Task-Routed Low-Rank Adaptation with Grounded Supervision for Medical Visual Question Answering

Mengkang Lu[1], Rongze Ma[1], Qingjie Zeng[1], Zilin Lu[1], and Yong Xia[1,2()]

[1] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
[2] Ningbo Institute of Northwestern Polytechnical University, Ningbo 315048, China.
yxia@nwpu.edu.cn

Abstract. Medical VQA combines discrete recognition with free-form evidence, but a shared adaptation objective can blur their requirements, while low-resolution preprocessing may remove diagnostic detail. We propose a parameter-efficient framework integrating data-calibrated native-detail preservation, concise image-grounded targets, and task-routed low-rank adapters. A general adapter is trained on all tasks while freezing the vision encoder and multimodal projector; an open-answer specialist is further trained on balanced free-form examples. At inference, question type deterministically selects the adapter, while schema-first generation and failure-triggered recovery isolate serialization errors. We evaluate Qwen3-VL on a 512-case holdout from 17,722 Med-CMR-derived cases across 12 modalities, 11 organ systems, and seven reasoning categories. Multiple-choice accuracy improves from 94.71% to 96.63% at 8B; 32B reaches 97.12% (404/416; exact paired $p = 0.0129$ versus baseline). On 96 open-ended cases, answer/evidence-trace token F1 rises from 0.231/0.240 to 0.253/0.250, although bootstrap intervals include zero. The 32B setup trains 536.9M parameters ($\sim 1.68\%$), avoiding full-model fine-tuning.

Keywords: Medical visual question answering · Multimodal learning · Parameter-efficient fine-tuning · Grounded reasoning

1 Introduction

Visual question answering (VQA) extends image recognition from a fixed label space to conditional, language-mediated decisions [1]. In medicine, the same interface may ask for a categorical diagnosis, the location of an abnormality, a short description of visual evidence, or a causal interpretation. Earlier medical VQA datasets emphasized radiology, pathology, or knowledge-enhanced questions [8, 6, 10]; newer resources broaden the image and question distribution substantially [17, 4]. This breadth is useful, but it makes adaptation difficult: multiple-choice (MCQ) questions reward exact option selection, whereas open-ended questions require concise language that remains supported by the image.

Two practical failure modes motivate this work. First, uniform resizing is not neutral. Small lesions, thin interfaces, cellular morphology, and spatial relations can disappear when images are forced below a conservative pixel budget. Second, a single adapter trained on a mixture dominated by MCQ examples receives relatively little signal about free-form answer style and visual evidence. Simply increasing trainable parameters does not necessarily resolve either problem, and full-model optimization is costly for modern multimodal large language models (MLLMs). Low-rank adaptation (LoRA) offers a memory-efficient alternative [7], but its application still requires decisions about which modules to adapt, what targets to supervise, and whether heterogeneous tasks should share one parameter update.

We address these issues with a task-routed, grounded adaptation framework for medical VQA. It preserves images up to a pixel cap selected from the empirical resolution distribution, learns an explicit but concise visual-evidence trace alongside each answer, and separates a general LoRA adapter from an open-answer specialist over one frozen multimodal backbone. Unlike token-level mixture-of-expert routing, the route is known from the requested output type, so no learned gate or additional inference model is required. A schema-first decoder and selective recovery policy further prevent malformed strings from being conflated with reasoning errors.

Our contributions are threefold. (1) We formulate a data-calibrated preprocessing and grounded-target pipeline that retains native detail for the complete released image distribution while discouraging unsupported clinical narrative. (2) We introduce a two-stage, task-routed LoRA procedure that specializes free-form behavior without duplicating the 32B backbone. (3) We provide controlled evaluation across 8B and 32B scales, paired significance tests, task-level error analysis, and ablations of adapter scope, resolution, specialization, and target cleaning. The resulting study is a method analysis rather than a report of a hidden evaluation or ranking.

2 Related Work

2.1 Medical Visual Question Answering

The clinician-authored VQA-RAD dataset covers radiology [8]; PathVQA focuses on pathology [6]; and SLAKE adds semantic labels and knowledge [10]. PMC-VQA expands scale using biomedical literature [17]. Med-CMR organizes multimodal reasoning by modality, organ system, and fine-grained capability, including long-tail, causal, spatial, temporal, small-object, and multi-source reasoning [4]. These resources shift the problem from short closed questions toward heterogeneous outputs requiring correctness and visual grounding.

Biomedical MLLMs such as LLaVA-Med [9] and Med-Flamingo [13] adapt general vision-language architectures through biomedical instruction data or few-shot multimodal context. The present work instead starts from Qwen3-VL [2] and studies how supervised adaptation should be organized when a single corpus

contains both exact-choice and free-form reasoning tasks. We do not interpret generated traces as a guarantee of faithful causal reasoning; they are supervised descriptions whose overlap with reference visual evidence can be measured.

2.2 Efficient and Task-Specialized Adaptation

LoRA represents an update using low-rank factors while leaving pretrained weights fixed [7]. Medical variants have explored parameter-efficient tuning of multimodal models [5] and mixture-of-expert LoRA modules for multitask medical language problems [11]. Our method differs in two respects. It freezes the vision tower and projector after an empirical adapter-scope screen, and it uses deterministic task-level routing rather than a learned token- or instance-level gate. This is appropriate because the output contract (one option versus an open response) is observed before inference. Chain-of-thought supervision can improve intermediate computation [15], but verbose rationales may introduce unsupported claims. We therefore supervise a short, ordered evidence description and audit its lexical agreement separately from answer agreement.

3 Method

3.1 Problem Formulation

Let $x = (I, q, o)$ contain a medical image I, question q, and optional choices o. The observable type is $z \in \{\text{mcq}, \text{open}\}$, and the target $y = (r, a)$ contains an evidence trace r and answer a. We model $p_{\theta,\phi_z}(r, a \mid x, z)$ autoregressively, with frozen backbone θ and task-selected update ϕ_z. Cross-entropy is applied only to assistant tokens. The presence of choices supplies z, so routing adds neither a classifier nor routing uncertainty.

3.2 Native-Detail Preprocessing

The image processor maps an input of height H and width W to (H', W') while preserving aspect ratio and satisfying

$$P_{\min} \le H'W' \le P_{\max}, \qquad P_{\min} = 4{,}096, \quad P_{\max} = 589{,}824. \tag{1}$$

The upper bound was selected before training by auditing all 17,722 available images. Their median size is 512×384, the maximum area is 523,264 pixels, and 2,814 images (15.88%) exceed the 262,144-pixel baseline cap. Thus, the new bound retains every available image at native area (subject to the model processor's patch divisibility) rather than enlarging the bound arbitrarily. The input sequence limit is increased from 2,048 to 4,096 tokens to accommodate the additional visual tokens and the structured target. The same bounds are used in training and evaluation.

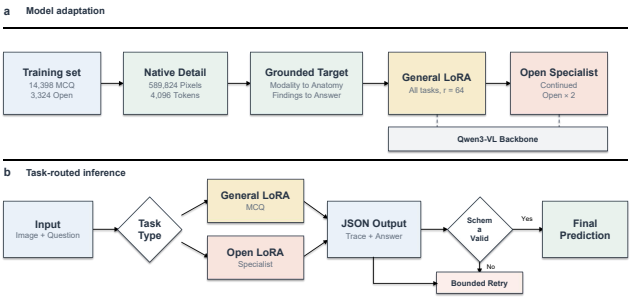

Fig. 1. Overview of the framework. (a) Images retain native detail up to a data-derived pixel cap; a frozen visual stack feeds a shared language backbone, while the dashed path denotes grounded training supervision. (b) Task type deterministically selects the general adapter for MCQ or the continued specialist for open-ended questions. Both produce a parseable evidence trace and answer.

3.3 Grounded Target Construction

Each example provides an answer and image description, serialized into *reasoning* and *answer* fields. The instruction requests a one-to-three-sentence chain ordered by modality, anatomy, visible location/laterality, morphology or spatial relation, and conclusion; it disallows unsupported history, procedures, measurements, and certainty. MCQ answers are normalized to the option alphabet, while open answers are concise spans. The trace is an auditable description target, not a claim to expose latent causal computation.

We retain the source visual descriptions after light serialization instead of aggressively deleting sentences that resemble case context. The decision is empirical: a cleaning ablation reduced both open-answer and evidence-trace F1 (Sec. 5.2). Escaping and field validation are performed before training, and examples with missing images or targets are rejected; none of the 17,722 released cases was missing an image.

3.4 Task-Routed Low-Rank Adaptation

For a language-layer projection $W_0 \in \mathbb{R}^{d\times k}$, LoRA defines

$$W_z = W_0 + \frac{\alpha}{r} B_z A_z, \quad A_z \in \mathbb{R}^{r\times k}, \quad B_z \in \mathbb{R}^{d\times r}, \tag{2}$$

with rank $r = 64$, scale $\alpha = 128$, and dropout 0.05. Updates are attached to the query, key, value, output, gate, up, and down projections of the language model. The vision encoder and multimodal projector remain frozen. The 32B

configuration trains 536.9M parameters, approximately 1.68% relative to the nominal backbone size, and stores two adapters over one shared backbone.

Training has two stages under $\mathcal{L}_\lambda = \sum_{\mathcal{D}_{\text{mcq}}} \ell(x, y) + \lambda \sum_{\mathcal{D}_{\text{open}}} \ell(x, y)$. The *general stage* learns ϕ_g with $\lambda = 1$. The *specialist stage* initializes $\phi_o \leftarrow \phi_g$ and continues with $\lambda = 2$, implemented by duplicating each open example. At inference, ϕ_g serves MCQ and ϕ_o serves open questions. Short continuation changes the relative open exposure from 18.8% to 31.6% without discarding closed questions. Exposure and stopping fraction are chosen on 8B, then fixed for 32B.

3.5 Schema-First Generation and Recovery

Greedy decoding emits the required JSON object with at most 384 new tokens. A deterministic parser extracts it, normalizes MCQ answers, and requires non-empty fields. Only a failed parse, invalid option, or severe repetition triggers regeneration. Three bounded tiers increase repetition penalty from 1.05 to 1.10 and 1.15 and set no-repeat n-grams to 8, 6, and 4. Valid outputs are never resampled; parse success is reported separately.

4 Experimental Design

4.1 Data and Leakage-Controlled Split

Experiments use 17,722 released cases derived from Med-CMR [4]: 14,398 MCQ and 3,324 open-ended cases. The images cover 12 modality groups and 11 organ systems. The largest modality groups are pathology/microscopy (4,767), photography/dermatology (3,986), CT (2,828), radiography/fluoroscopy (2,180), and MRI (1,623); the remaining seven groups contribute 2,338 cases. Reasoning labels include long-tail generalization (12,498), causal reasoning (2,046), small-object perception (1,087), fine-detail recognition (748), multi-source integration (709), spatial reasoning (340), and temporal reasoning (294).

A SHA-256 hash of the stable case identifier and seed 2026 defines a reproducible task-stratified holdout. It contains 512 cases (416 MCQ and 96 open); the other 17,210 cases form the training set. The manifest is fixed for every run. No holdout record is used for gradient updates, stopping selection, prompt editing, or the pixel-cap audit beyond global unlabeled image dimensions. For the specialist corpus, duplicating the 3,228 open training cases yields 20,438 training instances.

4.2 Metrics and Statistical Analysis

MCQ accuracy is exact match after option normalization. For open questions, we compute token F1 after lowercasing, punctuation/article removal, and whitespace normalization, following common extractive-QA practice [14]. Answer F1 compares a with the reference answer; trace F1 compares r with the reference

visual description. These are inexpensive lexical-overlap diagnostics, not semantic or clinical correctness measures. We therefore report them separately and avoid interpreting small changes as clinical improvements. JSON validity is the proportion satisfying the required schema after bounded recovery.

For MCQ proportions we report Wilson 95% confidence intervals [16]. Pairwise MCQ changes are tested with the exact two-sided McNemar test [12]. Open-ended changes use 20,000 paired bootstrap resamples with seed 2026 [3]; uncertainty is substantial because $n = 96$. All comparisons use identical examples and decoding limits.

4.3 Implementation

We instantiate Qwen3-VL-Thinking at 8B and 32B scales [2]. Training uses bfloat16, TF32 matrix operations, fused AdamW, cosine decay, 0.03 warmup ratio, weight decay 0.01, gradient checkpointing for 32B, and seed 2026. The 32B general stage runs two epochs at learning rate 5×10^{-5} with per-device batch one and four-step accumulation on seven H20 GPUs (global batch 28); it completes 1,230 updates in 1.97 h. The specialist runs 0.4 epoch at 10^{-5}, and checkpoint 250 is selected by held-out loss. The 8B high-resolution stage continues the selected low-resolution adapter for one epoch at 2×10^{-5}; its specialist uses the same rank and 10^{-5} learning rate. No quantization is used in training.

The primary baseline is the selected 8B language-only rank-64 LoRA trained for two epochs with a 262,144-pixel cap and 2,048-token limit. It uses the same backbone family, split, output schema, and recovery policy. We additionally compare a 32B general adapter without open specialization and report preliminary 8B adapter-scope screens. All reported scores are from local labeled data; no external or hidden-test scores enter the analysis.

5 Results

5.1 Main Comparison

Table 1 shows the effect of resolution, task routing, and scale. High-resolution continuation raises 8B MCQ accuracy by 1.92 percentage points, from 94.71% to 96.63%, and open specialization recovers the small answer/trace loss of the general high-resolution adapter. Scaling the same design to 32B yields 97.12% MCQ accuracy. Its routed open output reaches 0.253 answer F1 and 0.250 trace F1, improvements of 0.022 and 0.010 over the baseline, respectively.

The 32B general adapter has the highest lexical answer overlap, while the open specialist trades 0.0047 answer F1 for a 0.0055 gain in evidence-trace F1 and shortens mean open answers from 10.7 to 9.1 words. This trade-off is visible rather than collapsed into an arbitrary scalar. The routed system chooses the specialist because its purpose is grounded, concise free-form evidence; applications prioritizing answer-span overlap can instead route open questions to the general adapter without retraining.

Table 1. Results on the fixed 512-case holdout. Open metrics use the 96 open-ended cases; MCQ uses 416 cases. "General" denotes one adapter for both tasks; "routed" uses the general adapter for MCQ and the open specialist for open questions. Best values are bold.

Backbone / adaptation	Pixel cap	Route	MCQ acc.	Ans. F1	Trace F1	JSON
8B baseline LoRA	262k	General	94.71	0.231	0.240	99.41
8B high-resolution	590k	General	96.63	0.228	0.239	99.80
8B proposed	590k	Routed	96.63	0.236	0.246	**100**
32B high-resolution	590k	General	**97.12**	**0.258**	0.245	**100**
32B proposed	590k	Routed	**97.12**	0.253	**0.250**	**100**

Compared with the baseline, the proposed 32B system corrects 12 MCQ cases and regresses on two. The exact paired McNemar test gives $p = 0.0129$. Accuracy rises from 394/416 (Wilson 95% CI: 92.12–96.48%) to 404/416 (95.03–98.34%). Against routed 8B, it corrects seven and regresses on five ($p = 0.774$), so the observed two-case scale gain is not statistically resolved. For open questions, the paired bootstrap confidence intervals for the 32B-versus-baseline changes are $[-0.006, 0.050]$ for answer F1 and $[-0.011, 0.031]$ for trace F1. These intervals include zero and motivate cautious interpretation.

5.2 Ablation and Architecture Screening

Table 2 summarizes an initial 8B screen performed before high-resolution and specialization experiments. Language-only LoRA achieves the lowest validation loss and best open F1 while training only 174.6M parameters. Extending LoRA to the vision tower for one epoch does not improve robust MCQ accuracy and reduces open F1. One-epoch full-language GaLore trains 8.19B parameters but performs substantially worse. The epoch counts and trainable scopes differ, so this table supports configuration selection rather than a claim of an isolated algorithmic advantage.

Table 2. Preliminary 8B adapter-scope screen at the 262k pixel cap. Times are wall-clock training times on the same node. The designs are intentionally reported as a selection study, not a fully controlled component ablation.

Adaptation scope	Trainable	Epochs	Time	Eval loss	MCQ / open F1
Language LoRA, $r = 64$	174.6M	2	37:32	**0.6719**	**94.71** / **0.231**
Vision+language LoRA, $r = 32$	104.4M	1	20:54	0.6907	94.71 / 0.206
Full language, GaLore $r = 128$	8.19B	1	34:49	0.7062	82.69 / 0.180

The high-resolution stage and task routing are more directly comparable because they share the selected adapter lineage. High-resolution continuation corrects eight baseline MCQ errors without a paired regression on this hold-out. Adding the specialist leaves MCQ unchanged by construction and increases

open answer/trace F1 from 0.228/0.239 to 0.236/0.246. A separate specialist trained after removing identifier-like and history-like source sentences reaches 0.235 answer F1 but only 0.223 trace F1; indiscriminate cleaning evidently removes useful visible context as well as undesirable phrasing. We therefore retain the original visual descriptions and control hallucination through the concise grounding instruction.

5.3 Error and Reliability Analysis

Most MCQ corrections occur in long-tail generalization ($n = 300$): accuracy rises from 93.7% to 95.3% at 8B and 96.7% at 32B. Causal accuracy rises from 94.9% to 100% ($n = 39$), but 32B multi-source accuracy falls from 100% to 92% relative to routed 8B ($n = 25$). Small-object, fine-detail, spatial, and temporal subsets contain only 7–26 cases and are at ceiling, so they cannot establish that those capabilities are solved. Before recovery, JSON success is 98.05% for the 32B general adapter and 95.83% for the specialist; bounded retries raise both to 100% without resampling valid outputs. On a separate released unlabeled set, the routed models generate 2,532 unique, non-empty predictions (2,057 MCQ and 475 open), all accepted by the reference schema validator. This supports interface completeness, not correctness.

6 Discussion

The largest 8B MCQ step follows high-resolution continuation, consistent with nearly one sixth of images exceeding the baseline pixel budget. This is an association, not an isolated resolution effect, because context length and optimization exposure also change. Task routing addresses a separate imbalance: MCQ dominates the corpus, whereas open questions require both an answer and evidence. The specialist changes their exposure without duplicating the backbone, but its trace gain accompanies a small answer-F1 loss. Disaggregated endpoints therefore reveal a trade-off hidden by a single aggregate score.

The study has important limitations. It uses one split from one corpus and only 96 labeled open cases; external generalization remains unknown. Token F1 neither recognizes all valid paraphrases nor establishes clinical correctness, faithfulness, calibration, or safety. The models process single images rather than longitudinal series or volumes. Adapter-scope screening is confounded by epoch count, while resolution, context length, and continued training are not factorially separated. Finally, a plausible generated trace is not mechanistic interpretability. External datasets, clinician evidence ratings, semantic entailment checks, and controlled resolution experiments are necessary next steps; adaptive visual-token allocation could also reduce compute.

7 Conclusion

We combined data-calibrated detail preservation, grounded supervision, task-routed LoRA, and schema-aware recovery for heterogeneous medical VQA. On a

fixed holdout, the 32B system significantly improves MCQ accuracy over the 8B baseline while training about 1.68% of the backbone; open-metric gains remain uncertain. Visual resolution, supervision, task balance, and interface reliability should be designed together and evaluated separately.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 92470101, Science and Technology Program of Haicang District of Xiamen, China, under Grant 350205Z20242013, in part by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang, China, under Grant 2025C01201(SD2), and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025019 and Grant CX2026054.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2425–2433 (2015). https://doi.org/10.1109/ICCV.2015.279
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall/CRC, New York (1993)
4. Gong, H., Ji, X., Liu, Y., Wu, W., Yan, X., Liu, J., Wu, K., Pan, J., Jian, B., Zhang, J., Hu, X., Li, H.B.: Med-CMR: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning. arXiv preprint arXiv:2512.00818 (2025)
5. He, J., Li, P., Liu, G., He, G., Chen, Z., Zhong, S.: PeFoMed: Parameter-efficient fine-tuning of multimodal large language models for medical imaging. arXiv preprint arXiv:2401.02797 (2024)
6. He, X., Zhang, Y., Mou, L., Xing, E.P., Xie, P.: PathVQA: 30,000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
7. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
8. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**, 180251 (2018). https://doi.org/10.1038/sdata.2018.251

9. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
10. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. pp. 1650–1654 (2021). https://doi.org/10.1109/ISBI48211.2021.9434010
11. Liu, Q., Wu, X., Zhao, X., Zhu, Y., Xu, D., Tian, F., Zheng, Y.: When MoE meets LLMs: Parameter-efficient fine-tuning for multi-task medical applications. arXiv preprint arXiv:2310.18339 (2023)
12. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947). https://doi.org/10.1007/BF02295996
13. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Reis, E.P., Rajpurkar, P., Leskovec, J.: Med-Flamingo: A multimodal medical few-shot learner. arXiv preprint arXiv:2307.15189 (2023)
14. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392 (2016). https://doi.org/10.18653/v1/D16-1264
15. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 24824–24837 (2022)
16. Wilson, E.B.: Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association **22**(158), 209–212 (1927). https://doi.org/10.1080/01621459.1927.10502953
17. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Development of a large-scale medical visual question-answering dataset. Communications Medicine **4**, 277 (2024). https://doi.org/10.1038/s43856-024-00709-2