MIRA: Modality-Adaptive Image-Grounded Reasoning and Answering for Medical Visual Question Answering under Domain Shift

Sarth Santosh Shah[2,3], Adinath Madhavrao Dukre[1], and Imran Razzak[1]

[1] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] MedOS, Abu Dhabi, UAE
[3] National Institute of Technology Karnataka, Surathkal, India

Abstract. Medical visual question answering under domain shift asks that the answer be right *and* that the reasoning be faithful to the image, not fluent narration from language priors. The MedReason 2026 challenge makes the second requirement decisive, capping the visual accuracy of an answer whenever the reasoning is not grounded in the image. We present **MIRA** (Modality-adaptive Image-grounded Reasoning and Answering), an offline system built on a nine-billion-parameter multimodal model and designed backwards from that rule. Its centrepiece is *grounding-aware chain-of-thought*: rather than sampling a teacher and keeping the traces that happen to be correct, we hand a frozen teacher the reference answer and supervise only its visual justification, so the teacher explains a known answer instead of searching for one. A two-stage low-rank adaptation trains under a token-weighted loss mirroring the rubric; a metadata-free modality router and synthetic out-of-distribution acquisitions address the shifted test split. On an 886-case held-out slice MIRA answers **96.2%** of multiple-choice questions correctly (against 34.7% un-adapted, on a smaller subset) and reasons with a judge-rated faithfulness of **3.45/4**. A case-matched ablation ($n = 47$) associates grounded supervision with a rise in faithfulness $2.13 \rightarrow \mathbf{3.43}$ and in gated visual accuracy $2.11 \rightarrow \mathbf{2.40}$, while answer correctness moves little ($1.60 \rightarrow 1.70$), locating the remaining difficulty in the answers rather than in the grounding.

Keywords: Medical VQA · Grounded reasoning · Domain shift · Chain-of-thought · Parameter-efficient fine-tuning · Multimodal LLM.

1 Introduction

Medical visual question answering promises a natural interface to clinical decision support: a model reads an image, answers a question, and shows its reasoning so that a clinician can check it [10,13]. A model can nonetheless produce the right words for the wrong reasons, leaning on language priors until the explanation is untethered from the evidence. MedReason 2026 makes faithfulness part of the score: on open-ended questions it judges whether the answer's visual claims hold and whether the reasoning trace stays anchored to the image, and lets the second

gate the first, capping visual accuracy however plausible the answer sounds [12]. Grounding is a training target, not a presentational nicety.

Two obstacles stand in the way. Chain-of-thought supervision as usually practised is text-centric [15]: it teaches a model to emit reasoning tokens without redirecting where it looks, which fails on the small nodules and focal lesions on which diagnosis turns. The evaluation is also unforgiving of engineering slips: an earlier eight-billion-parameter attempt of ours answered only a third of the multiple-choice questions because a short budget truncated its thinking. Existing remedies miss what the challenge rewards: alignment methods improve answers at the response level [10,13], while reasoning distillation copies an answer-blind teacher's mistakes into the supervision. Neither trains a visual justification for a known-correct answer, which is what the gate measures. So we invert the recipe, supervising that justification to make visual faithfulness rather than answer correctness the explicit training target. **MIRA** is the resulting system (Fig. 1). Our contributions are:

(i) **Grounding-aware chain-of-thought** (GA-CoT), supervising the teacher only on its visual justification of a given answer; (ii) a **two-stage low-rank adaptation** under a rubric-aligned, token-weighted objective, lifting multiple-choice accuracy from 34.7% to 96.2%; (iii) a **metadata-free modality router**, synthetic out-of-distribution training, and a guaranteed-conclusion decoding rule, removing the earlier truncation failure; and (iv) a **case-matched ablation** attributing the gated visual score chiefly to grounded reasoning and locating the headroom in answer correctness.

2 Related Work

Benchmarks and reasoning supervision. VQA-RAD [9], PathVQA [7], and SLAKE [11] reward the final answer alone; MedReason 2026 also scores how faithfully the reasoning follows the image [12]. Chain-of-thought prompting [15] improves multi-step reasoning but remains text-centric, and the usual distillation recipe filters an answer-blind teacher by final correctness; we instead anchor the chain to explicit observations and condition the teacher on the answer.

Medical VLMs and efficient adaptation. LLaVA-Med [10] and Med-Flamingo [13] optimise response-level objectives rather than a gated faithfulness one. We adapt a member of the Qwen multimodal family [1,2] (which also supplies the challenge's visual-accuracy judge) with low-rank adapters [8,5] over its language, vision, and gated-linear-attention projections, accelerated by FlashAttention [4]. Rather than a learned domain-adaptation head we meet the shift with a metadata-free random-forest router [3], light contrast normalisation [14], and synthetic exposure to the target acquisitions.

3 Challenge and Problem Formulation

Task and data. Each case presents a medical image, an optional context, and a question. Closed-ended cases require selecting one option; open-ended cases

Table 1. Modality distribution of the 17,722-case development pool; counts cover the 17,223 single-modality cases. Radiography and MRI, in bold, are the source pools for synthetic augmentation and for the modality router.

Path./Micro.	Photo./Derm.	CT	**Radiog. & Fluoro.**	**MRI**	Ultrasound	Endoscopy	Ophthal./Opt.	Nucl. Med./PET
4,752	3,953	2,711	**2,061**	**1,545**	683	606	498	240

Physiologic signals: 174.

require *both* a reasoning trace and a final answer, the answer alone being insufficient [12]. The labelled development pool holds 17,722 cases (14,398 closed-ended, 3,324 open-ended) over ten imaging modalities (Table 1). We set aside a fixed five-percent internal slice of 886 cases before any training and adapt on the remaining 16,836. A validation set of 2,532 cases and the hidden test sets withhold their answers; the private split is drawn from out-of-distribution acquisitions, namely low-dose radiographs and high-field (7T) brain MRI. Evaluation runs on the organisers' side from a submitted container.

Metrics. Three quantities are reported [12]. *Multiple-choice accuracy* is the exact match of the chosen option. *Answer correctness* $\mathsf{GT}_{\mathsf{final}} \in \{0, \ldots, 4\}$ is the medical and semantic agreement of the final answer with the reference, judged by a text-only Llama-3.1-70B model [6]. *Visual accuracy* $\mathsf{VA}_{\mathsf{final}} \in \{0, \ldots, 4\}$ is judged by a Qwen-2.5-VL model [2] that sees the image and yields two sub-scores: $\mathsf{VA}_{\mathsf{ans}}$, whether the answer's visual claims are supported, and $\mathsf{RVF}_{\mathsf{trace}}$, whether the reasoning stays faithful. Crucially, the second gates the first:

$$\mathsf{VA}_{\mathsf{final}} = \begin{cases} \min(\mathsf{VA}_{\mathsf{ans}},\ 1) & \text{if } \mathsf{RVF}_{\mathsf{trace}} \leq 1, \\ \min(\mathsf{VA}_{\mathsf{ans}},\ 3) & \text{if } \mathsf{RVF}_{\mathsf{trace}} = 2, \\ \mathsf{VA}_{\mathsf{ans}} & \text{if } \mathsf{RVF}_{\mathsf{trace}} \geq 3. \end{cases} \tag{1}$$

The leaderboard averages per-metric ranks over the combined hidden test sets [12]. Equation (1) drives what follows: full visual-accuracy credit needs a faithfulness of at least three.

4 Method

Let π_θ be a multimodal model that, for a case $x = (x_v, x_q)$ of image x_v and question x_q, emits a reasoning trace τ within a thinking segment followed by an answer a; open-ended training cases carry a reference answer $a^\star$. Because Eq. (1) gates visual-accuracy credit on faithfulness, MIRA supervises τ and not only a (Fig. 1).

4.1 Base Model and Parameter-Efficient Adaptation

We build on a nine-billion-parameter multimodal model coupling a vision encoder to a gated linear-attention language backbone with a native thinking mode, loaded in four-bit precision [5] so it runs offline on a single accelerator; decoding relies on the backbone's fast linear-attention kernels, without which throughput

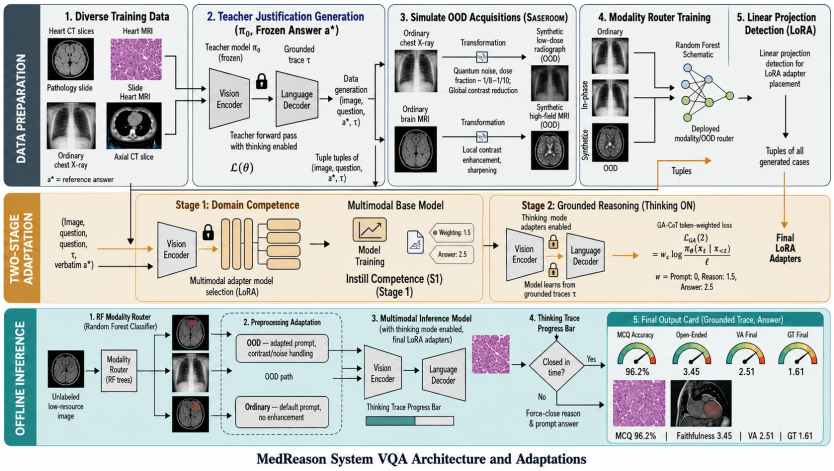

Fig. 1. MIRA end to end: data preparation builds two training corpora and a modality router; two-stage adaptation yields one 51 M-parameter adapter under a token-weighted loss; inference infers modality, adapts prompt and preprocessing, and closes the reasoning to guarantee a readable answer. Dashed arrows pass artefacts between phases.

falls roughly fifteenfold. We place low-rank adapters [8] on every linear projection of the three sub-networks (language attention and feed-forward layers, vision encoder, gated-linear-attention blocks), located automatically at load time. Adaptation touches 51 M parameters, 0.54% of the model; the output embedding stays frozen, since training a weight tied to the input embedding would break adapter portability at no gain.

4.2 Grounding-Aware Chain-of-Thought

Conventional distillation retains traces whose answer happens to be right, rewarding the teacher for solving the problem and inheriting its errors. We invert the arrangement: for each open-ended case we run the model as a frozen teacher π_0, giving it the image, the question, *and* the reference answer $a^\star$, and ask only that it account for $a^\star$ from the image. The teacher returns a two-part trace $\tau = (\tau_{\text{obs}}, \tau_{\text{rea}})$: first what is directly visible (structures, laterality, relative intensity, abnormalities, image quality), then a chain of inference in which each step refers back to one of those observations. Supplying $a^\star$ takes answer-finding out of the teacher's task and directs it towards an image-conditioned justification, but does not certify that each observation is present in the image: a teacher told the answer can still rationalise after the fact. Our filter is structural, in that we discard the teacher's own answer, keep the justification, append the verbatim reference $a^\star$ as the target, and retain a trace only if it opens from an explicit visual observation; the retained justifications were not separately verified against the images, so *grounded* should be read in that operational sense. This yields 3,149 traces, one per open-ended training case, of roughly 1,400 characters; a

Table 2. Principal settings; decoding is greedy with a bounded-then-closed budget.

Setting	Value	Setting	Value
Backbone / precision	nine-billion-param., 4-bit	Loss weights	reason. 1.5, ans. 2.5, prompt 0
Adapter rank / scale	16 / 32 (dropout 0.05)	Optimiser	8-bit AdamW, bf16
Trainable parameters	51 M (0.54%)	Grad. clip / decay	0.5 / 0.01
Effective batch / length	16 / 2048	Stage 1	2 epochs, lr 2×10^{-5}
Reasoning budget	768 / 1024 (closed/open)	Stage 2	2 epochs, lr 5×10^{-6}
Held-out slice	886 cases (fixed seed)	Router threshold	0.60

variant with the reasoning replaced by fixed boilerplate supports the ablation in Sect. 5.

4.3 Two-Stage Adaptation and the Rubric-Aligned Objective

Two stages separate two skills (Table 2). The *first stage is broad*: the model sees all training cases in a plain, answer-only format with thinking disabled, learning medical vocabulary, the concise answer style the judges expect, and the multiple-choice format; this is the stage that moves multiple-choice accuracy. The *second stage is narrow*: from the first stage's adapter at a fourfold smaller learning rate, it trains only on the grounded traces with thinking on, learning to reason from the image without unlearning what it knows. The loss matches the rubric, which reads the reasoning and the answer but not the prompt; writing w_t for the token weight, we mask the prompt and weight the reasoning and answer spans above the default:

$$\mathcal{L}(\theta) = -\frac{\sum_t w_t \, \log \pi_\theta(x_t \mid x_{<t})}{\sum_t w_t}, \qquad w_t = \begin{cases} 0 & t \in \text{prompt}, \\ 1.5 & t \in \text{reasoning}, \\ 2.5 & t \in \text{answer}. \end{cases} \tag{2}$$

4.4 Inferring Modality and Closing the Reasoning

At test time images arrive without the modality label available during training, and the hidden split is drawn from unseen acquisitions, so we infer modality from the image. Nine illumination- and scale-stable statistics (brightness, contrast, sharpness, noise, the balance of high- and low-frequency energy, colour saturation, intensity entropy, and the near-black and near-white pixel fractions) feed a random forest [3] separating ordinary images from the two synthetic out-of-distribution classes, reaching about 98% accuracy under five-fold cross-validation, a figure that measures recognition of our simulated transformations, not accuracy on real low-dose or 7T acquisitions. The router is cautious, leaving the ordinary-image default only when its confidence in an out-of-distribution class exceeds 0.60 *and* outweighs the ordinary-image score.

Given the inferred modality we adjust the system prompt (warning, for instance, of quantum-mottle noise on a low-dose radiograph) and apply a gentle, structure-preserving preprocessing step, kept mild because the faithfulness judge

reads the reasoning against the *original* image. Finally, the model thinks within a fixed budget; if it has not closed its reasoning by then we close it on its behalf and prompt directly for the answer, decoding only the tokens needed to name an option or state a short response. This guarantees a readable answer from completed and truncated reasoning alike, removing the failure that limited our earlier baseline.

4.5 Synthetic Out-of-Distribution Data and Answer Calibration

Synthetic acquisitions. We simulate the hidden acquisitions from ordinary training images: a low-dose radiograph by injecting dose-dependent quantum noise (at a sampled dose fraction between a tenth and a third of normal) and lowering global contrast toward mid-grey, a high-field scan by local contrast enhancement [14] then mild sharpening. Mixing about a fifth of the eligible post-split radiography and MRI cases as synthetic variants yields 671 images (377 low-dose, 294 high-field), entering the first stage with the matching modality prompt. The same transforms label the router, so detector and model share one definition of an out-of-distribution image.

Answer calibration. The answer-correctness judge reads text only and penalises verbose, reference-divergent answers, so we confine all reasoning to the thinking segment and require a concise, reference-style final answer, enforced when building the data by stripping the teacher's own answer: *what* the model concludes stays short enough for the text judge, while *why* stays in the trace for the visual judge.

5 Experiments and Results

Setup and data separation. We report the metrics of Sect. 3 on an *internal held-out slice*: the fixed five-percent sample of 886 cases (711 closed-ended, 175 open-ended), drawn with a fixed seed before any training and used neither for adaptation, nor for teacher traces, nor as a source for the synthetic acquisitions. Closed-ended cases use the official exact match; the open-ended judges replicate the official rubric and gating of Eq. (1), scoring visual accuracy with a model from the organisers' Qwen-2.5-VL family, a closely aligned proxy for the official judge rather than a reproduction of it, and answer correctness with a public Qwen2.5-72B model in place of the gated Llama-3.1-70B; both open-ended figures are internal estimates only. Separation is by released case identifier; as the release exposes no patient or study identifier, two images from one study may fall on opposite sides of the split, and we ran no near-duplicate detection over images or question templates.

Multiple-choice accuracy. On the held-out slice MIRA answers **96.2**% of multiple-choice questions correctly (684 of 711), far above the un-adapted base model (34.7% on a smaller subset) and our earlier eight-billion-parameter baseline (33.0%, lost to reasoning truncation). The first stage alone reaches **96.6**%, so specialising for reasoning does not cost the closed-ended skill (Table 3, left,

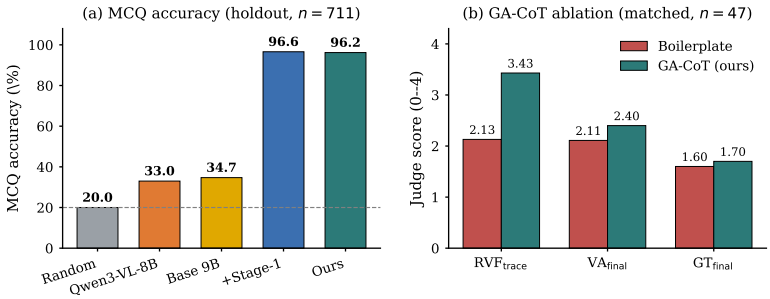

Fig. 2. Held-out slice results at a glance: multiple-choice accuracy across configurations (base model, earlier truncated 8B baseline, after stage 1, after both stages) alongside MIRA's mean open-ended scores ($\mathsf{RVF}_{\mathsf{trace}}$, $\mathsf{VA}_{\mathsf{ans}}$, $\mathsf{VA}_{\mathsf{final}}$, $\mathsf{GT}_{\mathsf{final}}$), summarising the numbers in Table 3.

Table 3. *Left*: multiple-choice accuracy ($n=711$; base model on a deterministic subset, 67/193, Wilson 95% interval [28.4, 41.7]%). *Centre*: MIRA's open-ended scores ($n=175$). *Right*: the grounding ablation ($n=47$, matched), same judges and cases; the control replaces each grounded trace with fixed boilerplate, varying the content of the supervision rather than its presence. Scores are 0–4.

Multiple choice ($n=711$)		**MIRA** ($n=175$)		**Grounding ablation** ($n=47$, matched)				
Configuration	Acc.	Score	Value	Reasoning	$\mathsf{RVF}_{\mathsf{trace}}$	$\mathsf{VA}_{\mathsf{ans}}$	$\mathsf{VA}_{\mathsf{final}}$	$\mathsf{GT}_{\mathsf{final}}$
Random choice (5 options)	20.0%	$\mathsf{RVF}_{\mathsf{trace}}$	3.45	Boilerplate	2.13	2.49	2.11	1.60
Earlier 8B baseline (truncated)	33.0%	$\mathsf{VA}_{\mathsf{ans}}$	2.51	**Grounded (ours)**	**3.43**	2.40	**2.40**	**1.70**
Base model, no adaptation	34.7%	$\mathsf{VA}_{\mathsf{final}}$	2.51	Difference	+1.30	−0.09	+0.30	+0.10
After first (domain) stage	96.6%	$\mathsf{GT}_{\mathsf{final}}$	1.61					
After both stages (ours)	**96.2%**							

Fig. 2). The base figure covers the $n=193$ closed-ended cases inside the first 240 of the slice, taken in deterministic order because un-adapted decoding is some twenty-five times slower per case (67/193, Wilson 95% interval [28.4, 41.7]%); not being case-matched against the full 711, it is a diagnostic reference point, not a controlled estimate of the gain.

Open-ended reasoning, and what grounding buys. For MIRA on open-ended cases faithfulness averages **3.45** of 4, gated and ungated visual accuracy both 2.51, and answer correctness 1.61, the weakest of the three (Table 3, centre; Figs. 2 and 3). To attribute the gain we retrain the grounding stage on the boilerplate traces and rescore, under the same judges, the 47 open-ended cases inside that 240-case slice (Table 3, right). Grounded traces raise faithfulness from 2.13 to **3.43** and, through the gate, visual accuracy from 2.11 to **2.40**, while answer correctness moves little (1.60 to 1.70). That the ungated answer-level score does not rise (2.49 to 2.40) while the gated one does is consistent with the improvement coming from fewer faithfulness caps rather than better visual claims, the mechanism Eq. (1) rewards. With $n=47$ on a five-point ordinal scale we report means without confidence intervals: the faithfulness difference (+1.30) is large relative to the scale, the answer-correctness one (+0.10) is not, and we draw no conclusion from the latter.

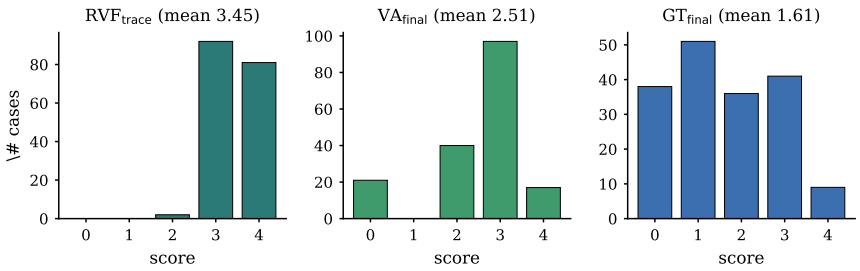

Fig. 3. Per-case open-ended score distributions on the held-out slice ($n = 175$). Judge-rated faithfulness and gated visual accuracy concentrate at the high end; answer correctness is the dispersed metric.

Validation coverage. With the validation answers withheld, we verify that the container's predictions cover all 2,532 cases (2,057 closed-ended, 475 open-ended) in the official format and spread evenly across the five options (390, 397, 441, 440, 389, no blanks), showing none of the single-letter collapse of the earlier truncated baseline.

6 Discussion and Conclusion

The results separate where each part of the design acts. **Adaptation earns the multiple-choice score**, closing the gap from near chance to 96.2%; **grounded supervision earns the gated visual score**, carrying faithfulness past the threshold Eq. (1) requires; and **answer correctness is the binding constraint**, at 1.61 of 4. Its small ablation difference (1.60 to 1.70) is the expected shape of this intervention: the second stage supervises the justification, and nothing in GA-CoT adds diagnostic knowledge the backbone lacks. $\mathsf{GT}_{\mathsf{final}}$ thus reflects a backbone-knowledge ceiling rather than a shortfall of reasoning supervision, and the levers that would move it, a larger backbone, retrieval, or self-consistency, are orthogonal to grounding.

Caveats bound the numbers. We report no matched comparison with published medical vision-language models, so our internal references bound our own components but not our standing in the literature. The held-out slice shares the training distribution and separates by identifier rather than patient; neither open-ended judge is the official one; synthetic acquisitions are not the real thing, and a different scanner would erode the router's precision; and the evidence rests on one split and a 47-case ablation without uncertainty estimates. Treating the gate as the design driver nonetheless yields **96.2**% accuracy at **3.45**/4 faithfulness. Adapter, router and container will be released.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023), https://arxiv.org/abs/2308.12966
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems **35**, 16344–16359 (2022)
5. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems **36**, 10088–10115 (2023)
6. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
7. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
8. 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. arXiv preprint arXiv:2106.09685 (2021)
9. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
10. 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. Advances in neural information processing systems **36**, 28541–28564 (2023)
11. 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 (ISBI). pp. 1650–1654. IEEE (2021)
12. MedReason 2026 Challenge (MICCAI): Benchmarking medical MLLM reasoning under domain shift. https://medreason26.github.io/ (2026)
13. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine learning for health (ML4H). pp. 353–367. PMLR (2023)
14. Reza, A.M.: Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology **38**(1), 35–44 (2004)
15. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)