OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM

Supplementary Material

7. The details of involved datasets

To construct comprehensive evaluation benchmark, we collect numerous medical datasets and convert them into VQA format. In Table 8 and Table 9, we provide a list of all the datasets included in OmniMedVQA, along with their modality information, the number of utilized images and QA items, and the access condition. It's worth noting that RadImageNet [80] stands out as one of the largest datasets in the biomedical field, containing 1.35 million radiologic images encompassing CT and MRI modalities, spanning 11 anatomical regions, and covering 165 different diseases. As a result, RadImageNet constitutes a significant portion of our OmniMedVQA dataset.

Additionally, 3D_Modality is our self-construction dataset, incorporating data from 17 different medical datasets. The involved 17 datasets are ISLES_SPES[79], ISLES2016[115], ISLES2017[115], ISLES2018[31, 54], ISLES2022[57], AMOS[61], Longitudinal Multiple Sclerosis Lesion Segmentation[28],VALDO[29], PICAI[98], ASC18[118], BraTS2013[65, 83], BraTS2018[22, 23, 83], MSSEG2016[38], CMRxMotions[114], MRBrainS13[82], BrainTumour[19], MRBrain18[68]. We leverage these datasets to create questions about more fine-grained modality recognition, such as Magnetic Resonance T1-weighted, Magnetic Resonance T2-weighted, Magnetic Resonance T1-weighted Inversion Recovery, Computed Tomography Cerebral Blood Volume, Computed Tomography Time to Maximum *et al.*

We release the OmniMedVQA according to the license and permission. Specifically, there are 42 dataset are completely open access. Thus, we directly provide the images with the corresponding QA items. Meanwhile, there are 31 datasets are restricted access. For these datasets, we only release the evaluation QA items and provide the instruction guidelines, based on which you can associate each QA item with the corresponding images. Researchers only need to download the original datasets and then combine them with the provided QA items according to the guidelines.

For the convenience of future research, in Table 11, Table 12 and Table 13, we provide the evaluation results on the completely open-access datasets. If you do not want to download each restricted access dataset one by one, these results could help you establish the benchmark and analyse the experimental results quickly.

8. The distribution of our dataset

We illustrate the distribution of different classes within Modality Recognition, Anatomy Identification and Disease Diagnosis in Fig. 4. We can find there is no significant bias in our OmniMedVQA and the distribution remains balanced. This demonstrates the effectiveness of the sampling process when we develop the dataset.

9. The details of modalities

In this section, we provide an overview of the number of images and QA items associated with various modalities in OmniMedVQA. As detailed in Table 10, we incorporate data from 12 different modalities. Moreover, to better present the characteristic of different modalities, we illustrate the images with the corresponding QA items in Fig. 5 and Fig. 6.

10. The details of multi-choice questions

As introduced in Sec 3, we generate a set of incorrect options for each item, which are utilized to construct multiple-choice question-answer pairs. The number of candidate options of each question ranges from 2 to 4. In Fig. 7, we illustrate the QA items with different number of options. As depicted, questions with two options are "Yes/No" selection. On the other hand, questions with three options predominantly focus on Lesion Grading, which judges the severity of the disease.

References

[1] Chest ct-scan images dataset. https://tianchi.aliyun.com/dataset/93929, . 2

[2] Covid ct dataset. https://tianchi.aliyun.com/dataset/106604, . 2

[3] Isic 2019 challenge. https://challenge.isic-archive.com/landing/2019/. 2

[4] Oral cancer (lips and tongue) images. https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images, . 2

[5] Dental condition dataset. https://www.kaggle.com/datasets/salmansajid05/oral-diseases, . 2

[6] Glaucoma grading based on multi-modality images. https://aistudio.baidu.com/competition/detail/119/0/task-definition, . 2

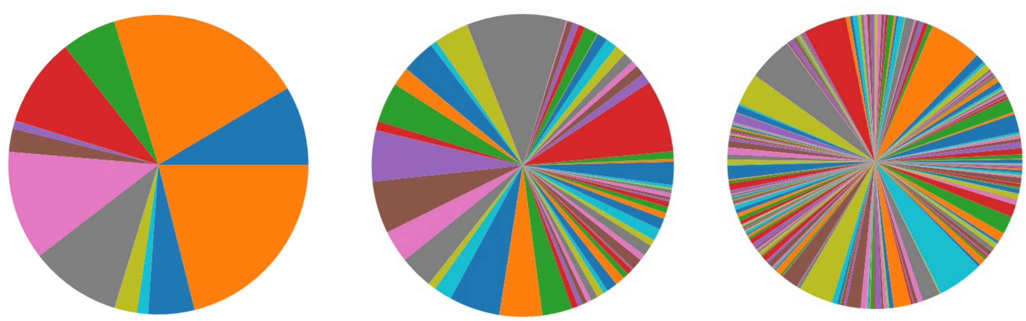
[7] Glaucoma detection. https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection, . 2

Table 8. The information of involved dataset in OmniMedVQA. Notably, **Co** denotes Colposcopy, **CT** denotes Computed Tomography, **DP** denotes Digital Photography, **FP** denotes Fundus Photography, **IRI** denotes Infrared Reflectance Imaging, **MR** denotes Magnetic Resonance Imaging, **OCT** denotes Optical Coherence Tomography, **Der** denotes Dermoscopy, **End** denotes Endoscopy, **Mic** denotes Microscopy Images, **US** denotes Ultrasound.

index	Dataset	Modality	# Imgs	# QA Items	Access
1	TCB_Challenge [56]	X-Ray	18	32	Restricted Access
2	Oral_Cancer_kaggle [4]	DP	27	34	Restricted Access
3	Dental_Condition_Dataset [5]	DP	2281	2752	Restricted Access
4	Cervical_Cancer_Screening [25]	Co	319	338	Restricted Access
5	Chest_CT_Scan [1]	CT	382	871	Open Access
6	Covid_CT [2]	CT	135	199	Open Access
7	SARS-CoV-2 CT-scan [107]	CT	461	910	Open Access
8	RadImageNet [80]	CT,MR,US	55443	56697	Open Access
9	Fitzpatrick 17k [50, 51]	Der	1450	1552	Open Access
10	ISBI2016 [53]	Der	348	681	Open Access
11	ISIC2018 [36, 112]	Der	185	272	Open Access
12	ISIC2019 [3]	Der	1860	1952	Open Access
13	ISIC2020 [97]	Der	1499	1580	Open Access
14	MED-NODE [49]	Der	34	38	Restricted Access
15	Monkeypox Skin Image 2022 [58]	Der	154	163	Open Access
16	PAD-UFES-20 [91]	Der	401	479	Open Access
17	PH^2 [81]	Der	36	45	Restricted Access
18	AIDA [8]	End	207	340	Restricted Access
19	Kvasir [94]	End	1225	1537	Restricted Access
20	ACRIMA [44]	FP	129	159	Open Access
21	Adam Challenge [45]	End	78	87	Open Access
22	AIROGS [42]	FP	3853	4004	Restricted Access
23	APTOS2019_Blindness [62]	FP	544	625	Restricted Access
24	AVN Assessment [87]	FP	18	22	Restricted Access
25	DeepDRiD [78]	FP	131	131	Open Access
26	Diabetic Retinopathy [9]	FP	1996	2051	Open Access
27	DRIMDB [102]	FP	122	132	Open Access
28	GAMMA [6]	FP	20	20	Restricted Access
29	Glaucoma_Detection [7]	FP	121	142	Restricted Access
30	JSIEC [30]	FP	177	220	Open Access
31	Messidor-2 [17, 43]	FP	270	321	Restricted Access
32	OLIVES [95]	FP	534	593	Open Access
33	PALM2019 [46]	FP	451	510	Open Access
34	Refuge2 [72, 88]	FP	128	145	Restricted Access
35	Cataract_dataset_kaggle [10]	FP	120	138	Restricted Access
36	Yangxi [76]	FP	1446	1515	Open Access
37	BCNB [119]	MR	4334	4806	Restricted Access
38	BRIGHT Challenge [11]	MR	675	890	Restricted Access
39	BreakHis [108]	MR	684	735	Open Access
40	NLM- Malaria Data [12]	MR	67	75	Open Access
41	CRC100k [63]	MR	1186	1322	Open Access
42	DigestPath19 [40]	MR	81	95	Restricted Access
43	His_Can_Det [39]	MR	7381	7572	Restricted Access
44	lc25000 [26]	MR	1796	1903	Restricted Access
45	MAlig_Lymph [89]	MR	75	149	Open Access
46	MRL_Eye [47]	IRI	9477	9785	Restricted Access
47	BioMediTech [86]	Mic	345	511	Open Access
48	Blood_Cell [13]	Mic	1092	1175	Open Access
49	CornealNerve [99]	Mic	18	25	Restricted Access
50	Cervix93 [93]	Mic	434	664	Restricted Access
51	HuSHeM [103]	Mic	41	89	Open Access
52	BACH2018 [20]	Mic	80	102	Restricted Access
53	ALL Challenge [52]	Mic	295	342	Open Access
54	MHSMA [60]	Mic	1196	1282	Open Access
55	Nerve_Tortuosity [100]	Mic	5	6	Restricted Access
56	Br35h [55]	MR	382	429	Restricted Access
57	OCT & X-Ray 2017 [64]	OCT,X-Ray	1066	1301	Open Access
58	Retinal OCT-C8 [14]	OCT	3224	4016	Open Access
59	Knee_Osteoarthritis [33]	X-Ray	518	518	Open Access
60	RUS_CHN [15]	X-Ray	1642	1982	Open Access

Table 9. Continued from Table 8. The information of involved dataset in OmniMedVQA. Notably, **Co** denotes Colposcopy, **CT** denotes Computed Tomography, **DP** denotes Digital Photography, **FP** denotes Fundus Photography, **IRI** denotes Infrared Reflectance Imaging, **MR** denotes Magnetic Resonance Imaging, **OCT** denotes Optical Coherence Tomography, **Der** denotes Dermoscopy, **End** denotes Endoscopy, **Mic** denotes Microscopy Images, **US** denotes Ultrasound.

index	Dataset	Modality	# Imgs	# QA Items	Access
61	Pulmonary_Chest_Shenzhen [59]	X-Ray	131	296	Open Access
62	Chest_X-Ray_PA [21]	X-Ray	664	850	Open Access
63	CoronaHack [37]	X-Ray	476	684	Open Access
64	Covid-19_tianchi [16]	X-Ray	66	96	Open Access
65	Covid19_heywhale [35]	X-Ray	550	690	Open Access
66	COVIDGR [110]	X-Ray	156	220	Restricted Access
67	COVIDx CXR-4 [113]	X-Ray	335	485	Open Access
68	MINIJSRT [106]	X-Ray	133	257	Restricted Access
69	MIAS [109]	X-Ray	65	142	Open Access
70	Mura [96]	X-Ray	1277	1464	Open Access
71	Pulmonary_Chest_MC [59]	X-Ray	28	38	Open Access
72	SIIM-ACR [124]	X-Ray	1036	1286	Restricted Access
73	3D_Modality	MR,CT	426	426	Partially-Open Access



(a) Modality Recognition (b) Anatomy Identification (c) Disease Diagnosis

Figure 4. We illustrate the distribution of different classes within Modality Recognition, Anatomy Identification and Disease Diagnosis.

Table 10. The numbers of images and QA items sourced from different modalities in our OmniMedVQA.

Modality	# Images	# QA Items
Colposcopy	319	338
CT	14457	15836
Digital Photography	2308	2786
Fundus Photography	10108	10815
Infrared Reflectance Imaging	9477	9785
MR	31917	32705
Optical Coherence Tomography	3791	4646
Dermoscopy	5967	6762
Endoscopy	1432	1877
Microscopy Images	19785	21743
X-Ray	7594	9711
Ultrasound	10855	10991

[8] Analysis of images to detect abnormalities in endoscopy. https://aidasub-cleceliachy.grand-challenge.org/Description/, 2016. 2

[9] Diabetic retinopathy arranged - retina images with class labels for classification. https://tianchi.aliyun.com/dataset/93926, 2023. 2

[10] Cataract image dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset, 2023. 2

[11] Bright challenge: Breast tumor image classification on gigapixel histopathological images. https://research.ibm.com/haifa/Workshops/BRIGHT/, 2023. 2

[12] Nlm - malaria data. https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html, 2023. 2

[13] Blood cell images. https://www.kaggle.com/datasets/paultimothymooney/blood-cells, 2023. 2

[14] Retinal oct - c8 dataset. https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8/data, 2023. 2

[15] X-ray hand small joint classification dataset (based on bone age scoring method rus-chn). https://aistudio.baidu.com/datasetdetail/69582/0, 2023. 2

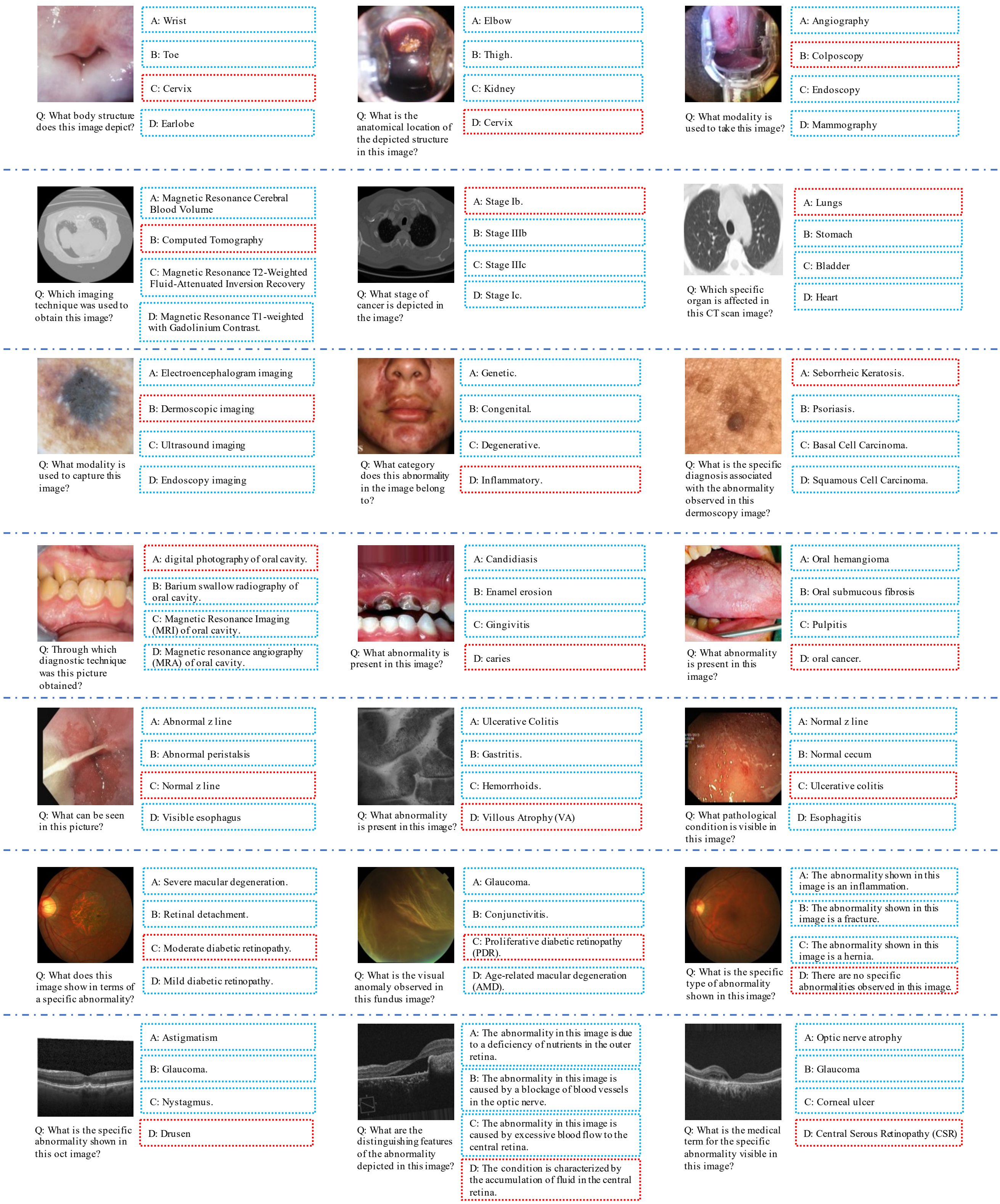


Figure 5. The representative samples from different modalities. From the above to bottom, we illustrate the samples from 7 different modalities in each row, *i.e.,* Colposcopy, CT, Dermoscopy, Digital Photography, Endoscopy, Fundus Photography and OCT. Notably, each dashed box corresponds to a specific option, and the red dashed box indicates the correct option.

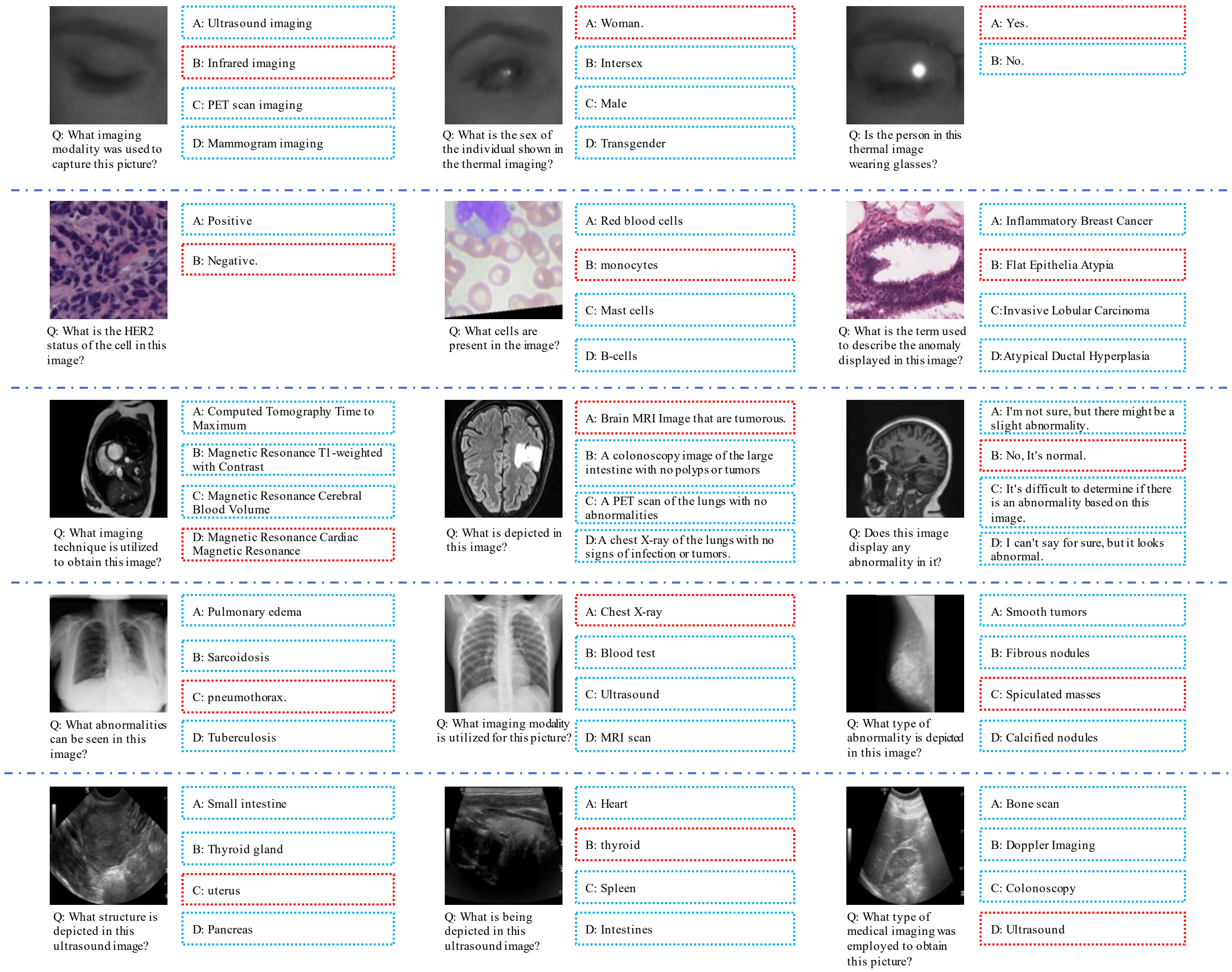


Figure 6. Continued from Fig. 5. The representative samples from different modalities. From the above to bottom, we illustrate the samples from 5 different modalities in each row, *i.e.,* Infrared Reflectance Imaging, Microscopy Images, MR, X-Ray and Ultrasound. Notably, each dashed box corresponds to a specific option, and the red dashed box indicates the correct option.

[16] Covid-19 image dataset: 3 way classification - covid-19, viral pneumonia, normal. https://tianchi.aliyun.com/dataset/93853, 2023. 3

[17] Michael D Abràmoff, James C Folk, Dennis P Han, Jonathan D Walker, David F Williams, Stephen R Russell, Pascale Massin, Beatrice Cochener, Philippe Gain, Li Tang, et al. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA ophthalmology*, 131 (3):351–357, 2013. 2

[18] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, 2022. 3

[19] 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. 1

[20] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019. 2

[21] Amanullah Asraf and Zabirul Islam. Covid19, pneumonia and normal chest x-ray pa dataset. *Mendeley Data*, 1, 2021. 3

[22] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4

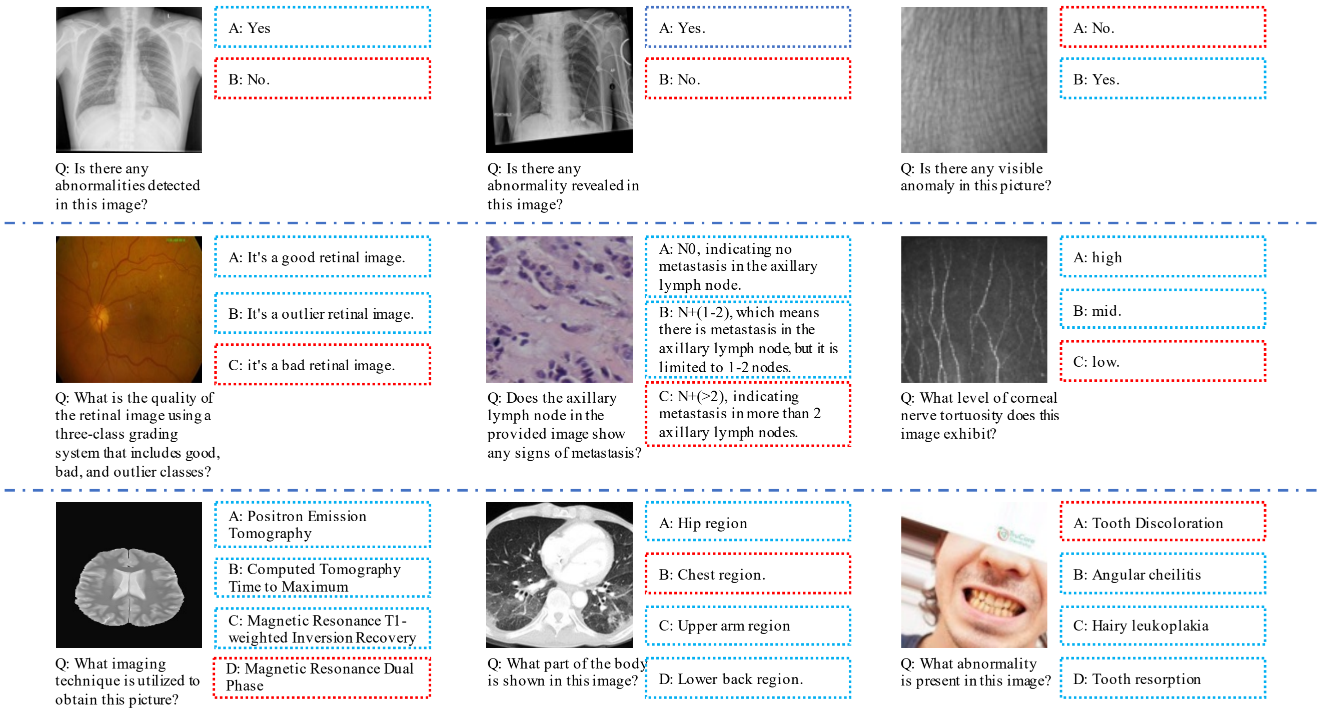


Figure 7. Illustration of representative samples with different numbers of candidate options.

Table 11. The accuracy of representative LVLMs on completely open-access data of our OmniMedVQA in terms of five different question types. Notably, we report the Question-answering Score and Prefix-based Score before and after "/", respectively. Meanwhile, in each column, the best performance is marked in red, while the second best performance is marked in blue.

Model	Modality Recognition	Anatomy Identification	Disease Diagnosis	Lesion Grading	Other Biological Attributes	Overall
Random Guess	25.00	25.67	25.12	27.86	25.48	26.91
MiniGPT-4 [128]	26.43 / 25.85	28.88 / 30.19	30.47 / 19.31	34.56 / 40.04	30.36 / 39.25	29.74 / 23.44
BLIP-2 [73]	68.19 / 46.75	44.39 / 72.43	44.51 / 23.64	29.03 / 24.31	67.95 / 33.85	48.12 / 36.08
InstructBLIP [41]	75.27 / 24.15	44.35 / 57.56	32.29 / 24.97	59.25 / 56.91	23.72 / 36.65	40.40 / 32.10
mPLUG-Owl [123]	28.95 / 9.30	24.83 / 32.18	30.13 / 25.39	43.61 / 86.84	28.62 / 34.25	29.25 / 26.35
Otter [70]	24.50 / 7.96	25.81 / 25.24	27.74 / 22.73	37.37 / 41.71	26.33 / 28.07	27.13 / 21.93
LLaVA [77]	21.36 / 14.94	25.86 / 15.16	29.10 / 20.50	43.95 / 24.07	31.90 / 36.11	27.96 / 19.49
LLaMA_Adapter_v2 [48]	37.29 / 40.36	33.72 / 43.30	31.19 / 23.52	41.99 / 37.13	34.22 / 41.25	32.82 / 30.38
VPGTrans [125]	26.80 / 31.81	31.06 / 37.95	30.05 / 17.62	30.60 / 40.75	29.67 / 39.31	29.81 / 24.62
Med-Flamingo [84]	30.19 / 24.25	24.93 / 25.57	38.90 / 24.29	30.74 / 52.48	14.18 / 38.25	34.03 / 25.73
RadFM [117]	13.31 / 51.76	21.69 / 30.11	30.35 / 28.30	26.64 / 33.56	43.85 / 40.28	26.99 / 32.27
MedVInT [127]	68.10 / 33.61	40.26 / 23.27	35.78 / 28.29	12.77 / 5.29	30.30 / 23.58	40.04 / 27.32
LLaVA-Med [71]	26.93 / 13.06	29.53 / 22.47	29.22 / 32.50	34.18 / 30.41	33.08 / 22.70	29.25 / 27.69

(1):1–13, 2017. 1

[23] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 1

[24] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021. 1, 3

[25] BenO, JL Jones, H Kumar, Meg Risdal, M Rao,

Table 12. The overall accuracy of representative LVLMs on completely open-access data of our OmniMedVQA in terms of different modalities . Here, we report the accuracy of all items within each modality when utilizing the **Question-answering score**. Specifically, **Co** denotes Colposcopy, **CT** denotes Computed Tomography, **DP** denotes Digital Photography, **FP** denotes Fundus Photography, **IRI** denotes Infrared Reflectance Imaging, **MR** denotes Magnetic Resonance Imaging, **OCT** denotes Optical Coherence Tomography, **Der** denotes Dermoscopy, **End** denotes Endoscopy, **Mic** denotes Microscopy Images, **US** denotes Ultrasound. Meanwhile, in each column, the best and second-best performance are marked in red and blue, respectively.

Model	Co	CT	DP	FP	IRI	MR	OCT	Der	End	Mic	X-Ray	US
MiniGPT-4 [128]	-	22.81	-	38.33	-	27.48	31.40	40.25	-	36.23	38.30	25.50
BLIP-2 [73]	-	56.74	-	46.24	-	41.32	68.08	40.65	-	50.40	67.58	37.27
InstructBLIP [41]	-	28.72	-	50.31	-	33.15	42.59	62.22	-	46.29	61.04	41.25
mPLUG-Owl [123]	-	24.54	-	36.92	-	29.90	43.76	36.10	-	27.25	28.92	21.40
Otter [70]	-	18.53	-	37.51	-	26.06	29.64	42.64	-	27.48	31.85	23.49
LLaVA [77]	-	17.73	-	47.11	-	26.72	33.73	49.74	-	28.87	30.70	18.66
LLaMA_Adapter_v2 [48]	-	21.41	-	50.74	-	26.63	33.00	51.76	-	38.66	46.44	34.05
VPGTrans [125]	-	21.26	-	45.02	-	25.44	25.14	45.01	-	34.70	46.64	25.45
Med-Flamingo [84]	-	38.47	-	30.12	-	40.56	26.51	32.43	-	19.93	30.34	24.64
RadFM [117]	-	27.56	-	36.89	-	24.06	32.80	39.21	-	27.96	30.95	16.57
MedVInT [127]	-	40.74	-	31.84	-	43.10	23.26	29.11	-	32.00	55.10	41.26
LLaVA-Med [71]	-	18.69	-	39.03	-	27.47	34.61	44.95	-	33.29	30.68	29.88

Table 13. The overall accuracy of representative LVLMs on completely open-access data of our OmniMedVQA in terms of different modalities. Here, we report the accuracy of all items within each modality when utilizing **Prefix-based score**. Specifically, **Co** denotes Colposcopy, **CT** denotes Computed Tomography, **DP** denotes Digital Photography, **FP** denotes Fundus Photography, **IRI** denotes Infrared Reflectance Imaging, **MR** denotes Magnetic Resonance Imaging, **OCT** denotes Optical Coherence Tomography, **Der** denotes Dermoscopy, **End** denotes Endoscopy, **Mic** denotes Microscopy Images, **US** denotes Ultrasound. Meanwhile, in each column, the best and second-best performance are marked in red and blue, respectively.

Model	Co	CT	DP	FP	IRI	MR	OCT	Der	End	Mic	X-Ray	US
MiniGPT-4 [128]	-	29.46	-	29.36	-	12.63	3-	25.47	-	31.81	34.10	27.20
BLIP-2 [73]	-	38.87	-	28.83	-	20.64	18.70	19.90	-	49.91	48.14	81.79
InstructBLIP [41]	-	35.66	-	38.85	-	14.86	51.74	29.81	-	41.32	36.76	58.51
mPLUG-Owl [123]	-	37.00	-	49.00	-	13.10	41.76	18.94	-	36.73	28.78	29.14
Otter [70]	-	32.55	-	25.94	-	10.64	45.98	22.64	-	24.93	27.99	20.88
LLaVA [77]	-	38.27	-	20.10	-	4.36	51.23	14.27	-	23.57	23.67	20.74
LLaMA_Adapter_v2 [48]	-	36.03	-	30.34	-	17.35	54.22	21.81	-	35.23	37.54	47.51
VPGTrans [125]	-	30.30	-	28.83	-	10.89	31.60	22.43	-	34.15	34.02	40.89
Med-Flamingo [84]	-	22.25	-	36.74	-	14.07	58.57	39.78	-	45.95	38.09	17.42
RadFM [117]	-	45.47	-	28.22	-	24.70	37.40	25.79	-	28.72	54.66	24.78
MedVInT [127]	-	37.77	-	9.87	-	30.51	18.54	20.97	-	23.05	21.86	25.43
LLaVA-Med [71]	-	36.75	-	23.60	-	24.83	51.51	25.54	-	30.12	25.04	16.75

Vadim Sherman, Vipul, Wendy Kan, and Yau Ben-Or. Intel & mobileodt cervical cancer screening. https://kaggle.com/competitions/intel-mobileodt-cervical-cancer-screening, 2017. 2

[26] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019. 2

[27] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 2

[28] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017. 1

[29] Kimberlin van Wijnen Carole Sudre. Where is valdo - vascular lesions detection challenge 2021. https://valdo.grand-challenge.org/, 2021. 1

[30] Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, 12(1):4828, 2021. 2

[31] Carlo W Cereda, Søren Christensen, Bruce CV Campbell, Nishant K Mishra, Michael Mlynash, Christopher Levi, Matus Straka, Max Wintermark, Roland Bammer, Gregory W Albers, et al. A benchmarking tool to evaluate computer tomography perfusion infarct core predictions against a dwi standard. *Journal of Cerebral Blood Flow & Metabolism*, 36(10):1780–1789, 2016. 1

[32] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 2

[33] Pingjun Chen. Knee osteoarthritis severity grading dataset. *Mendeley Data*, 1:21–23, 2018. 2

[34] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[35] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 3

[36] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019. 2

[37] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020. 3

[38] Olivier Commowick, Michaël Kain, Romain Casey, Roxana Ameli, Jean-Christophe Ferré, Anne Kerbrat, Thomas Tourdias, Frédéric Cervenansky, Sorina Camarasu-Pop, Tristan Glatard, et al. Multiple sclerosis lesions segmentation from multiple experts: The miccai 2016 challenge dataset. *Neuroimage*, 244:118589, 2021. 1

[39] Will Cukierski. Histopathologic cancer detection. https://kaggle.com/competitions/histopathologic-cancer-detection, 2018. 2

[40] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485, 2022. 2

[41] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 3, 7, 8, 6

[42] Coen De Vente, Koenraad A Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, et al. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 2023. 2

[43] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014. 2

[44] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M Mossi, and Amparo Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, 18: 1–19, 2019. 2

[45] Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenole Quellec, Sarah Matta, et al. Adam challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022. 2

[46] Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Palm: Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *arXiv preprint arXiv:2305.07816*, 2023. 2

[47] Radovan Fusek. Pupil localization using geodesic distance. In *Advances in Visual Computing: 13th International Symposium, ISVC 2018, Las Vegas, NV, USA, November 19–21, 2018, Proceedings 13*, pages 433–444. Springer, 2018. 2

[48] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2, 3, 7, 8, 6

[49] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert systems with applications*, 42(19):6578–6585, 2015. 2

[50] 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*, pages 1820–1828, 2021. 2

[51] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942*, 2022. 2

[52] Anubha Gupta and Ritu Gupta. Isbi 2019 c-nmc challenge: Classification in cancer cell imaging. *Select Proceedings*, 2019. 2

[53] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical

imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 2

[54] Arsany Hakim, Søren Christensen, Stefan Winzeck, Maarten G Lansberg, Mark W Parsons, Christian Lucas, David Robben, Roland Wiest, Mauricio Reyes, and Greg Zaharchuk. Predicting infarct core from computed tomography perfusion in acute ischemia with machine learning: Lessons from the isles challenge. *Stroke*, 52(7):2328–2337, 2021. 1

[55] Ahmed Hamada. Br35h:: Brain tumor detection 2020. *Version 5*, 2020. 2

[56] Khaled Harrar. Texture characterization of bone radiograph images. application to osteoporosis diagnosis. 2014. 2

[57] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 1

[58] Towhidul Islam, Mohammad Arafat Hussain, Forhad Uddin Hasan Chowdhury, and BM Riazul Islam. A web-scraped skin image database of monkeypox, chickenpox, smallpox, cowpox, and measles. *biorxiv*, pages 2022–08, 2022. 2

[59] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4 (6):475, 2014. 3

[60] Soroush Javadi and Seyed Abolghasem Mirroshandel. A novel deep learning method for automatic assessment of human sperm images. *Computers in biology and medicine*, 109:182–194, 2019. 2

[61] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in Neural Information Processing Systems*, 35: 36722–36732, 2022. 1

[62] Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. Kaggle, 2019. 2

[63] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018. 2

[64] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018. 2

[65] Michael Kistler, Serena Bonaretti, Marcel Pfahrer, Roman Niklaus, and Philippe Büchler. The virtual skeleton database: an open access repository for biomedical research and collaboration. *Journal of medical Internet research*, 15 (11):e245, 2013. 1

[66] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 2

[67] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. *arXiv preprint arXiv:2311.15826*, 2023. 1

[68] Hugo J Kuijf and E Bennink. Grand challenge on mr brain segmentation at miccai 2018, 2019. 1

[69] 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. 1, 3

[70] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2, 3, 7, 8, 6

[71] 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*, 2023. 1, 2, 3, 7, 8, 6

[72] Fei Li, Diping Song, Han Chen, Jian Xiong, Xingyi Li, Hua Zhong, Guangxian Tang, Sujie Fan, Dennis SC Lam, Weihua Pan, et al. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. *NPJ digital medicine*, 3(1):123, 2020. 2

[73] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5, 6, 7, 8

[74] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023. 3

[75] 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)*, pages 1650–1654. IEEE, 2021. 1, 3

[76] Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *Plos one*, 14(9):e0222025, 2019. 2

[77] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 7, 8, 6

[78] Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6), 2022. 2

[79] Oskar Maier, Bjoern H Menze, Janina Von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017. 1

[80] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5): e210315, 2022. 1, 2

[81] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph^2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013. 2

[82] Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. Mrbrains challenge: online evaluation framework for brain image segmentation in 3t mri scans. *Computational intelligence and neuroscience*, 2015:1–1, 2015. 1

[83] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 1

[84] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. 1, 2, 3, 7, 8, 6

[85] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[86] Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS One*, 11(2): e0149399, 2016. 2

[87] Thivya Narendran. Image set for retinal artery-vein nicking assessment. https://people.eng.unimelb.edu.au/thivun/projects/AV_nicking_quantification/. 2

[88] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570, 2020. 2

[89] Nikita V Orlov, Wayne W Chen, David Mark Eckley, Tomasz J Macura, Lior Shamir, Elaine S Jaffe, and Ilya G Goldberg. Automatic classification of lymphoma images with transform-based global features. *IEEE Transactions on Information Technology in Biomedicine*, 14(4):1003–1013, 2010. 4, 2

[90] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 3

[91] Andre GC Pacheco and Renato A Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in biology and medicine*, 116:103545, 2020. 2

[92] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 2

[93] Hady Ahmady Phoulady and Peter R. Mouton. A new cervical cytology dataset for nucleus detection and image classification (cervix93) and methods for cervical nucleus detection, 2018. 2

[94] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 164–169, 2017. 2

[95] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-yee Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022. 2

[96] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017. 3

[97] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, Liam Caffery, Emmanouil Chousakos, Noel Codella, Marc Combalia, Stephen Dusza, Pascale Guitera, David Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):34, 2021. 2

[98] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. End-to-end prostate cancer detection in bpmri via 3d cnns: effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical image analysis*, 73: 102155, 2021. 1

[99] Fabio Scarpa, Enrico Grisan, and Alfredo Ruggeri. Automatic recognition of corneal nerve structures in images from confocal microscopy. *Investigative ophthalmology & visual science*, 49(11):4801–4807, 2008. 2

[100] Fabio Scarpa, Xiaodong Zheng, Yuichi Ohashi, and Alfredo Ruggeri. Automatic evaluation of corneal nerve tortuosity

in images from in vivo confocal microscopy. *Investigative ophthalmology & visual science*, 52(9):6404–6408, 2011. 2

[101] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[102] Uğur Şevik, Cemal Köse, Tolga Berber, and Hidayet Erdöl. Identification of suitable fundus images using automated quality assessment methods. *Journal of biomedical optics*, 19(4):046006–046006, 2014. 2

[103] Fariba Shaker, S Amirhassan Monadjemi, Javad Alirezaie, and Ahmad Reza Naghsh-Nilchi. A dictionary learning approach for human sperm heads classification. *Computers in biology and medicine*, 91:181–190, 2017. 2

[104] Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. Tiny lvlm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729*, 2023. 6

[105] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2

[106] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000. 3

[107] Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pages 2020–04, 2020. 4, 2

[108] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015. 2

[109] John Suckling. The mammographic images analysis society digital mammogram database. In *Exerpta Medica. International Congress Series, 1994*, pages 375–378, 1994. 3

[110] Siham Tabik, Anabel Gómez-Ríos, José Luis Martín-Rodríguez, Iván Sevillano-García, Manuel Rey-Area, David Charte, Emilio Guirado, Juan-Luis Suárez, Julián Luengo, MA Valero-González, et al. Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images. *IEEE journal of biomedical and health informatics*, 24(12):3595–3605, 2020. 3

[111] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[112] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 2

[113] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020. 3

[114] Shuo Wang, Chen Qin, Chengyan Wang, Kang Wang, Haoran Wang, Chen Chen, Cheng Ouyang, Xutong Kuang, Chengliang Dai, Yuanhan Mo, et al. The extreme cardiac mri analysis challenge under respiratory motion (cmrxmotion). *arXiv preprint arXiv:2210.06385*, 2022. 1

[115] Stefan Winzeck, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov, Egor Krivov, Mikhail Belyaev, Miguel Monteiro, et al. Isles 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral mri. *Frontiers in neurology*, 9:679, 2018. 1

[116] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023. 1

[117] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023. 1, 2, 3, 7, 8, 6

[118] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021. 1

[119] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 11:759007, 2021. 2

[120] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023. 2, 5, 6

[121] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 1

[122] He Xuehai, Zhang Yichen, Mou Luntian, Xing Eric, and Xie Pengtao. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1, 3

[123] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers

large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 3, 7, 8, 6

[124] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, Paras Lakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation, 2019. 3

[125] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023. 2, 3, 7, 8, 6

[126] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 3

[127] 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. *arXiv preprint arXiv:2305.10415*, 2023. 1, 2, 3, 6, 7, 8

[128] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 3, 7, 8, 6