yanghaihua@bytedance.com

ICLR International Conference On Learning Representations

ByteDance

Dataset (coming soon)

Introduction of MedLesionVQA

- A body-surface benchmark aligned with visual diagnostic workflow
- Expert-level and fine-grained annotation system
- Evaluations of 20+ state-of-the-art MLLMs and physicians

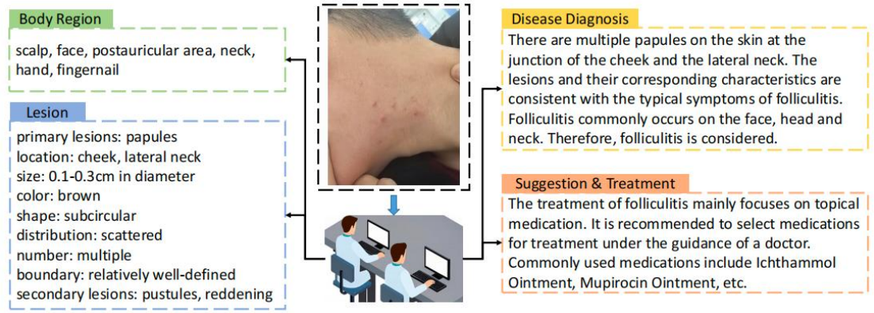
Statistics

- 19K expert-verified question–answer pairs, 12K in-house images
- 7 stepwise diagnostic multimodal evaluation abilities
- 96 prevalent diseases, 94 lesion types, and 110 body regions
- 7 lesion attributes: size, color, shape, quantity, distribution, and boundary

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*	✗	in-house	disease (2)
SNU-134 (Han, 2019)	2101 / *null*	✗	in-house	disease (134)
MedLesionVQA	12K / 19K	✓	in-house	lesion (94) and attribute (7) body region (110) disease (96) suggestion & treatment

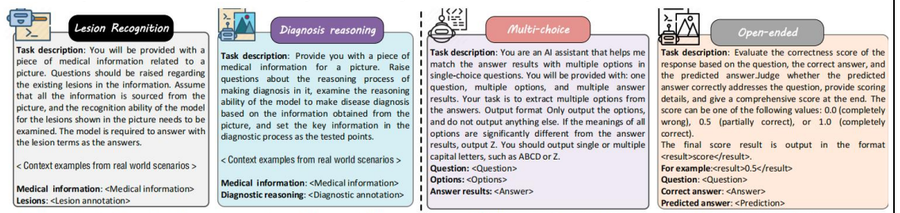
Annotation Protocol

- Body region: Annotate visible body parts and oral cavity regions via lexical trees.
- Lesion: Label 94 lesion types with attributes, location, and relationships.
- Disease: Provide ranked differential diagnoses merged from two physicians' opinions.
- Suggestion & Treatment: Give treatment advice, medication, and daily care recommendations.



Question-answer construction

- Evaluation category balance: Balance question categories to match real clinical distributions.
- QA construction prompts: Design ability-specific prompts from real-world clinical questions.
- Diverse question types: Use multiple-choice and concise open-ended question formats.
- Manual review and improvement: Physicians review QA pairs for accuracy, clarity, and difficulty.

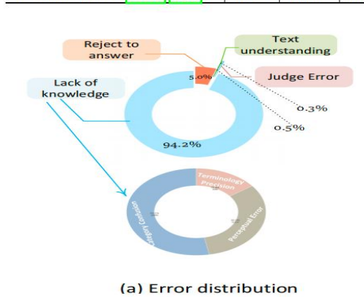


(a) Prompts of automatic QA construction for evaluation abilities. (b) Prompts for extracting answers and scoring predictions.

Evaluation Results

- MLLMs Cannot Function as Body Surface Health Doctors
- Textual Capabilities Can Cause MLLMs to Appear More Competent Than They Are
- Performance Improves as Model Size Increases
- The Need to Rethink Domain-Specific Models

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-flash (DeepMind, 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., 2024f)	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-small (Wu 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
Lingshu-32B (Xu et al., 2025)	0.3541	0.3539	0.21	0.2723	0.3796	0.5551	0.30	0.2811	0.7276
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
Lingshu7b (Xu et al., 2025)	0.3398	0.3448	0.2058	0.271	0.379	0.504	0.2822	0.2794	0.7078
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., 2024f)	0.3096	0.2994	0.3035	0.3247	0.2796	0.3109	0.1474	0.2772	0.4082
ShareGPT4V-7B (Chen et al., 2024c)	0.2897	0.2831	0.2232	0.2914	0.2656	0.4158	0.1476	0.3256	0.4235
HuatuoGPT-vision (Chen et al., 2024a)	-	0.279	0.2281	0.1111	0.2997	0.5649	0.1897	0.282	0.676
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
Biomedix2 (Mullappilly et al., 2024)	0.2603	0.2676	0.171	0.1647	0.3118	0.4095	0.1563	0.349	0.5977
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., 2024f)	0.2632	0.254	0.1895	0.3151	0.2428	0.2172	0.1336	0.3121	0.2965
MedGemma-4b (Sellergren et al., 2025)	0.195	0.203	0.1212	0.0915	0.1697	0.4007	0.1908	0.2339	0.5564
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., 2024f)	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-7B (Wang 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., 2024c)	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



(a) Error distribution