MedVLThinker: Simple Baselines for Multimodal Medical Reasoning

Xiaoke Huang xhuan192@ucsc.edu
UC Santa Cruz
Juncheng Wu jwu418@ucsc.edu
UC Santa Cruz
Hui Liu huiliulayne@gmail.com
Amazon Research
Xianfeng Tang tangxianfeng@outlook.com
Amazon Research
Yuyin Zhou yzhou284@ucsc.edu
UC Santa Cruz

Abstract

Large Reasoning Models (LRMs) have introduced a new paradigm in AI by enabling models to "think before responding" via chain-of-thought reasoning. However, the absence of open and reproducible recipes for building reasoning-centric medical LMMs hinders community-wide research, analysis, and comparison. In this paper, we present MedVLThinker, a suite of simple yet strong baselines. Our fully open recipe consists of: (1) systematic data curation for both text-only and image-text medical data, filtered according to varying levels of reasoning difficulty, and (2) two training paradigms: Supervised Fine-Tuning (SFT) on distilled reasoning traces and Reinforcement Learning with Verifiable Rewards (RLVR) based on final answer correctness. Across extensive experiments on the Qwen2.5-VL model family (3B, 7B) and six medical QA benchmarks, we find that RLVR consistently and significantly outperforms SFT. Additionally, under the RLVR framework, a key, counter-intuitive finding is that training on our curated text-only reasoning data provides a more substantial performance boost than training on multimodal image-text data. Our best open 7B model, trained using the RLVR recipe on text-only data, establishes a new state-of-the-art on existing public VQA benchmarks, surpassing all previous open-source medical LMMs. Furthermore, scaling our model to 32B achieves performance on par with the proprietary GPT-4o. We release all curated data, models, and code to provide the community with a strong, open foundation for future research in multimodal medical reasoning.

Keywords: Medical, Reasoning, Multimodal, Vision-Language, Health Care

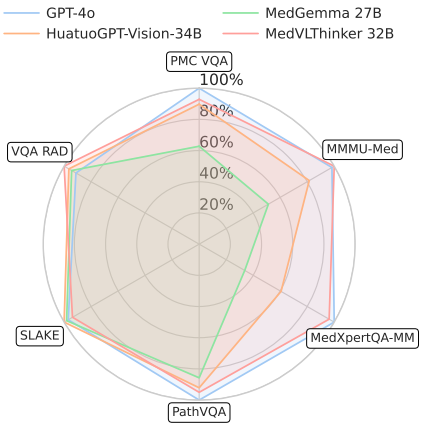


Figure 1: MedVLThinker provides a simple yet strong baseline for multimodal medical reasoning. Notably, MedVLThinker-32B yields performance **on par with the closed-source GPT-4o model**.

Data and Code Availability Our code, models, and data are publicly available at https://github.com/UCSC-VLAA/MedVLThinker.

Institutional Review Board (IRB) Our research does not require IRB approval.

© 2025 X. Huang, J. Wu, H. Liu, X. Tang & Y. Zhou.

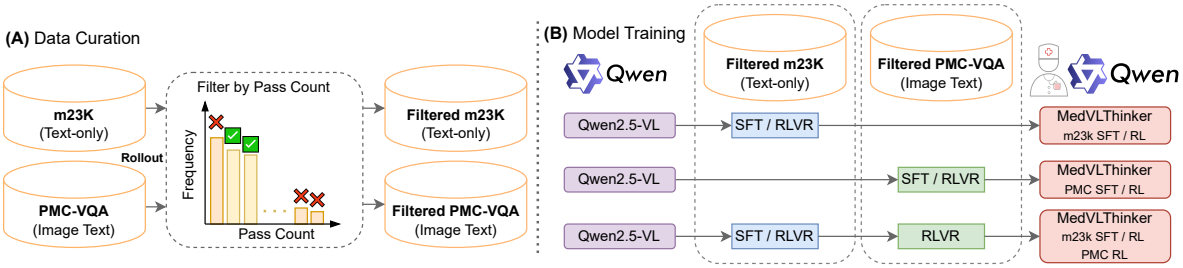


Figure 2: The data filtering and training pipeline. (A) We first filter both text-only m23k dataset and image-text PMC-VQA dataset, by generating multiple answers per question with Qwen2.5-VL-Instruct. Then we filter those questions are answered all wrong or almost correct. (B) Based on the filtered two datasets, we conduct supervised finetuning (SFT), reinforcement learning with verfiable rewaresd (RLVR), and their combination to train a herd of multimodal medical large reasoning models.

1. Introduction

The practice of healthcare increasingly involves processing vast amounts of multimodal medical data (e.g., text, imaging, lab results). Clinicians must integrate information from different sources (clinical notes, radiology images, lab reports) to make diagnoses and treatment decisions. Large Multimodal Models (LMMs) have recently emerged as general-purpose foundation models that can perceive and reason about visual inputs Li et al. (2023); Liu et al. (2023b, 2024); Hurst et al. (2024); Chen et al. (2024b); Xie et al. (2024b). Given that medical data are natively multimodal (e.g., microscopy slides, CT and MRI scans, X-rays), LMMs have a natural appeal for medical AI and have begun to be adopted in modality-rich clinical settings with the potential to improve diagnosis Chen et al. (2024b); Li et al. (2023); Liu et al. (2023a), treatment planning Zhou et al. (2023), and patient monitoring Alshibli et al. (2025).

Parallel to this, Large Reasoning Models (LRMs) extend large language models with a new response paradigm: the model "thinks" through a chain-of-thought before producing a final answer. This allows the model to devote more computation at inference time to reasoning, often improving performance on complex tasks Guo et al. (2025); Guha et al. (2025); Jaech et al. (2024). Early medical adaptations of text-only LRMs have demonstrated strong performance on medical QA tasks Huang et al. (2025); Chen et al. (2024a); Wu et al. (2025); Jiang et al. (2025); Xie et al. (2024a). The ability to generate detailed reasoning steps at test time appears to confer significant gains in accuracy on challenging questions Zuo et al. (2025). However, how to best combine this reasoning paradigm with multimodal understanding remains underexplored. While there have been efforts to build medical multimodal reasoning models, they are often limited in openness—being either entirely closed-source Su et al. (2025); Liu et al. (2025), releasing only model weights without data or training code Sellergren et al. (2025), or, if fully open, are confined to narrow datasets or specific domains (e.g., CT or MRI only) Lai et al. (2025); Pan et al. (2025). As a consequence, the field lacks a comprehensive analysis of how critical factors such as data modality, curation pipelines, and training strategies affect model performance.

In this paper, we provide MedVLThinker, the very first fully open-source recipe for building and evaluating generalized Medical Vision-Language Reasoning Models. Our comprehensive framework provides a complete workflow, from data curation and training pipelines to a standardized evaluation protocol. This enables, for the first time, *a fair and systematic comparison across diverse multimodal medical QA benchmarks*. Figure 2 provides an overview of our approach. We first curate **two types of training data**: a *text-only* QA dataset and an *image-text* (multimodal) QA dataset. Using a general-purpose multimodal LLM (Qwen2.5-VL-Instruct) Bai et al. (2025), we probe each question with multiple trials to estimate its difficulty. Specifically, for each question, we generate multiple candidate answers and count how many times the model answers correctly (the "pass count"). Questions that are consistently answered correctly (too easy) or never answered correctly (too hard) are filtered out, yielding a focused training set of medium-difficulty questions. We then

employ strong teacher models to generate detailed reasoning chains (long *chains-of-thought*, CoTs) for the remaining questions. For text-only questions, we use the *DeepSeek* Guo et al. (2025) model (a powerful text-based LRM) as the CoT teacher, and for image-based questions, we use GPT-4o Hurst et al. (2024) (a vision-enabled GPT-4 variant).

Using these data, we train the base multimodal LLM under **two paradigms**: (1) Supervised fine-tuning (SFT) on the teacher-generated CoT traces, and (2) Reinforcement Learning with Verifiable Rewards (RLVR) on the question-answer pairs (without CoTs). SFT directly teaches the model to reproduce high-quality reasoning and answer traces, whereas RLVR uses only binary rewards from answer correctness to encourage the model's own reasoning. We implement RLVR via Group Relative Policy Optimization (GRPO) Shao et al. (2024), an efficient policy-gradient algorithm that requires no value estimator or critic model. In RLVR training, the model generates multiple reasoning traces for each question; each trace is verified for correct answer format and correctness of the final answer, yielding a +1 or -1 reward. These binary rewards are normalized (whitened) across the batch and fed into the GRPO update step, which applies a PPO-style clipped objective. This process gradually concentrates the model's generation probability mass on verifiably correct reasoning traces while limiting divergence from the original model output distribution.

We conduct extensive experiments on **six multimodal medical QA benchmarks** to investigate the properties of our MedVLThinker. We use the Qwen2.5-VL series as the base models (in 3B, 7B, and 32B parameter sizes). Our evaluations cover both general medical visual QA and modality-specific QA (covering specialties like pathology, radiology, etc.). The results reveal several important, and at times counter-intuitive, findings: First, regarding training paradigms, models trained with RLVR consistently outperform those trained with SFT across both 3B and 7B scales. Second, in terms of data modality, text-only training outperforms image-text training. Notably, SFT on distilled text-only CoT data degrades performance relative to the base model (e.g., MedVLThinker-7B accuracy drops from 53.5% to 43.8%), whereas SFT on image-text data yields performance similar to the untrained base model. In contrast, RLVR on text-only data provides the largest performance boost, improving the 7B model from 53.5% to 54.9%. RLVR on image-text data also improves performance, but to a lesser extent. Moreover, combining text, only and image-text data—either through SFT+RL or sequential RL, does not yield additional gains beyond using text-only data alone. Third, model scale has a clear impact: 7B models consistently outperform their 3B counterparts across all configurations.

Among existing open-source 7B medical LMMs, MedVLThinker-7B (trained with RLVR on text-only data) achieves a new state-of-the-art average accuracy of 54.9% across six benchmarks. To evaluate the effect of model scaling, we further train a 32B variant. As shown in Figure 1, MedVLThinker-32B performs competitively with the proprietary GPT-4o, demonstrating the potential of open models to close the performance gap with commercial systems. To accelerate community-driven development and foster future innovation, we will release our complete research toolkit, including all models, code, and pipelines for data curation, training, and evaluation.

2. Related Works

2.1. Large Reasoning Models and Medical Adaptation

Large Reasoning Models (LRMs) endow large language models with the ability to articulate step-by-step reasoning before finalizing an answer Wei et al. (2022); Guo et al. (2025); Team et al. (2025); Jaech et al. (2024). This test-time "think then answer" approach allows extended reasoning and has yielded impressive gains in domains such as mathematical problem Zeng et al. (2025); Yang et al. (2025); Muennighoff et al. (2025) solving and code generation Jaech et al. (2024); Yang et al. (2024). One way to train LRMs is via Reinforcement Learning with Verifiable Rewards (RLVR), which forgoes supervised chain-of-thought data and instead uses binary feedback on answer correctness Chen et al. (2025); Yu et al. (2025). RLVR eliminates the need to curate lengthy reasoning exemplars; it directly incentivizes correct reasoning by rewarding only the final outcome. In practice, an efficient implementation of RLVR is crucial. Group Relative Policy Optimization (GRPO) Shao et al. (2024) has been adopted for its efficiency, removing the need for a separate value network (critic) during RL updates. An alternative approach is to distill the reasoning traces of stronger models via supervised fine-tuning (SFT). For example, one can use a GPT-4 level model to

generate high-quality explanations (CoTs) for medical questions, and then fine-tune a smaller model on this data Chen et al. (2024a). Recent work shows that fine-tuning medical-focused LRMs (either via SFT on expert traces or via RL on answer rewards) can significantly improve medical question answering performance Wu et al. (2025); Huang et al. (2025); Jiang et al. (2025). Our work extends these ideas to the multimodal realm, examining whether similar reasoning enhancements hold when visual information is involved.

2.2. Multimodal Medical Large Language Models

Given that clinical data often includes images (radiology Lau et al. (2018), pathology Ikezogwo et al. (2023), etc.), there is growing interest in extending LLMs to handle visual inputs for medical applications. Med-Flamingo Moor et al. (2023) was among the first to propose an interleaved vision-language training pipeline for a medical LLM, enabling it to handle image-text pairs in a single prompt. LLaVA-Med Li et al. (2023) introduced a two-stage approach: first, connect a vision encoder with an LLM via a learned projection (connector) and fine-tune on general images; second, fine-tune the combined model on medical image–text instruction data to specialize it. *PMC-VQA* Zhang et al. (2023) is one such large-scale multimodal instruction dataset, constructed from PubMed Central articles (figures and captions) using GPT-3.5 as an annotator. However, the quality of GPT-3.5-generated questions and answers in PMC-VQA is limited by the base model's capacity, and the dataset likely contains noise or insufficiently detailed questions. Other contemporaneous efforts include HuatuoGPT-Vision Chen et al. (2024b), which scales up LLaVA-Med's pipeline by generating a much larger set of QA pairs from a medical corpus and training larger models (up to 34B parameters). There are also modality-specific medical VLMs such as RadFM Wu et al. (2023) and SkinGPT Zhou et al. (2024) that follow similar pipelines but focus on particular domains (e.g., radiology, dermatology) with domain-specific image-text data. In summary, several open-source medical LMMs have been proposed, but integrating an explicit reasoning mechanism (as in LRMs) into these models has not been thoroughly studied prior to our work.

2.3. Concurrent Works

Very recently, a few works have begun exploring the idea of eliciting *medical reasoning* in LLMs. For text-only medical QA, HuatuoGPT-o1 uses a PPO-based RL approach Schulman et al. (2017) with an external reward model to train a medical reasoning LLM Chen et al. (2024a), and MedS3 leverages Process-Reward Models (PRMs) for RL to improve stepwise reasoning Jiang et al. (2025). Another approach, denoted M1 in a recent preprint, distills the reasoning traces of a GPT-4-based model (denoted R1) into a smaller model via SFT Huang et al. (2025). In the multimodal domain, MedVLM-R1 Pan et al. (2025) demonstrates the effectiveness of RLVR on a small scale of multimodal data (fewer than 1K training samples), and Med-R1 Lai et al. (2025) applies a similar RLVR scheme on separate modality-specific datasets. However, these models are trained on limited data and are not generalizable across different types of medical visual questions. GMAI-VL-R1 Su et al. (2025) is a general multimodal medical LLM trained with an RLVR paradigm, but its training data and code are not publicly available. In contrast, our work provides an **open-source** recipe for building multimodal medical reasoning models with both SFT and RL techniques, and we conduct a thorough experimental study across varying model scales (3B, 7B, 32B) and diverse benchmarks. To our knowledge, this is the first work to systematically compare supervised CoT distillation and RLVR for multimodal medical QA, and to benchmark the resulting models against prior open medical LMMs and closed models like GPT-4.

3. Methods

We describe our data curation process and training methodologies for MedVLThinker. Figure 2 illustrates the overall pipeline of data filtering and model training.

Data Curation and Filtering. We gather two datasets for training: a text-only medical QA dataset and a multimodal (image+text) medical QA dataset. For text-only data, we use the `m23k` Huang et al. (2025), which compiles 23,493 multiple-choice medical questions from the training splits of MedQA Jin et al. (2021), MedMCQA Pal et al. (2022), and HeadQA Vilares and Gómez-Rodríguez (2019). Each question in m23k is accompanied by a set of candidate answers, and we have access to high-quality reason-

ing chains (CoTs) for these questions distilled from the DeepSeek-R1 Guo et al. (2025) model. For multimodal data, we adopt PMC-VQA Zhang et al. (2023), a large dataset of 176,948 visual QA pairs derived from biomedical literature figures and captions (covering about 149k images). PMC-VQA was generated using GPT-3.5 and covers a broad range of medical topics, making it a general-purpose multimodal medical QA resource (unlike modality-specific datasets such as PathVQA He et al. (2020), SLAKE Liu et al. (2021), VQA-Rad Lau et al. (2018), which target one type of image).

Not all questions in these datasets are equally useful for training a reasoning model; some are too easy (already trivial for the base model) and some are too hard (unsolvable even with reasoning). Following recent curriculum learning insights Muennighoff et al. (2025), we perform a difficulty-based filtering on both datasets. We prompt three variants of a general multimodal model (Qwen2.5-VL-Instruct with 3B, 7B, 32B parameters) to answer each question 16 times (using nucleus sampling with temperature 1.0). For each question, we record the *pass count*, i.e. the number of trials (out of 16) that produced the correct answer. Figure 3 shows the distribution of pass counts on the text-only m23k and image-based PMC-VQA, for each model size. As model scale increases, more questions achieve high pass counts (e.g. the 32B model answers a larger fraction of questions correctly in a majority of trials). This indicates that the base model's capability improves with scale, which in turn means that a sufficiently large model can solve many of the questions reliably given enough attempts. For the purposes of training data selection, we focus on medium-difficulty questions that are neither always solved nor hopelessly unsolved. Concretely, we use the results of the 3B model to filter the data: any question with pass count $= 0$ (all trials wrong) or ≥ 7 (correct in at least 7 out of 16 trials) is removed. This retains questions that a smaller model finds neither trivial nor impossible, under the assumption that these medium-difficulty questions will benefit most from reasoning training. After filtering, the text-only dataset is reduced to 16,512 questions and the image-text dataset to 115,456 questions. These filtered datasets are used for all subsequent training of 3B, 7B, and 32B models, ensuring a fair comparison across model scales.

3.1. Training Strategies

We train our MedVLThinker models on the filtered data under different strategies, as outlined above. We perform SFT and RLVR on the text-only and image-text datasets *separately* to isolate the effect of each data modality. In addition, we experiment with two combined strategies: (a) SFT on text-only data followed by RL on image-text data (denoted SFT$_{\text{T}}$ + RL$_{\text{I}}$), and (b) RL on text-only data followed by RL on image-text data (RL$_{\text{T}}$ + RL$_{\text{I}}$). Figure 2(B) illustrates the training variants. Below, we describe the two core training paradigms in detail:

Supervised Fine-Tuning (SFT). Supervised fine-tuning forms the foundation of our pipeline. Starting from a general-purpose pretrained multimodal language model (Qwen2.5-VL), we minimize the token-level cross-entropy loss on the curated question-answer pairs (with their reasoning traces). Using teacher-forced learning on the high-quality CoT annotations provides a dense supervision signal, allowing the model to quickly internalize domain-specific medical knowledge, terminology, answer formatting, and the nuanced conventions of clinical explanations. For text-only questions, we use long-form rationales generated by the DeepSeek-R1 model as targets, and for image-based questions, we use GPT-4o-generated rationales. This SFT step teaches the model to emulate the step-by-step reasoning of superior teachers.

Reinforcement Learning with Verifiable Rewards (RLVR). After SFT, we further refine the model using RL on answer correctness as feedback. We adopt Group Relative Policy Optimization (GRPO), a variant of PPO that operates on a group of sampled outputs. For each question, we sample N reasoning trace rollouts from the model (we use $N = 8$ in our experiments). A deterministic verifier then checks each output: if the answer is given in the expected format (e.g., the model produces a chain-of-thought delineated by special tokens and then a final answer choice) and the final answer is correct, a reward $+1$ is assigned; otherwise, a reward -1 is assigned. We normalize (whiten) these binary rewards across the group of outputs to obtain advantage estimates. The GRPO algorithm then updates the model policy using a PPO-style clipped objective, where the usual learned value function is replaced by group-based advantage computation. This yields a KL-regularized, contrastive policy update

that steadily pushes the model to generate more *verifiably correct reasoning traces* (i.e. reasoning that leads to the correct answer) while constraining it to stay close to the behavior policy (to avoid degeneration). Importantly, RLVR does not require explicit CoT annotations, only a reliable way to verify final answer correctness, making it an appealing method to enhance reasoning using the same data. In our setting, all questions are multiple-choice or otherwise have objectively correct answers, so the reward signal is automatically obtained.

4. Experiments

4.1. Implementation Details

We initialize our models from the Qwen2.5-VL checkpoint. For SFT, we fine-tune the model for 3 epochs with a batch size of 32 and learning rate 1×10^{-4}. For RLVR, we train using GRPO for 5 epochs on the text-only data and 1 epoch on the image-text data, with a learning rate of 1×10^{-6}. We set the total batch size to 128 for text-only RL (sufficient to sample 8 rollouts per question) and 256 for image-text RL (since each sample includes image features). For experiments where RL is continued on a second dataset (e.g., applying RL on PMC-VQA after an SFT on m23k), we reduce the batch size (to 64) during the second stage to accommodate the longer sequence lengths (the combined image+CoT+answer sequence can reach ∼2048 tokens). All models are trained on 8×H100 GPUs using mixed precision, except the 32B model, which is trained on 32 GPUs.

4.2. Evaluation

We evaluate our models on a suite of **six** multimodal medical QA benchmarks, which can be divided into two categories: (1) *general-domain medical QA* and (2) *modality-specific QA*. The general-domain evaluations include the test set of PMC-VQA Zhang et al. (2023) (for direct comparison, since our models train on a filtered subset of its training data), the validation set of MMMU-Health Yue et al. (2024) (the health and medicine portion of the MMMU benchmark), and MedXpert-MM Zuo et al. (2025), a challenging benchmark requiring complex reasoning over multimodal inputs. The modality-specific evaluations include PathVQA He et al. (2020) (pathology images), SLAKE Liu et al. (2021) (slit-lamp ophthalmology images) and VQA-Rad Lau et al. (2018) (radiology X-rays). Together, these six datasets cover a broad range of medical visual question answering scenarios, from generic biomedical knowledge to highly specialized imaging tasks.

For each benchmark, we report the accuracy (% of questions answered correctly). Model responses are generated using greedy decoding (temperature 0) to evaluate base capability without sampling variance. We note that even with deterministic decoding, slight nondeterminism in the inference engine (due to floating-point precision) can cause minimal variability; thus, we run each evaluation 3 times and report the average accuracy (the standard deviation was below 0.1 and is provided in the appendix for completeness). In the result tables, we use the notation T to indicate models trained on the text-only (m23k) data and I for models trained on the image-text (PMC-VQA) data. For example, "SFT$_T$" denotes a model fine-tuned on text-only CoT data, and "RL$_T$ +RL$_I$" denotes a model first trained with RL on text-only data then further with RL on image-text data.

4.3. Results

Impact of Training Paradigm (SFT vs. RLVR). Table 1 summarizes the performance of the Qwen2.5-VL 3B and 7B models under various training recipes. We observe that RLVR-trained models consistently outperform SFT-trained models of the same size across all benchmarks. For the 3B base, RLVR on text-only data (RL$_T$) achieves 53.19% average accuracy, versus 32.80% for SFT on text-only (SFT$_T$) (a dramatic drop below the 49.14% base performance). Similarly, the 7B RL$_T$ model reaches 54.88% average, compared to 43.83% for SFT$_T$ (again, SFT underperforms even the 53.50% base model). These results confirm that simply fine-tuning on distilled CoT data does *not* guarantee better performance – in fact, it may overload the model with long, possibly mismatched rationales that hurt its effectiveness on multimodal QA. In contrast, RLVR directly optimizes the model's own reasoning policy and proves markedly more effective at improving accuracy.

Impact of Training Data (Text-only vs. Image-text). The choice of training data modality also has a significant effect. From Table 1, training on the text-only data tends to yield better results than training on the image-text data. For instance, the 7B RL$_T$ (54.88% avg) outperforms RL$_I$ (53.66% avg). However, SFT on the text-only CoT data consistently *harms* performs relative to the base model (43.83%

Table 1: Performance on multimodal medical benchmarks for our baselines. We use greedy decoding to evaluate the ability of the models. (T) means text-only data; (I) means image-text data.

Model	PMC	MMMU	MedX-M	PathVQA	SLAKE	VQA-Rad	Avg.
Qwen2.5-VL-3B-Instruct	44.77	44.12	20.69	61.96	61.30	62.01	49.14
SFT((T) m23k)	28.53	32.55	16.00	42.74	43.91	33.09	32.80
SFT((I) PMC)	54.55	47.84	21.46	52.76	65.79	58.58	50.16
SFT((T) m23k)+RL((I) PMC)	46.32	44.31	20.52	43.85	58.49	50.98	44.08
RL((T) m23k)	47.32	52.16	22.90	62.28	63.38	71.08	53.19
RL((I) PMC)	54.22	48.43	21.51	51.61	75.56	62.38	52.28
RL((T) m23k)+RL((I) PMC)	51.33	48.43	22.60	49.71	66.11	60.17	49.72
Qwen2.5-VL-7B-Instruct	49.30	52.94	18.89	65.39	65.71	68.75	53.50
SFT((T) m23k)	34.58	46.86	16.40	56.35	54.97	53.80	43.83
SFT((I) PMC)	54.67	49.80	21.39	53.02	67.71	57.72	50.72
SFT((T) m23k)+RL((I) PMC)	43.18	47.84	21.84	51.43	60.34	55.15	46.63
RL((T) m23k)	50.67	56.86	24.43	66.83	65.79	64.71	54.88
RL((I) PMC)	55.38	55.29	24.11	57.09	66.59	63.48	53.66
RL((T) m23k)+RL((I) PMC)	56.37	50.98	25.80	48.24	59.13	58.09	49.77

for 7B SFT(T)), whereas SFT on the multimodal data yields a slight improvement over base on some benchmarks (e.g., +1-2% on PathVQA, SLAKE) but overall comparable average (50.72% SFT(I) vs 53.50% base). We hypothesize that the long, text-only rationales distilled from a text-based LRM (DeepSeek) may not align well with the needs of a multimodal model that also has to interpret images. The image-based data, while noisy, at least engages the model's visual processing during training, which might explain why SFT(I) does not drastically degrade performance. Nonetheless, the strongest gains come from RLVR on text-only data, which boosts performance substantially (e.g., +4.05% for 3B, +1.38% for 7B, compared to base). RLVR on the multimodal data also improves over base, but to a lesser degree. These results highlight that in RLVR, high-quality *textual* QA data (with verifiable answers) can be more valuable than larger but noisier image-based data for training the reasoning capability of multimodal models. Improving the quality of multimodal training data remains an important challenge (see Discussion).

Combined Training Strategies. We also evaluated whether combining text-only and image-text training yields further benefits. Two combinations were tried: SFT on text then RL on images (SFT(T) +RL(I)), and RL on text then RL on images (RL(T) +RL(I)). As Table 1 shows, neither strategy provided gains over the single-modality RL training. In fact, for the 7B model, SFT(T) +RL(I) (53.07% avg) was worse than RL(I) alone (53.66%), and RL(T) +RL(I) (49.77%) fell behind RL(T) (54.88%). For the 3B model, similar results are observed. It appears that after a model has been optimized on the text-only data, adding the image-text data (even via RL) can hinder the reasoning capability, resulting in a net drop in performance. We conclude that the best recipe in our study is to apply RLVR directly on a high-quality text-only reasoning dataset. This produces the top results for both 3B and 7B. In most cases, adding an SFT stage or an extra RL stage on image data does not help, and in the worst case, it reduces accuracy.

Effect of Model Scale. Increasing the model size clearly improves performance across the board. The 7B models outperform the 3B models in every corresponding setting (comparing rows in Table 1). For example, the base 7B is 4.36% higher on average than base 3B; the RL(T) 7B is +1.69% higher than its counterpart 3B; and SFT(I) 7B is +0.56% higher than SFT(I) 3B. On certain benchmarks like MedXpert-MM (which is especially challenging and requires complex reasoning), the gap is more pronounced: the best 7B (RL(T)) attains 24.43% versus 22.90% for the best 3B, and 7B SFT(T) achieves 16.40% vs 16.00% for 3B (both quite low). This trend suggests that larger models have more capacity to learn medical knowledge and to benefit from the reasoning train-

Table 2: Performance on multimodal medical benchmarks with other methods. We use greedy decoding to evaluate the ability of the models. Ⓣ means text-only data. Open Weights (OW): only the model parameters are released; Open Recipe (OR): data, code, and training details are released, enabling full reproducibility.

Model	OW	OR	PMC	MMMU	MedX-M	PathVQA	SLAKE	VQA-Rad	Avg.
General LMM									
GPT-4o-mini	✗	✗	51.90	63.53	28.55	63.33	75.24	66.91	58.24
GPT-4o	✗	✗	58.55	68.82	35.95	72.43	76.44	70.22	63.74
Gemme 3 4B	✓	✗	44.42	46.67	21.89	59.24	66.59	56.86	49.28
Gemme 3 27B	✓	✗	52.05	60.78	30.80	65.70	72.60	65.20	57.86
Qwen2.5-VL-3B-Instruct	✓	✗	44.77	44.12	20.69	61.96	61.30	62.01	49.14
Qwen2.5-VL-7B-Instruct	✓	✗	49.30	52.94	18.89	65.39	65.71	68.75	53.50
Qwen2.5-VL-32B-Instruct	✓	✗	53.28	63.92	27.68	67.98	73.24	75.12	60.20
Medical LMM									
MedGemma 4B	✓	✗	42.73	32.55	8.17	59.64	83.49	78.55	50.86
MedGemma 27B	✓	✗	36.75	35.88	12.13	62.09	77.40	72.67	49.49
Llava Med v1.5 Mistral 7B	✓	✓	34.28	31.37	22.56	56.52	62.82	56.74	44.05
HuatuoGPT-Vision-7B	✓	✓	53.39	50.59	22.00	63.53	75.00	63.60	54.69
HuatuoGPT-Vision-34B	✓	✓	52.54	57.06	21.80	66.72	78.85	74.26	58.54
MedVLThinker-3B RL(Ⓣ m23k)	✓	✓	47.32	52.16	22.90	62.28	63.38	71.08	53.19
MedVLThinker-7B RL(Ⓣ m23k)	✓	✓	50.67	56.86	24.43	66.83	65.79	64.71	54.88
MedVLThinker-32B RL(Ⓣ m23k)	✓	✓	54.37	70.00	34.60	68.82	73.96	76.96	63.12

ing. Pushing to even larger scales may continue to yield gains (we test a 32B model below).

Comparison to Previous Models. In Table 2, we compare our MedVLThinker models against prior open-source medical VLMs and against GPT-4-based models. Our 7B RLVR-trained model achieves an average score of 54.88%, which is 3–4% higher than the reported performance of HuatuoGPT-Vision-7B-Qwen2.5 (48.60% avg) and also above LLaVA-Med v1.5 (Mistral-7B). On general-domain benchmarks like MedXpert-MM, our advantage is even larger: MedVLThinker-7B scores 24.43% vs HuatuoGPT-Vision's 22.00%. This demonstrates the benefit of our focused reasoning training. HuatuoGPT-Vision was primarily trained with instruction tuning on multimodal data (and a bit of RLHF), and it underperforms on challenging reasoning questions. We also note that HuatuoGPT-Vision reportedly suffered a large performance drop on generic medical QA after its multimodal fine-tuning (similar to our observation that SFT on image data can hurt general QA). In contrast, our RLVR approach improved performance without such trade-offs. Finally, our MedVLThinker-32B (RL on text-only) reaches 63.12% average accuracy, surpassing the GPT-4o-mini model (58.24%) and essentially matching the full GPT-4o (63.74%) on these benchmarks. This is a notable result: it suggests that with sufficient model size and proper training, open models can approach the performance of proprietary models like GPT-4 on specialized tasks. We emphasize that our entire training pipeline, data, and models are open-source, providing a foundation for the community to build upon.

Qualitative Results. We provide a few anecdotal examples of our model's outputs in Figure 4 to illustrate the reasoning quality of text-only RLVR training. More qualitative results of 3B, 7B, and 32B models can be found in the supplemental materials.

5. Conclusion

In this work, we presented MedVLThinker, a set of baseline multimodal medical reasoning models built by combining large vision-language models with advanced reasoning training paradigms. We carried out a systematic study of supervised CoT fine-tuning versus reinforcement learning (GRPO-based RLVR) for teaching a multimodal model to reason about medical questions. Our experiments show that RLVR is markedly more effective than CoT fine-tuning in

improving model performance, especially when using high-quality text-only medical QA data. We also found that models trained on text-only data generalize better than those trained on image-text data, highlighting a data quality issue in current multimodal corpora. By training models at multiple scales, we demonstrated a clear benefit to larger model size: our 7B MedVLThinker achieves state-of-the-art results among open models on six benchmarks, and a 32B variant reaches parity with a GPT-4-based competitor. Our work provides not only strong baseline models for the community but also insights into training strategies for multimodal reasoning. In future work, we plan to address the limitations identified (data quality, curriculum, broader tasks) and hope that MedVLThinker will inspire further research in reliable and transparent medical AI.

Acknowledgments

This work was partially funded by an unrestricted gift from Google.

References

Ahmad Alshibli, Yakoub Bazi, Mohamad Mahmoud Al Rahhal, and Mansour Zuair. Vision-biollm: Large vision language model for visual dialogue in biomedical imagery. *Biomedical Signal Processing and Control*, 103:107437, 2025.

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

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024a.

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

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

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

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

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

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *arXiv preprint arXiv:2504.00869*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Shuyang Jiang, Yusheng Liao, Zhe Chen, Ya Zhang, Yanfeng Wang, and Yu Wang. Meds3: Towards medical small language models with self-evolved slow thinking. *arXiv preprint arXiv:2501.12051*, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

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.

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

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.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025.

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

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin

Ye, Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

David Vilares and Carlos Gómez-Rodríguez. Head-qa: A healthcare dataset for complex reasoning. *arXiv preprint arXiv:1906.04701*, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*, 2024a.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Guang Yang, Yu Zhou, Xiang Chen, Xiangyu Zhang, Terry Yue Zhuo, and Taolue Chen. Chain-of-thought in neural code generation: From and for lightweight language models. *IEEE Transactions on Software Engineering*, 2024.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

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

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

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.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. *arXiv preprint arXiv:2304.10691*, 2023.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding,

and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

Appendix A. Implementation Details

We implement our training pipeline using two stages: supervised fine-tuning (SFT) followed by reinforcement learning with verifiable rewards (RLVR). For SFT, we employ distributed training using PyTorch's torchrun with FSDP (Fully Sharded Data Parallel) configuration across 8 GPUs per node. The SFT stage uses a learning rate of 1e-5 with cosine scheduler, warmup ratio of 0.05, weight decay of 1e-4, and trains for 5 epochs with a global batch size of 16. We utilize gradient checkpointing and bf16 precision to optimize memory usage. For the RLVR stage, we use the VERL framework with GRPO (Group Relative Policy Optimization) as the advantage estimator. The RL training employs a smaller learning rate of 1e-6, KL divergence regularization with coefficient 0.01, and generates 8 samples per prompt during rollout. We implement a custom reward function that combines format adherence (ensuring responses follow the `<think>...</think> <answer>...</answer>` structure) with accuracy rewards based on exact answer matching. The training uses VLLM for efficient inference during rollout generation with tensor model parallelism across 2 GPUs and 60% GPU memory utilization. Throughout both stages, we use the instruction prompt template below.

Prompt Template

```
You will solve a problem/request.
You should provide your thoughts
within <think> </think> tags before
providing the answer.\nWrite your
final answer within <answer> </answer>
tags.\n{{ Question }}\n{{ Options }}
```

Appendix B. Qualitative Results

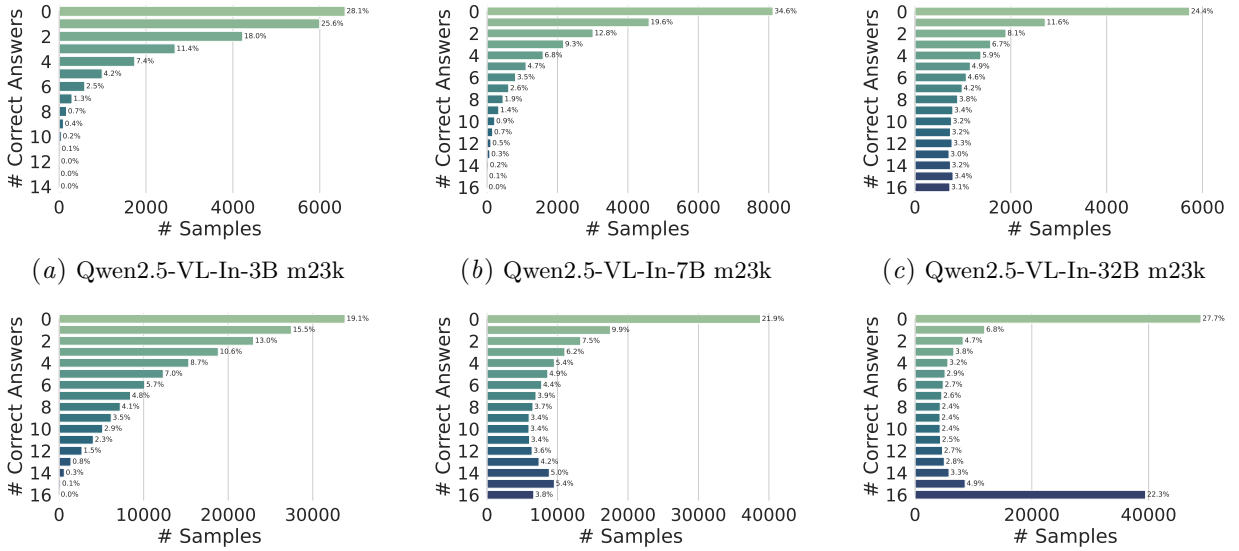
More qualitative results (text-only RLVR 3B, 7B, and 32B) can be found in the **supplemental material**.

Appendix C. Discussions

Quality of Training Data. One striking observation is the performance gap between models trained on text-only data versus image-text data. In our experiments, models trained purely on the PMC-VQA image-text corpus often *lost* exhibit capability relative to their starting point (especially for general QA tasks), whereas models trained on the m23k text-only corpus made clear gains. We suspect the primary cause is the quality of the training data. The PMC-VQA dataset was generated automatically by GPT-3.5 from journal figures and captions. Many of the questions may be simplistic or flawed, and the answers might not always require deep reasoning (or could even be incorrect in some cases). In contrast, the text-only m23k dataset is derived from human-authored exam questions and has higher factual and linguistic quality, supplemented by expert-generated reasoning chains. Unfortunately, besides PMC-VQA, other "general" multimodal medical QA datasets compiled from various sources (e.g., OmniMedQA Hu et al. (2024), GMAI-Bench Ye et al. (2024)) currently do not contain training splits and thus are not directly usable for model training. This highlights the need for better multimodal medical QA data. An encouraging direction is the emergence of high-quality, human-curated medical image report datasets (radiology reports, pathology reports, etc.), which could be leveraged to generate more realistic multimodal QA pairs. In future work, we plan to incorporate such sources to improve the image-text training signal.

Difficulty-Based Filtering. In applying RLVR, it is important to present the model with training examples of appropriate difficulty. We performed a simple filtering by removing questions that a smaller model got either 0/16 or >7/16 correct. This "one-size-fits-all" filter was then applied uniformly for training all model sizes. In reality, different model scales have different ability levels, and an optimal curriculum might adjust the filtering threshold per model (a form of capacity-aware data selection). More adaptive curriculum learning strategies could further improve RLVR training by continuously calibrating question difficulty to the model's growing competence. We leave a full exploration of curriculum learning for medical reasoning to future work.

Effectiveness of RLVR in the Medical Domain. Our results validate that RLVR is a powerful approach for improving reasoning in medical QA, consistent with findings in other domains. With only a few epochs of RL (a relatively small compute budget compared to pretraining), we observed signifi-



(a) Qwen2.5-VL-In-3B m23k (b) Qwen2.5-VL-In-7B m23k (c) Qwen2.5-VL-In-32B m23k

(d) Qwen2.5-VL-In-3B PMC-VQA (e) Qwen2.5-VL-In-7B PMC-VQA (f) Qwen2.5-VL-In-32B PMC-VQA

Figure 3: Probing the questions difficulty with Qwen2.5-VL-Instruct. For each question, we generate 16 answers. Then we draw the pie plots for the pass count. When the scale of the multimodal LLM increased, the number of high pass count questions increased. This indicate the potential of the models, especially for latter RLVR training, which encourage the models improve this possibility to answer questions correctly. The pass count are used for latter data filtering.

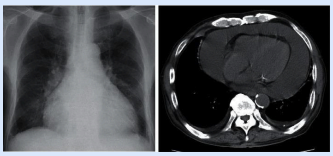
cant gains in the model's ability to arrive at correct answers. This improvement can be interpreted as an increase in *sampling efficiency*: after RLVR, the model is far more likely to produce a correct answer in a single try, whereas the base model might need multiple attempts (as illustrated by Figure 3). Of course, RLVR is not a magic bullet; its success still depends on the diversity and difficulty of the training questions and the reliability of the reward signal. In our case, we used exact-match answer checking, which is straightforward for multiple-choice questions. Extending RLVR to open-ended generation or multi-step clinical reasoning (where reward shaping is harder) is an interesting challenge. We believe scaling up the RLVR approach, with larger verification datasets and more compute, could yield even stronger medical reasoners. This work provides a first step in that direction using openly available tools.

Task Scope and Future Work. So far, we have limited training and evaluation to single-turn question-answering tasks. However, real-world medical scenarios often involve more interactive and diverse tasks: multi-turn dialogues with patients, grounding textual descriptions in images (e.g., locating findings in an image), and reasoning about temporal sequences of images or data. Our current MedVLThinker could potentially be adapted to some of these tasks, but we have not specifically trained or tested it on them. In the future, we aim to extend our framework beyond QA format, incorporating vision-language grounding tasks, dialogue agents for patient interaction, and other medically relevant capabilities. We are hopeful that the combination of explicit reasoning and multimodal understanding, as demonstrated in MedVLThinker, can serve as a foundation for more advanced clinical AI systems.

Appendix D. Limitations

While MedVLThinker aims to advance the state of multimodal medical QA, there are several limitations to acknowledge. First, the quality of the image-text training data (PMC-VQA) is a concern. Because this dataset was synthetically generated by MLLMs, it likely contains noise and biases, which in turn limited the gains from training on it. Future improvements will require either cleaning this data or obtaining higher-quality human-curated multimodal data. Second, our difficulty-based filtering strategy was static

and based on the 3B model's performance. This may not have been optimal for the larger models; a more dynamic or model-specific curriculum could yield better results. Third, our RLVR training was relatively short and task-specific, leveraging only QA-format rewards. As a result, the models are specialized to the QA task format and may not generalize to other interactions (e.g., multi-turn conversations or explanatory responses) without additional fine-tuning. Finally, although we tested a 32B model, even larger models or more diverse pretraining might be needed to capture the full complexity of real-world medical reasoning. We release our models and code in hopes that the community can build on them to address these limitations.

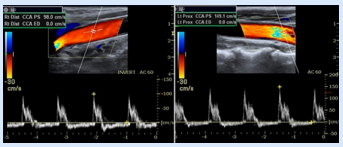


Dataset: PMC VQA

Question: What did the imaging reveal about the patient's pericardial fluid?

Options: A. Very little; B. A moderate amount; C. A massive amount; D. None

<think> To determine the amount of pericardial fluid in the patient, let's analyze the provided images: 1. **Chest X-ray (Left Image)**: The chest X-ray shows a widened mediastinum, which can be indicative of pericardial effusion. However, the X-ray alone is not definitive for quantifying the amount of fluid. 2. **CT Scan (Right Image)**: The CT scan provides a more detailed view of the pericardial space. The image shows a significant amount of fluid surrounding the heart, which is consistent with a pericardial effusion. The fluid appears to be extensive, filling the pericardial space and compressing the heart, which is characteristic of a massive pericardial effusion. Given the visual evidence from both the chest X-ray and the CT scan, the pericardial fluid appears to be extensive and significant. </think> <answer>C. A massive amount</answer>

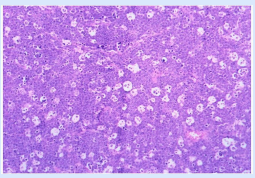


Dataset: MedXpertQA-MM

Question: You are shown images of the right and left distal common carotid arteries, respectively. Which of the following is the MOST likely diagnosis?

Options: A. Intra-aortic balloon pump; B. Carotid artery dissection; C. Mitral insufficiency; D. Subclavian steal; E. Aortic insufficiency

<think> To determine the most likely diagnosis, let's analyze the provided images and the options: 1. **Images Analysis**: The left image shows the right distal common carotid artery (CCA) with a velocity of 98.0 cm/s in the proximal segment (PS) and 0.0 cm/s in the distal segment (ED). This suggests a significant reduction or absence of flow in the distal segment. The right image shows the left distal common carotid artery (CCA) with a velocity of 149.1 cm/s in the proximal segment (PS) and 0.0 cm/s in the distal segment (ED). This also indicates a significant reduction or absence of flow in the distal segment. 2. **Key Observations**: The absence of flow in the distal segments of both carotid arteries is a critical finding. The presence of a "reversed" flow pattern in the distal segments (indicated by the color inversion in the images) suggests retrograde flow, which is often associated with conditions that cause a pressure gradient or flow reversal. 3. **Diagnosis Consideration**: This condition causes a diastolic runoff of blood from the aorta back into the left ventricle, leading to a characteristic diastolic flow reversal in the carotid arteries. The images show a significant reduction or absence of flow in the distal segments, which is consistent with aortic insufficiency. 4. **Conclusion**: The most likely diagnosis is **Aortic insufficiency** because it explains the flow reversal and the significant reduction in flow in the distal segments of both carotid arteries. </think> <answer>E. Aortic insufficiency</answer>



Dataset: PathVQA

Question: Does this image show burkits lymphoma?

Options: A. Intra-aortic balloon pump; B. Carotid artery dissection; C. Mitral insufficiency; D. Subclavian steal; E. Aortic insufficiency

<think> To determine whether the image shows Burkitt's lymphoma, let's analyze the key features: 1. **Histological Appearance**: Burkitt's lymphoma is characterized by a highly mitotic, starry-sky appearance. This is due to the presence of numerous mitotic figures and scattered macrophages (often referred to as "starry sky" because of the scattered appearance of these cells). 2. **Starry-Sky Pattern**: In Burkitt's lymphoma, the lymphoma cells are small to medium-sized, with a high nuclear-to-cytoplasmic ratio, and are often arranged in a diffuse pattern. The presence of tingible body macrophages, which are macrophages that engulf apoptotic bodies, is a hallmark feature that gives the "starry-sky" appearance. 3. **Image Analysis**: The provided image shows a diffuse population of cells with a high mitotic rate. There are scattered cells with a more prominent appearance, which could be tingible body macrophages. The overall pattern is consistent with the "starry-sky" appearance typical of Burkitt's lymphoma. 4. **Conclusion**: Based on the histological features observed in the image, it is consistent with the appearance of Burkitt's lymphoma. </think> <answer>A. Yes</answer>

Figure 4: Case study on multiple medical VQA benchmarks with our 32B text-only RLVR model. Out MedVLThinker demonstrates robust reasoning capability across various imaging modalities.