# Incentivizing Medical Vision Capabilities from Large-Scale Multimodal Pre-training

**Junying Chen**[1,2], **Zhenyang Cai**[1], **Yunjin Yang**[2], **Kunyuan Cai**[1], **Rongsheng Wang**[1], **Xidong Wang**[1], **Geyang Yu**[1], **Xiangyi Feng**[3], **Benyou Wang**[1,2,3,4*]
[1]The Chinese University of Hong Kong, Shenzhen [2]Shenzhen Loop Area Institute
[3]Shenzhen Research Institute of Big Data [4]National Health Data Institute, Shenzhen

## Abstract

Building capable medical vision-language models (VLMs) is challenging due to the scarcity of domain data and the difficulty of eliciting pre-trained medical visual knowledge. We present HuatuoGPT-Vision-2, a medical VLM built on a simple yet effective recipe: large-scale multimodal pre-training followed by reinforcement learning (RL). We first construct PubMedVision-Plus, the largest medical image-text dataset with 60M pairs from public PubMed literature, along with a clean 12M high-quality subset. After pre-training on this data, we find that traditional supervised fine-tuning (SFT) struggles to elicit the pre-trained knowledge. In contrast, by applying RL with verifiable and rubric-based instances synthesized directly from the pre-training data, we successfully incentivize the model to use this pre-trained medical visual knowledge. On PubMed-derived data, HuatuoGPT-Vision-2 outperforms similarly sized VLMs on multiple medical benchmarks and expert evaluations, significantly narrowing the gap between medical VLMs and leading proprietary VLMs. We hope these data, models, and findings can help accelerate open research on capable medical VLMs. Data, code, and models are available at `https://github.com/FreedomIntelligence/HuatuoGPT-Vision-2`.

## 1 Introduction

General-purpose vision-language models (VLMs) have demonstrated impressive capability in visual understanding, multimodal reasoning, and dialogue [1, 13, 29]. Yet, in the medical visual domain, these models still fall short of the specialized knowledge required for clinical interpretation. This has motivated recent medical VLMs, such as HuatuoGPT-Vision [7], HealthGPT [26], Lingshu [60], and MedGemma [45], to adapt general-purpose VLMs to the medical domain. Their common recipe is to first *pre-train* on large-scale medical image-text data to inject domain knowledge, and then apply supervised fine-tuning (SFT) for downstream tasks or conversational use.

However, this paradigm faces two critical bottlenecks. The first is **data scarcity**: to our knowledge, the scale of available medical image-text data is severely limited. Existing public medical image-text data are predominantly derived from PubMed papers. Even the largest prior public collections contain only around 24M biomedical image-caption pairs [31]. Moreover, many training datasets, such as those used in MedGemma, are not publicly available. This scarcity of publicly accessible resources fundamentally constrains the amount of domain knowledge that can be injected during pre-training. The second is **capability activation**: it remains unclear whether knowledge acquired through large-scale pre-training can be effectively elicited by SFT. SFT trains the model to imitate demonstrations under teacher forcing [44], rather than to actively exploit latent capabilities acquired during pre-training. Therefore, eliciting pre-trained medical visual knowledge remains an open research area.

*Correspondence to: Benyou Wang <wangbenyou@cuhk.edu.cn>

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

To address them, we propose a simple but effective recipe: **scale medical multimodal pre-training, and then elicit medical vision capabilities directly with reinforcement learning (RL)**. Concretely, we construct a significantly larger medical multimodal dataset and develop a more effective approach to incentivize the use of pre-trained medical visual knowledge during post-training.

To overcome the data scarcity bottleneck, we construct **PubMedVision-Plus**, the largest medical image-text dataset to date. We collect 60M medical image-caption pairs from public PubMed papers spanning 2013 to 2026, substantially expanding beyond the scale of prior open resources. From this corpus, we further derive **PubMedVision-Plus-Core**, a 12M high-quality subset filtered through a rigorous multi-stage pipeline. We then perform medical multimodal pre-training on a general-purpose VLM with this dataset.

To address the activation bottleneck, we propose to elicit pre-trained medical visual knowledge via reinforcement learning (RL) with training instances synthesized from the pre-training corpus itself. Specifically, we transform image-caption pairs into two complementary forms of RL supervision: multiple-choice questions with verifiable labels, and open-ended questions with rubric-based evaluation criteria, both grounded in the original caption descriptions. The RL stage is explicitly optimized to incentivize the use of medical visual knowledge injected during the pre-training stage.

Building on these components, we present **HuatuoGPT-Vision-2**[2], a medical VLM trained with large-scale medical multimodal pre-training followed directly by reinforcement learning. Using PubMed-derived data as the medical data source, HuatuoGPT-Vision-2 (8B and 32B) improve their Qwen3-VL backbones by an average of nearly 10 points on medical benchmarks, significantly narrowing the gap between medical VLMs and leading proprietary VLMs (e.g., GPT-5.4, Gemini-3-Pro). Expert-rubric and physician-preference evaluations reveal that HuatuoGPT-Vision-2-32B matches GPT-5.4-mini, while HuatuoGPT-Vision-2-8B surpasses some 32B-scale VLMs such as Qwen3-VL-32B-Instruct.

Our main contributions are as follows: **(1) Largest medical image-text dataset.** We construct PubMedVision-Plus with 60M image-caption pairs, and its 12M high-quality subset. **(2) Effective activation through RL.** We reveal that, under matched prompts and data budgets, SFT struggles to elicit pre-trained medical visual knowledge, while RL effectively activates the resulting capabilities. This yields a simple pre-training + RL recipe. This work contributes a data and training paradigm for medical VLMs, with RL serving as the activation mechanism. **(3) RL instance synthesis from pre-training data.** We introduce PubMedVision-RL, where both verifiable and rubric-based RL training instances are synthesized from pre-training captions, creating a unified pipeline from knowledge injection to activation. **(4) Strong empirical performance.** HuatuoGPT-Vision-2-32B outperforms open-source and medical VLMs of similar size, closing the gap with leading proprietary VLMs to the level of GPT-5.4-mini.

## 2 Preliminaries and Challenges for Medical VLMs

### 2.1 Preliminaries

Modern VLMs are typically developed through a two-stage pipeline consisting of pre-training and post-training. We outline and formalize the common training strategies as follows.

**Pre-training.** Given a domain-specific multimodal dataset $\mathcal{D}_{\text{PT}} = \{(I, c)\}$ where $I$ denotes input images (a set of one or more images) and $c$ the corresponding caption, the pre-training objective follows an autoregressive language modeling objective on multimodal data:

$$\mathcal{L}_{\text{PT}}(\theta) = -\mathbb{E}_{(I,c)\sim\mathcal{D}_{\text{PT}}}\left[\sum_{t=1}^{|c|} \log \pi_\theta(c_t \mid I, c_{<t})\right]. \tag{1}$$

This stage injects domain knowledge by learning to predict text context conditioned on images in an unsupervised autoregressive manner.

**Post-training strategy I: supervised fine-tuning (SFT).** SFT trains the model to follow instructions by imitating reference answers. Given a dataset $\mathcal{D}_{\text{SFT}} = \{(I, q, a)\}$ where $q$ is a question about

[2]HuatuoGPT-Vision-2 is the second generation of our HuatuoGPT-Vision series [7]. HuatuoGPT-Vision-7B in Table 2 is the first-generation model, which was trained with the pre-training + SFT recipe.

images $I$ and $a$ is a reference answer, the objective is:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(I,q,a)\sim\mathcal{D}_{\text{SFT}}}\left[\sum_{t=1}^{|a|}\log \pi_\theta(a_t \mid I, q, a_{<t})\right], \tag{2}$$

where $a_t$ denotes the $t$-th token of the reference answer. SFT learns to map queries to responses in a behavior-cloning fashion. Despite its efficiency and widespread use, SFT as an imitation learning method tends to generalize poorly compared to RL-based approaches [10, 17].

**Post-training strategy II: reinforcement learning (RL).** RL optimizes a policy $\pi_\theta$ to maximize a reward signal $r(I, q, o)$ that measures the quality of output $o$ given image $I$ and query $q$. We adopt Group Relative Policy Optimization (GRPO) [47], which for each query samples $G$ outputs $\{o_1, \ldots, o_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$ and optimizes:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_{i,t}A_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)A_i\right) - \beta\mathbb{D}_{KL}(\pi_\theta\|\pi_{\text{ref}})\right], \tag{3}$$

where $\rho_{i,t} = \pi_\theta(o_{i,t} \mid I, q, o_{i,<t})/\pi_{\theta_{\text{old}}}(o_{i,t} \mid I, q, o_{i,<t})$, $A_i$ is the group-normalized advantage shared across tokens in response $o_i$, and $\pi_{\text{ref}}$ is a reference policy. Unlike SFT, RL enables the model to explore and optimize, and is more effective at eliciting the latent potential of the pre-trained model [17].

## 2.2 Challenges

### 2.2.1 Challenge 1: Data Scarcity

Table 1: Comparison of large-scale medical image-text datasets. "Source papers" denotes the number of papers from which image-caption pairs are extracted. PubMedVision-Plus provides substantially larger-scale medical image-text data and stricter quality control.

| Dataset | # Source Papers | # Image-Caption Pairs | Papers Source | Data Cleaning / Filtering |
|---|---|---|---|---|
| ROCO [38] | 1.8M | 0.08M | PubMed | CNN-based radiology filtering |
| MEDICAT [49] | 0.13M | 0.217M | PubMed | Keyword + ResNet filtering |
| PMC-OA [27] | 2.4M | 1.65M | PubMed | Keyword + ResNet filtering |
| PMC-15M [69] | 4.4M | 15M | PubMed | Parser-based extraction |
| BIOMEDICA [31] | 6M | 24M | PubMed | Expert-guided concept filtering |
| **PubMedVision-Plus** | **15.2M** | **60M** | **PubMed** | **Multi-stage LLM/VLM quality filtering** |

Current large-scale medical multimodal datasets are primarily derived from PubMed papers [38, 49, 27, 70, 31]. PubMed data are well suited for medical multimodal pre-training because they are large-scale, diverse, publicly available, and free from private clinical records. However, existing public datasets remain limited in scope. As shown in Table 1, open resources cover only up to 24M image-text pairs and rely on simple filtering methods (e.g., rule-based or CNN-based). This indicates substantial room for improvement in both data scale and quality, which limits the community's research on medical VLMs. Motivated by this, we introduce PubMedVision-Plus to alleviate this limitation by providing 60M medical image-caption pairs extracted from over 15M public PubMed papers, with multi-stage LLM/VLM-based quality filtering.

### 2.2.2 Challenge 2: The Activation Problem

Even when medical multimodal knowledge is injected through large-scale pre-training, effectively eliciting this knowledge for downstream tasks remains an urgent and understudied challenge. The dominant paradigm adopts SFT as the bridge between pre-training and interactive behavior [24, 7]. *However, does vanilla SFT truly elicit pre-trained medical visual knowledge?* We question this through a controlled pilot study.

**Setup.** We use *Qwen3-VL-8B-Instruct* as the base model and pre-train it on a large-scale dataset, **PubMedVision-Plus-Core** (see Section 3.1), at varying data scales: 0%, 20%, ..., 100%. After pre-training, we apply two independent post-training strategies to elicit medical vision capabilities:

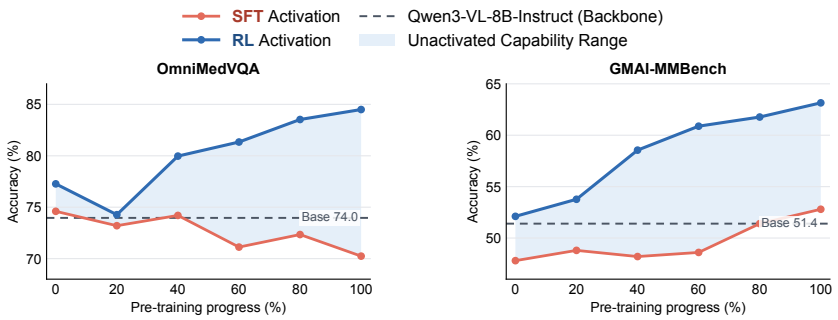



Figure 1: Performance as a function of pre-training progress. **Left**: OmniMedVQA. **Right**: GMAI-MMBench. The RL curve (blue) demonstrates substantially greater gains than the SFT curve (red) as pre-training data scales up, revealing a large region of unactivated capabilities (light blue).

**(1) Activation with RL**: The pre-trained model is trained with GRPO on 70K PubMedVision-RL instances (Section 3.2), using verifiable and rubric-based rewards. PubMedVision-RL is synthesized from pre-training data. **(2) Activation with SFT**: The pre-trained model is trained with SFT on 70K prompt-answer pairs, using the same prompts as PubMedVision-RL and answers generated by the stronger VLM *Qwen3-VL-32B-Instruct*, ensuring a fair comparison. We evaluate on OmniMed-VQA [19] and GMAI-MMBench [9], two broad-coverage medical multimodal benchmarks.

**Key findings.** Results in Figure 1 reveal two main insights. **(1) RL effectively elicits pre-trained medical visual knowledge, while SFT activates substantially less of it.** Previous SFT-based approaches may therefore underuse pre-trained potential, leaving a large space of unactivated capabilities. Notably, Section 4.2 shows that this gap persists even when SFT answers come from frontier teachers such as GPT-5.4 and Gemini-3-Pro. **(2) Larger medical pre-training datasets enable RL to elicit stronger medical visual capabilities.** This highlights the critical importance of scaling medical multimodal pre-training data.

## 3 Method

To tackle the challenges from Section 2.2, we propose a two-stage method. Stage 1 addresses data scarcity by scaling medical multimodal pre-training with substantially larger image-caption datasets and a high-quality core subset. Stage 2 solves the activation problem by converting the pre-training multimodal data into a verifiable format, allowing RL to elicit medical vision capabilities from knowledge injected during pre-training.

### 3.1 Stage 1: Scaling Medical Multimodal Pre-training

The first stage is pre-training on a large-scale medical multimodal dataset to inject domain-specific medical visual knowledge. However, current open resources are limited in both scale and quality. To address this, we first construct a larger-scale dataset, **PubMedVision-Plus**, and apply a multi-stage cleaning strategy to obtain a high-quality subset, **PubMedVision-Plus-Core**, for pre-training. See Figure 2 for an overview.

**PubMedVision-Plus: scaling up medical multimodal data.** To scale up, we first collect publicly available PubMed papers from 2013 to 2026. From these 15.2M papers, we extract every figure image and its associated caption, yielding **60 million image-caption pairs**. This scale is more than twice that of the previously largest open medical image-text corpus. We focus only on captions, as they are self-contained, while inline mentions are loosely associated and may introduce noise. During extraction, we remove artifacts (e.g., "Fig. 3"), correct encoding errors, and discard broken entries. Unlike prior works [7], we keep original expert-authored descriptions without any paraphrasing to ensure scalability and diversity. This yields **PubMedVision-Plus**, the largest medical image-caption corpus to our knowledge.

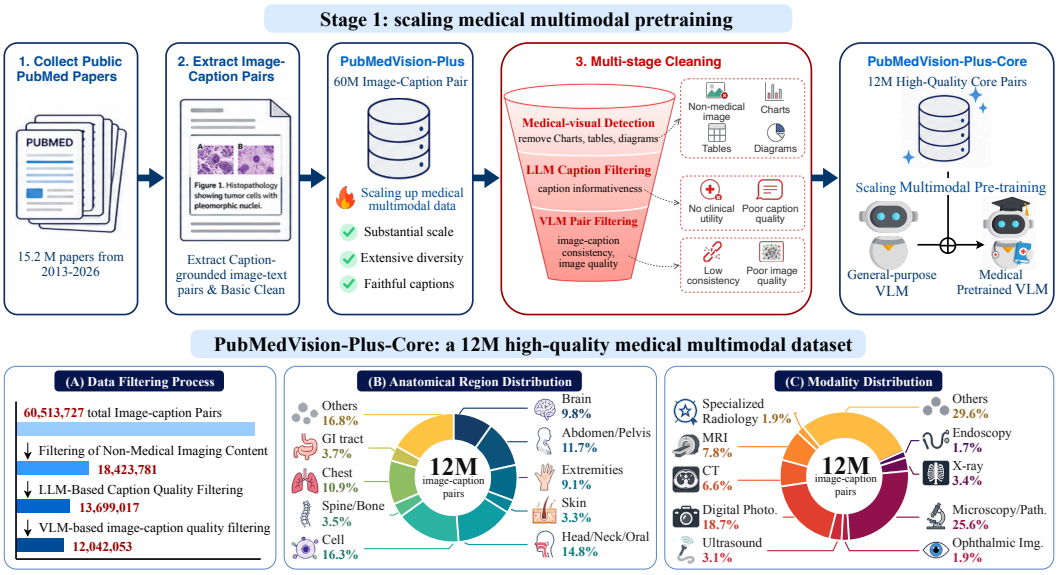



Figure 2: Stage 1 – Scaling medical multimodal pre-training with **PubMedVision-Plus-Core**. We first curate PubMedVision-Plus from public PubMed papers, yielding 60M image-caption pairs. A three-stage cleaning pipeline then yields a high-quality subset, PubMedVision-Plus-Core, which is used to pre-train a VLM for injecting medical visual knowledge.

**PubMedVision-Plus-Core: a 12M high-quality subset.** Despite the unprecedented scale of 60M pairs, not all PubMed image-text data benefits medical visual pre-training. Some figures are charts, diagrams, or schematic illustrations rather than medical visual evidence, and some captions contain little information for learning clinically visual knowledge, while still consuming substantial training compute. Thus, curating a high-quality subset for pre-training is essential. Given the data diversity, we design a three-stage cleaning pipeline using LLM and VLM, as illustrated in Figure 2. **(1) Medical-visual content detection.** We train a caption classifier based on the *all-MiniLM-L12-v2* [42] embedding model on 20K GPT-5.4-labeled samples to filter out non-medical content (e.g., charts, diagrams) based on captions, retaining clinically relevant figures. **(2) LLM-based caption filtering.** We train an LLM classifier based on *Qwen3-4B-Instruct* to remove captions lacking sufficient medical visual information, those of low quality, or missing clinically meaningful findings. **(3) VLM-based pair filtering.** We train VLM classifiers based on *Qwen3-VL-4B-Instruct* for image-caption consistency and image quality assessment, removing low-quality images and mismatched pairs. To enable scalable filtering, we use GPT-5.4 to label 20K seed examples for each filtering task and train a compact classifier with a task-specific binary head on them. See Figure 2 and Appendix E for details. The resulting PubMedVision-Plus-Core contains 12M medically relevant pairs with broad coverage of modalities and anatomical systems (see Figure 2), making it ideal for medical multimodal pre-training. On a physician-annotated gold set of 500 pairs sampled from the raw corpus, the full pipeline retains valid pairs with 93.3% precision and 84.3% recall (Appendix E).

**Pre-training on PubMedVision-Plus-Core.** We conduct continued multimodal pre-training on *PubMedVision-Plus-Core* to inject medical visual knowledge into a general-purpose VLM (e.g., *Qwen3-VL-8B-Instruct*). Given the dataset $\mathcal{D}_{\text{core}} = \{(I, c)\}$, where $I$ is an image and $c$ is its caption text, we train with the autoregressive caption prediction objective defined in Eq. 1. For efficiency, we concatenate multiple pairs into length-4096 sequences using `<end-of-text>` delimiters. We keep standard causal attention without cross-sample masking, rely on the delimiter to mark sample boundaries, and apply the loss only to caption tokens. To mitigate general capability degradation of the VLM, we interleave 1.5B general-domain text tokens from FineWeb [39]. The final pre-training mixture comprises 6.8B tokens from PubMedVision-Plus-Core and 1.5B general tokens, totaling 8.3B tokens.

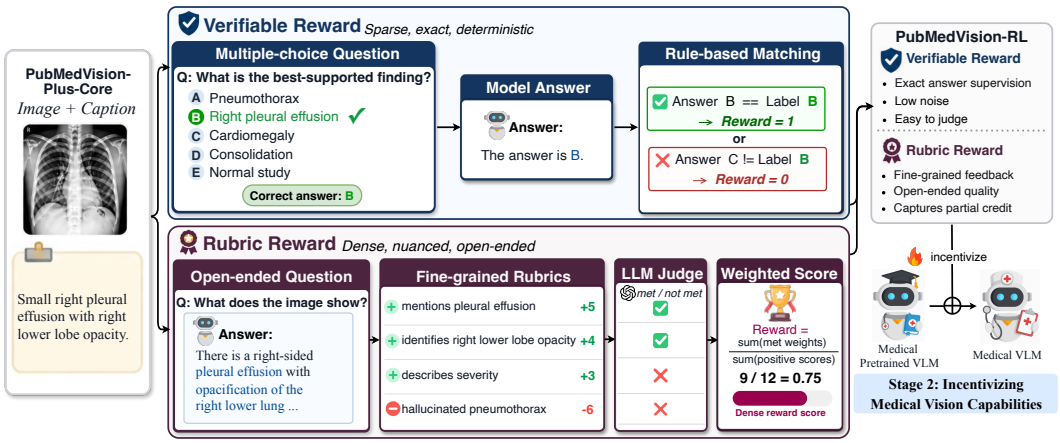



Figure 3: Stage 2 – Incentivizing medical vision capabilities with PubMedVision-RL. Image-caption pairs are converted into two complementary RL instance types, closed-ended verifiable rewards and open-ended rubric rewards. These RL instances incentivize the model to use knowledge learned during first-stage pre-training, thereby eliciting medical vision capabilities. The illustrated examples are simplified for readability, while the actual RL instances contain more challenging questions and more detailed rubrics. We provide actual RL instance samples in Appendix D.

### 3.2 Stage 2: Incentivizing Medical Vision Capabilities

The second stage uses RL to elicit medical vision capabilities from knowledge learned during medical multimodal pre-training. Rather than relying on surface imitation, we construct RL instances from the same pre-training image-caption pairs, where expert captions provide caption-grounded medical evidence. We use two complementary reward types: closed-ended verifiable rewards and open-ended rubric rewards. Figure 3 illustrates this stage.

**Verifiable Reward.** Verifiable rewards provide deterministic, binary feedback on whether the model's answer matches a ground-truth label. As shown in the upper branch of Figure 3, we convert pre-training pairs $\mathcal{D}_{\text{core}} = \{(I, c)\}$ into a verifiable RL set

$$\mathcal{D}_{\text{RL}}^{\text{verify}} = \{(I, q, \mathcal{A}, y)\}, \tag{4}$$

where $q$ is a multiple-choice medical visual question, $\mathcal{A}$ is its answer options, and $y \in \mathcal{A}$ is the ground-truth label. Given a model output $o$, the reward only depends on whether the selected option matches $y$:

$$R_{\text{verify}}(o, y) = \mathbf{1}[f_{\text{choice}}(o) = y], \tag{5}$$

where $f_{\text{choice}}(\cdot)$ extracts the selected option from $o$ by rule-based parsing. **(1) Question generation.** For each image-caption pair, GPT-5.4 generates two distinct multiple-choice questions (4–10 options each) focusing on different medical concepts. A second GPT-5.4 pass validates each question, ensuring the correct answer is uniquely derivable from the caption and distractors are reasonable. **(2) Difficulty filtering.** Using *Qwen3-VL-8B-Instruct* with three independent passes, we discard questions answered correctly in two or more attempts. This removes about 74% of generated questions, yielding 35K verifiable RL instances. Detailed construction procedures and prompts are provided in Appendix I.1. The verifiable reward is simple, rule-checkable, and less prone to reward-model bias, incentivizing the model to perform medical visual reasoning before answering.

**Rubric Reward.** Verifiable RL offers clear feedback, but its binary rewards are too sparse for fine-grained guidance and cannot handle open-ended scenarios. To enable VLMs to produce high-quality, instruction-following responses, we introduce rubric rewards. Using the same pre-training pairs $\mathcal{D}_{\text{core}} = \{(I, c)\}$, we construct a rubric-based RL dataset:

$$\mathcal{D}_{\text{RL}}^{\text{rubric}} = \{(I, q, \mathcal{C})\}, \quad \mathcal{C} = \{(r_i, w_i)\}_{i=1}^{K}, \tag{6}$$

where $q$ is an open-ended medical visual question, $r_i$ a scoring criterion, and $w_i \in \mathbb{R}$ its weight. Each $q$ is associated with multiple rubrics $\{(r_i, w_i)\}_{i=1}^{K}$. Given a model response $o$, a VLM judges each

criterion as met or not: $J_{\text{rubric}}(I, q, o, r_i) \in \{0, 1\}$. Following HealthBench [2], we use GPT-4.1 as the judge. The final reward is

$$R_{\text{rubric}}(I, q, o; \mathcal{C}) = \frac{\sum_{i=1}^{K} w_i \, J_{\text{rubric}}(I, q, o, r_i)}{\sum_{i=1}^{K} \max(w_i, 0)}, \tag{7}$$

where positive weights define the normalization scale, and negative weights decrease the reward when undesirable behaviors are detected. GPT-5.4 converts each image-caption pair into an open-ended question and a set of rubrics. A second GPT-5.4 pass filters out ambiguous or unreasonable rubrics and those that go beyond the caption. We then filter out cases where Qwen3-VL-8B-Instruct already scores $> 0.9$ or GPT-5.4 scores $0$, resulting in 35K rubric-based RL instances. Details are in Appendix I.2. The rubric reward provides dense signals to guide VLMs toward high-quality and accurate medical responses.

**RL Training with PubMedVision-RL.** We synthesize 35K verifiable instances and 35K rubric-based instances, yielding 70K medical multimodal RL examples, termed **PubMedVision-RL**. For RL training, we additionally mix in 15K general-domain rubric instances from RubricHub [25] to preserve general text capability, resulting in 85K total RL training instances. RL training uses GRPO with the DAPO recipe [63]. Both medical instance types are scored by their corresponding verifiable or rubric reward. All medical RL instances are synthesized from PubMedVision-Plus-Core, avoiding external contamination. Since LLM-based rubric scoring is expensive at RL scale, we train a small VLM scorer based on Qwen3-VL-8B-Instruct using 1M GPT-4.1 judgment labels, achieving 91.5% agreement with GPT-4.1 on held-out test data. This scorer is used only to compute rewards during RL training, while all reported rubric evaluations are judged by GPT-4.1. Its human validation and the API costs are reported in Appendix H.2 and Appendix H.6.

# 4 Experiments

## 4.1 Experimental Setup

**Models and training.** We use *Qwen3-VL-8B-Instruct* and *Qwen3-VL-32B-Instruct* as base models for multimodal pre-training and RL. Pre-training uses a batch size of 128, a learning rate of $2 \times 10^{-5}$, and one training epoch, totaling 8.3B tokens. RL uses GRPO with a learning rate of $1 \times 10^{-6}$, a batch size of 128, a rollout of $n = 8$, and a mini-batch size of 16. The SFT baseline uses the same 85K prompts as RL (70K multimodal medical from PubMedVision-RL + 15K text-only general from RubricHub), answered by Qwen3-VL-32B-Instruct. We train SFT for three epochs with a learning rate of $5 \times 10^{-6}$ and a batch size of 128. To test whether teacher quality limits SFT, we additionally train two 8B SFT baselines whose answers come from GPT-5.4 and Gemini-3-Pro. All experiments are conducted on two 8-GPU NVIDIA H200 machines. Note that HuatuoGPT-Vision-2 is intended for academic research only and must not be used for any clinical decision-making.

**Benchmarks.** We evaluate medical multimodal capability on GMAI-MMBench [9], the medical subset of the MMMU-Pro test set [66], MedXpertQA-MM [72], OmniMedVQA [19], PMC-VQA [70], PathVQA [18], SLAKE [28], and VQA-RAD [23]. For PathVQA, SLAKE, and VQA-RAD, we evaluate only the closed-set questions. We report accuracy on these closed-form benchmarks. Details are provided in Appendix C. For open-ended evaluation, we further introduce **MedMM-ExpertRubric**, a medical multimodal benchmark with physician-verified rubrics. It evaluates fine-grained answer quality for open-ended medical questions. Details are provided in Appendix F.

**Baselines.** We compare three groups. **(1) Proprietary VLMs:** Gemini [16, 15] and GPT [34, 35, 36]. All GPT-5.4 models, including the SFT teacher, use high reasoning effort, marked as "(High)" in Table 2. **(2) Open-source VLMs:** Gemma-4 [14], InternVL3.5 [54], Qwen3.5 [41], and Qwen3-VL [4]. **(3) Medical VLMs:** HuatuoGPT-Vision [7], Lingshu [60], and MedGemma [46] (marked with † in Table 2).

## 4.2 Experimental Results

**Baseline comparison.** Table 2 compares HuatuoGPT-Vision-2 with representative proprietary, open-source, and medical-domain VLMs. Among open-source VLMs, HuatuoGPT-Vision-2-32B

Table 2: Main results on medical VLM benchmarks. All scores are accuracies (%). GMAI, MPro, MedXQA, OMVQA, PMC, Path, and RAD denote GMAI-MMBench, MMMU-Pro-Med, MedXpertQA-MM, OmniMedVQA, PMC-VQA, PathVQA, and VQA-RAD, respectively. † marks medical VLMs. In the recipe comparison, **SFT** and **RL** mark the post-training method, the subscript of SFT denotes the teacher that generates the SFT answers (Qwen: Qwen3-VL-32B-Instruct, Gemini: Gemini-3-Pro, GPT: GPT-5.4), and shaded rows are our HuatuoGPT-Vision-2 models. Parentheses in Avg. show changes over the base model.

| Model | GMAI | MPro | MedXQA | OMVQA | PMC | Path | SLAKE | RAD | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| ***Proprietary VLMs*** | | | | | | | | | |
| GPT-4.1 [34] | 58.4 | 51.6 | 45.4 | 71.6 | 62.8 | 68.6 | 75.8 | 70.3 | 63.1 |
| GPT-5.4-nano (High) [36] | 40.4 | 57.9 | 46.8 | 56.2 | 52.0 | 64.4 | 67.2 | 71.0 | 57.0 |
| GPT-5.4-mini (High) [36] | 56.7 | 74.0 | 69.4 | 76.2 | 66.0 | 66.8 | 80.0 | 79.4 | 71.1 |
| GPT-5.4 (High) [35] | 64.6 | 80.9 | 78.2 | 75.6 | 70.3 | 68.7 | 83.4 | 78.4 | 75.0 |
| Gemini-3-Flash [15] | 65.5 | 82.7 | 79.3 | **86.4** | **73.8** | 77.0 | 82.6 | 80.1 | 78.4 |
| Gemini-3-Pro [16] | **76.2** | **85.8** | **81.9** | 80.8 | 72.4 | **77.4** | **88.9** | **85.7** | **81.1** |
| ***Open-source VLMs*** | | | | | | | | | |
| † HuatuoGPT-Vision-7B [7] | 42.0 | 29.5 | 21.0 | 73.6 | 61.0 | 62.8 | 71.8 | 69.7 | 53.9 |
| † MedGemma-27B-IT [46] | 37.2 | 34.7 | 33.6 | 61.0 | 43.6 | 56.2 | 68.4 | 71.0 | 50.7 |
| InternVL3.5-8B [54] | 52.4 | 46.3 | 22.6 | 77.4 | 54.2 | 64.6 | 75.0 | 70.3 | 57.9 |
| Qwen3.5-27B [41] | 54.4 | 59.2 | 43.0 | 70.6 | 56.2 | 67.8 | 80.6 | 74.9 | 63.3 |
| Gemma-4-31B-IT [14] | 58.4 | 51.9 | **49.4** | 77.0 | 64.8 | 74.4 | 81.8 | 81.0 | 67.3 |
| † Lingshu-32B [60] | 55.4 | 44.6 | 33.1 | 79.2 | 66.0 | **81.8** | 84.0 | 82.3 | 65.8 |
| *Our experiments on 8B scale* | | | | | | | | | |
| Qwen3-VL-8B-Instruct [4] | 51.4 | 52.6 | 27.9 | 74.0 | 53.1 | 65.2 | 73.4 | 71.0 | 58.6 |
| w/ **RL** | 52.1 | 50.5 | 30.8 | 77.3 | 55.1 | 67.2 | 76.0 | 74.8 | 60.5 (+1.9) |
| w/ Pre-training + $\textbf{SFT}_{\text{Qwen}}$ | 52.8 | 46.3 | 26.4 | 70.2 | 60.6 | 70.2 | 77.4 | 73.6 | 59.7 (+1.1) |
| w/ Pre-training + $\textbf{SFT}_{\text{Gemini}}$ | 51.6 | 46.8 | 27.6 | 71.4 | 60.9 | 71.8 | 77.8 | 72.0 | 60.0 (+1.4) |
| w/ Pre-training + $\textbf{SFT}_{\text{GPT}}$ | 55.6 | 45.6 | 27.8 | 73.1 | 61.3 | 69.2 | 79.4 | 74.7 | 60.8 (+2.2) |
| **w/ Pre-training + RL** *(† HuatuoGPT-Vision-2-8B)* | 63.2 | 56.1 | 34.2 | 84.5 | 69.2 | 74.0 | 84.6 | 82.3 | 68.5 (+9.9) |
| *Our experiments on 32B scale* | | | | | | | | | |
| Qwen3-VL-32B-Instruct [4] | 57.0 | 55.7 | 36.5 | 80.8 | 59.8 | 66.6 | 78.0 | 76.8 | 63.9 |
| w/ **RL** | 61.8 | 63.5 | 43.4 | 83.9 | 63.0 | 72.0 | 82.6 | 78.3 | 68.6 (+4.7) |
| w/ Pre-training + $\textbf{SFT}_{\text{Qwen}}$ | 55.2 | 52.2 | 30.1 | 73.1 | 61.0 | 65.6 | 81.9 | 78.4 | 62.2 (-1.7) |
| **w/ Pre-training + RL** *(† HuatuoGPT-Vision-2-32B)* | **65.8** | **69.8** | 48.9 | **88.4** | **71.2** | 77.8 | **86.6** | **83.5** | **74.0** (+10.1) |

achieves an average score of 74.0, clearly outperforming similarly sized general VLMs such as Gemma-4-31B-IT and medical VLMs such as Lingshu-32B. The 8B model also reaches 68.5, surpassing several larger open models, which highlights the value of domain adaptation. Results also show that proprietary VLMs, such as Gemini-3-Pro and GPT-5.4, are significantly stronger. HuatuoGPT-Vision-2-32B further narrows the gap to proprietary VLMs and exceeds GPT-5.4-mini on average. With transparent training data and a simple two-stage recipe, HuatuoGPT-Vision-2 provides a reproducible path for open medical VLM research.

**Ablation study.** Table 2 also compares different training recipes. Applying RL directly to a general VLM gives limited gains, with only +1.9 points at the 8B scale. After medical multimodal pre-training, RL expands the gain to about 10 points at both 8B and 32B, showing the importance of injecting domain knowledge before RL. In contrast, SFT gives much smaller gains and even degrades the 32B model, suggesting that it does not reliably turn pre-trained medical visual knowledge into downstream capabilities. The results reveal that under the same prompts and data volume, RL is more effective at eliciting capabilities in our evaluated setting. Stronger teachers help SFT only modestly. With answers from GPT-5.4 or Gemini-3-Pro, Pre-training + SFT reaches 60.8 and 60.0, still far below Pre-training + RL (68.5).

**General capability preservation.** Table 3 evaluates general capability retention after medical adaptation on MMLU-Pro [55], IFEval [71], GPQA [43], MMStar [8], RealWorldQA [57], and MathVision [52]. HuatuoGPT-Vision-2-8B slightly improves the overall average, while HuatuoGPT-Vision-2-32B remains nearly unchanged. These results indicate that the medical gains in Table 2 do not come from broad forgetting of general text or vision capability. This also suggests that continued

Table 3: General-domain capability evaluation. HuatuoGPT-Vision-2 (shaded) is compared with its Qwen3-VL backbone at each scale. RWQA and MathV denote RealWorldQA and MathVision. The better score in each pair is in bold, and parentheses in Avg. show changes over the backbone.

| Model | Text-only | | | General Vision | | | Avg. |
|---|---|---|---|---|---|---|---|
| | MMLU-Pro | IFEval | GPQA | MMStar | RWQA | MathV | |
| Qwen3-VL-8B-Instruct | 71.12 | **81.70** | 44.95 | **68.12** | 68.76 | 52.62 | 64.55 |
| † HuatuoGPT-Vision-2-8B | **71.38** | 81.10 | **46.01** | 66.00 | **71.90** | **53.20** | **64.93** (+0.38) |
| Qwen3-VL-32B-Instruct | 79.80 | 86.04 | **51.01** | **74.88** | 75.82 | **61.78** | **71.56** |
| † HuatuoGPT-Vision-2-32B | **82.25** | **87.64** | 49.61 | 73.25 | **75.95** | 59.97 | 71.44 (-0.12) |

large-scale pre-training does not severely harm the original backbone, and the latent capabilities can still be recovered.

**MedMM-ExpertRubric evaluation.** We evaluate open-ended medical multimodal answer quality using MedMM-ExpertRubric, which contains 500 held-out multimodal medical questions, each with physician-verified rubrics. For evaluation, GPT-4.1 directly scores responses against physician-verified rubrics, measuring coverage of physician-verified key points. Figure 4 shows that HuatuoGPT-Vision-2 improves over the Qwen3-VL backbones, and HuatuoGPT-Vision-2-32B exceeds GPT-5.4-mini, indicating stronger alignment with expert rubrics. Human validation of the GPT-4.1 judge is provided in Appendix H.2. This benchmark reorders the proprietary model ranking in Table 2: GPT-5.4 now surpasses Gemini-3-Flash but trails Gemini-3-Pro. See Appendix F for detailed procedures.

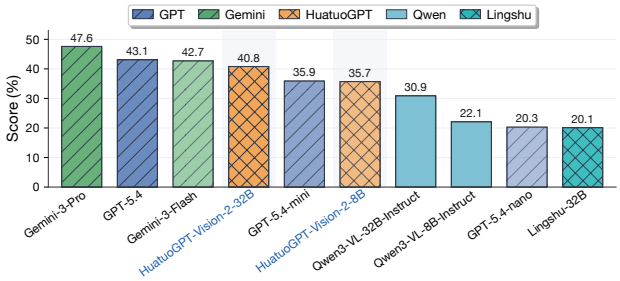


Figure 4: MedMM-ExpertRubric evaluation with expert-validated rubrics.

**Physician preference evaluation.** We conduct a physician preference study on 100 open-ended questions from MedMM-ExpertRubric. Two attending physicians compare anonymized responses from two VLMs in randomized order, choosing the preferred answer or marking a tie. Appendix G provides detailed annotation guidelines, inter-rater agreement analysis, and confidence intervals. Figure 5 shows that HuatuoGPT-Vision-2-32B is clearly preferred over Qwen3-VL-32B-Instruct and Lingshu-32B, and is comparable to GPT-5.4-mini with a slightly higher win rate, while still trailing GPT-5.4. This further demonstrates that medical training significantly improves physician preference. These physician preferences align with MedMM-ExpertRubric, supporting its reliability and potential to reduce evaluation costs for the community.

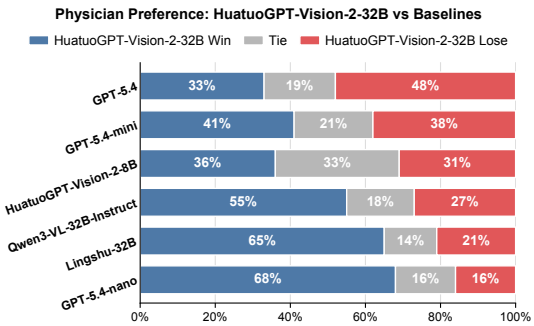


Figure 5: Physician pairwise preference study.

**Reward complementarity.** Figure 6 compares reward compositions on the 8B medical multimodal pre-trained VLM. Verifiable rewards excel in closed-ended benchmarks but struggle in open-ended scenarios. Rubric rewards better support open-ended response quality but are less effective on closed-ended benchmarks. Combining both yields the best RL performance, indicating that the two rewards provide complementary benefits in our experiments.

**Pre-training scale and RL reward gains.** Figure 7 analyzes RL reward gains under different pre-training data scales. Without continued pre-training, RL brings only marginal reward improvements.

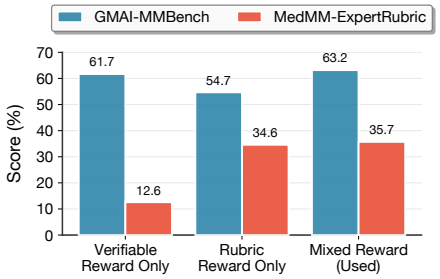



Figure 6: Reward-type ablation on the 8B medical multimodal pre-trained VLM.

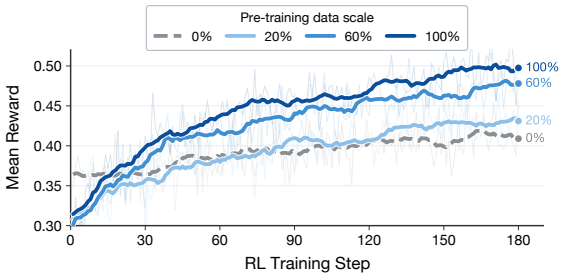



Figure 7: RL reward gains vs. pre-training data scale on Qwen3-VL-8B-Instruct.

With more medical pre-training data, the reward gains show a stronger upward trend. This supports our central view: pre-training builds the medical visual knowledge base, and RL elicits downstream capabilities from it. This also reveals that domain pre-training is essential when the backbone model is weak in domain capabilities.

Table 4: Zero-shot chest X-ray report generation at 8B. Cells show CheXbert micro-F1 / RadGraph F1.

| Dataset | Base | Pre-training + SFT | Pre-training + RL (Ours) |
|---|---|---|---|
| MIMIC-CXR | 35.0 / 12.9 | 39.9 / 13.6 | **44.1 / 16.2** |
| IU-Xray | 31.7 / 14.1 | 22.9 / 15.5 | **38.6 / 21.5** |
| CheXpert Plus | 35.9 / 13.0 | 38.7 / 13.8 | **44.4 / 16.1** |

**Generalization to clinical reports.** All medical training data come from PubMed figures. To test whether the elicited capabilities generalize to real clinical data and a new task format, we evaluate zero-shot report generation on MIMIC-CXR [22], IU-Xray [11], and CheXpert Plus [6]. HuatuoGPT-Vision-2-8B improves over its Qwen3-VL-8B backbone on all three datasets and both metrics, while Pre-training + SFT gives mixed results.

**Additional analyses.** Appendix H further reports bootstrap confidence intervals, human validation of the rubric judges, RL data scaling, failure analysis, and API costs.

## 5 Conclusion

Building strong medical VLMs requires both medical visual knowledge and an effective way to elicit it. We proposed a simple recipe: scale medical multimodal pre-training, then apply RL to turn the acquired knowledge into medical vision capabilities. This recipe is supported by **PubMedVision-Plus**, a 60M-pair medical image-text corpus, **PubMedVision-RL**, a caption-grounded RL dataset, and the resulting **HuatuoGPT-Vision-2** models. Experiments show that HuatuoGPT-Vision-2 achieves the strongest performance among similarly sized medical VLMs, and improves open-ended answer quality under expert-rubric and physician-preference evaluation. The results reveal that: 1. Injecting foundational domain capabilities is important for RL enhancement; 2. Under matched prompts and data budgets, RL elicits pre-trained knowledge more effectively than SFT. We hope these data, models, and findings can help accelerate open research on capable and transparent medical VLMs.

## Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China (Grant No. 2024YFE0203000); the National Natural Science Foundation of China (Grant No. 72495131); the Shenzhen Science and Technology Program—General Program of Basic Research (Grant No. JCYJ20250604141206009); the Shenzhen Medical Research Fund (Grant Nos. B2503005 and C2406002); the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (Grant 2023B1212010001 and ZSZDSYS20251011093302004); and the 1+1+1 CUHK–CUHK(SZ)–GDSTC Joint Collaboration Fund (Grant No. 2025A0505000078). The authors declare no competing interests.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, et al. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

[3] Fan Bai, Yuxin Du, Tiejun Huang, et al. M3D: Advancing 3D medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[4] Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Aurelia Bustos, Antonio Pertusa, José-María Salinas, and María de la Iglesia-Vayá. PadChest: A large chest X-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[6] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

[7] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370. Association for Computational Linguistics, 2024.

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pages 27056–27087. Curran Associates, Inc., 2024.

[9] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Yu Qiao, and Junjun He. GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. In *Advances in Neural Information Processing Systems*, volume 37, pages 94327–94427. Curran Associates, Inc., 2024.

[10] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10818–10838. PMLR, 2025.

[11] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23 (2):304–310, 2016.

[12] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. In *Advances in Neural Information Processing Systems*, volume 38, pages 123817–123846. Curran Associates, Inc., 2025.

[13] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[14] Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

[15] Google. Gemini 3 Flash: Frontier intelligence built for speed, 2025. `https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/`.

[16] Google. Start building with Gemini 3, 2025. `https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-developers/`.

[17] Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[18] Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards Visual Question Answering on Pathology Images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718. Association for Computational Linguistics, 2021.

[19] 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 (CVPR)*, pages 22170–22183, 2024.

[20] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597. AAAI Press, 2019.

[21] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[22] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019.

[23] 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):180251, 2018.

[24] 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. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc., 2023.

[25] Sunzhu Li, Jiale Zhao, Miteto Wei, et al. RubricHub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation. *arXiv preprint arXiv:2601.08430*, 2026.

[26] Tianwei Lin, Wenqiao Zhang, Sijing Li, et al. HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 37975–37995. PMLR, 2025.

[27] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-CLIP: Contrastive language-image pre-training using biomedical documents. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 525–536, 2023.

[28] 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.

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[30] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual Reinforcement Fine-Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2034–2044, 2025.

[31] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J. Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, Yuhui Zhang, Collin Chiu, Xiaohan Wang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19724–19735, 2025.

[32] Fanqing Meng, Lingxiao Du, Zongkai Liu, et al. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[33] 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 *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 2023.

[34] OpenAI. Introducing GPT-4.1 in the API, 2025. `https://openai.com/index/gpt-4-1/`.

[35] OpenAI. Introducing GPT-5.4, 2026. `https://openai.com/index/introducing-gpt-5-4/`.

[36] OpenAI. Introducing GPT-5.4 mini and nano, 2026. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[38] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M. Friedrich. Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, volume 11043 of *Lecture Notes in Computer Science*, pages 180–189. Springer International Publishing, 2018.

[39] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc., 2024.

[40] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, et al. LMM-R1: Empowering 3B LMMs with strong reasoning abilities through two-stage rule-based RL. *arXiv preprint arXiv:2503.07536*, 2025.

[41] Qwen Team. Qwen3.5-27B model card, 2026. https://huggingface.co/Qwen/Qwen3.5-27B. Accessed: 2026-11-06.

[42] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019.

[43] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

[44] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635. PMLR, 2011.

[45] Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

[46] Andrew Sellergren, Sahar Kazemzadeh, Fereshteh Mahvar, et al. An open vision-language model for diverse medical applications. *Nature Medicine*, 32(10):3695–3704, 2026.

[47] Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[48] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519. Association for Computational Linguistics, 2020.

[49] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120. Association for Computational Linguistics, 2020.

[50] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning of Vision Language Models. In *Advances in Neural Information Processing Systems*, volume 38, pages 5772–5822. Curran Associates, Inc., 2025.

[51] Tao Tu, Shekoofeh Azizi, Danny Driess, et al. Towards Generalist Biomedical AI. *NEJM AI*, 1(3): AIoa2300138, 2024.

[52] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, pages 95095–95169. Curran Associates, Inc., 2024.

[53] Weiyun Wang, Zhe Chen, Wenhai Wang, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.

[54] Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[55] Yubo Wang, Xueguang Ma, Ge Zhang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc., 2024.

[56] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 16(1):7866, 2025.

[57] xAI. RealWorldQA, 2024. `https://huggingface.co/datasets/xai-org/RealworldQA`. Accessed: 2026-11-07.

[58] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. MedTrinity-25M: A Large-scale Multimodal Dataset with Multigranular Annotations for Medicine. In *The Thirteenth International Conference on Learning Representations*, 2025.

[59] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2087–2098, 2025.

[60] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Can Hu, Siwei Pan, Jie Tan, Tingyang Xu, Deli Zhao, Hao Zhang, and Yu Rong. Lingshu: Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. Early access.

[61] Lin Yang, Shawn Xu, Andrew Sellergren, et al. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[62] Yi Yang, Xiaoxuan He, Hongkun Pan, et al. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2376–2385, 2025.

[63] Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244. Curran Associates, Inc., 2025.

[64] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13807–13816, 2024.

[65] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19985–19995, 2025.

[66] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186. Association for Computational Linguistics, 2025.

[67] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1859–1869, 2025.

[68] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.

[69] Sheng Zhang, Yanbo Xu, Naoto Usuyama, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2025.

[70] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1): 277, 2024.

[71] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. arXiv:2311.07911, 2023.

[72] Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80961–80990. PMLR, 2025.

# A Discussions

## A.1 Limitations

Our work has several limitations. First, PubMedVision-Plus is built from English biomedical literature. Although PubMed provides broad coverage and avoids private clinical data, the resulting models still need validation in multilingual and region-specific clinical settings. PubMed figures and captions also differ from real-world clinical workflows, where images are often paired with longitudinal history, laboratory results, reports, and treatment context. Moreover, PubMed favors abnormal and unusual findings, and our filtering pipeline keeps 84.3% of valid pairs on the physician-annotated gold set. Both factors can bias the corpus against normal cases and some modalities.

Second, MedMM-ExpertRubric uses physician-verified rubrics but still relies on GPT-4.1 for scalable scoring, and both GPT-4.1 and our 8B scorer show systematic errors on negated, conjunctive, and paraphrased criteria (Appendix H.2). Physician preference evaluation supports the reliability of this protocol, but our study covers only 100 questions with two physicians, and their agreement is moderate. Larger physician studies remain important for safety-sensitive medical evaluation. Our clinical transfer evaluation is also limited to zero-shot chest X-ray report generation. In addition, our SFT comparison covers three teacher models under matched prompts and data budgets, and other SFT designs, such as larger SFT sets or reasoning-trace distillation, may narrow the gap.

Finally, HuatuoGPT-Vision-2 is intended only for academic research as a general medical multimodal model. It is not intended for diagnosis, treatment, triage, or any other medical decision, and its outputs must not replace qualified clinical judgment.

## A.2 Future Directions

Several directions are worth exploring. A first direction is to extend PubMedVision-Plus beyond English literature and evaluate cross-lingual medical visual understanding. A second direction is to connect literature-derived supervision with richer clinical contexts, including multi-image studies, longitudinal records, and structured reports. A third direction is to scale physician feedback beyond evaluation, for example by training reward models or preference models from larger expert judgments. Finally, the pre-training + RL recipe can be extended to more specialized settings such as radiology, pathology, ophthalmology, ultrasound, and 3D medical imaging. These extensions would test whether the same knowledge-injection and capability-activation pattern holds across clinical modalities and workflows.

## A.3 Impact Statement

This work aims to advance medical AI by providing open datasets and models that lower the barrier to research in medical VLMs. All medical training data is sourced from publicly available PubMed literature, avoiding privacy concerns associated with clinical data. However, medical VLMs should not be used for direct clinical diagnosis without appropriate validation and oversight. We encourage responsible use and emphasize that model outputs are intended to assist, not replace, professional medical judgment.

# B Related Work

**Large-scale medical multimodal datasets.** Large-scale medical multimodal data are the foundation of medical VLMs, as reflected by clinical image-report corpora such as MIMIC-CXR [22], CheXpert [20], and PadChest [5], though these resources mainly focus on chest radiography. Literature-derived datasets broaden modality coverage: ROCO [38] and MedICaT [49] collect biomedical figures with captions or textual references, while PMC-15M/BiomedCLIP [69] scales scientific image-text pre-training to millions of pairs. More recent resources further transform PubMed or heterogeneous medical images into VQA or instruction data, including PMC-VQA [70], PubMedVision [7], MedTrinity-25M [58], and BIOMEDICA [31]. PubMedVision-Plus complements this line by scaling directly to 60M expert-authored PubMed image-caption pairs and curating a 12M high-quality subset with multi-stage LLM/VLM filtering.

**Medical VLMs.** Medical VLMs typically adapt general-purpose VLMs through domain data construction, continued pre-training, and supervised instruction tuning, as seen in early biomedical VQA systems such as LLaVA-Med [24] and MedVInT [70]. Med-Flamingo [33], RadFM [56], and BiomedGPT [68] further explore broader medical or radiology foundation models. Recent open models expand the modality and task space, including 3D medical imaging with M3D-LaMed [3], heterogeneous medical understanding and generation with HealthGPT [26], and generalist medical VLMs such as HuatuoGPT-Vision [7], Lingshu [60], and MedGemma [45]. Proprietary systems such as Med-PaLM M [51] and Med-Gemini [61] further show the potential of generalist biomedical AI at larger scale. Our work focuses on the next bottleneck: after medical knowledge is injected by pre-training, SFT may not elicit it effectively, while caption-grounded RL provides stronger post-training supervision.

**Reinforcement learning for VLMs.** RL has been widely used for LLM alignment, from RLHF for instruction following [37] to verifiable-reward RL for math and code reasoning [47, 17]. In VLMs, RLHF-V [64] and RLAIF-V [65] focus on alignment, trustworthiness, and hallucination reduction through human or AI feedback. Another line improves multimodal reasoning with preference optimization and structured reasoning supervision, such as MPO [53] and LLaVA-CoT [59]. Recent GRPO-style or rule-based methods further extend RL to visual perception and multimodal reasoning, including Visual-RFT [30], Reason-RFT [50], R1-VL [67], LMM-R1 [40], MM-Eureka [32], R1-Onevision [62], and OpenVLThinker [12]. These methods mainly activate reasoning patterns in models that already contain the required knowledge, often with SFT warm-up, distilled reasoning traces, or task-specific verifiers [59, 50, 62, 12]. Our work instead reveals a complementary insight: RL alone is limited without sufficient medical visual knowledge. This suggests that scaling medical visual knowledge first, followed by visual RL, is a promising path for building stronger medical VLMs.

# C Evaluation Benchmark Details

Table 5 summarizes the medical multimodal benchmarks used in our main evaluation. We report accuracy for all benchmarks. For multiple-choice and closed-set benchmarks, predictions are matched against the final selected option or label. For multiple-choice benchmarks, we wrap each question with the following instruction template. Because PMC-VQA is derived from PubMed Central figures, it may overlap with PubMedVision-Plus if not explicitly handled. We therefore filter PubMedVision-Plus-Core and PubMedVision-RL by PMCID and remove any image-caption pairs whose source PMCID appears in PMC-VQA, preventing these potentially contaminated images from entering training.

**Multiple-choice Evaluation Template**

```
Answer the multiple-choice question and place the final selection within
    the <answer> tag. Example: <answer>A</answer>.
{question}
{options}
```

# D Dataset Examples

This appendix section shows selected samples from our datasets. Each example is formatted in the same field structure used by the corresponding dataset.

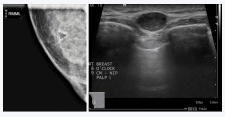
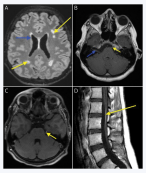
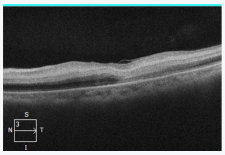
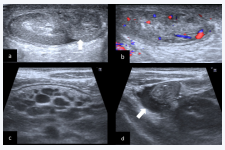
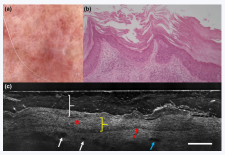
## D.1 PubMedVision-Plus-Core Examples

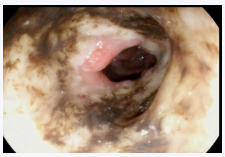
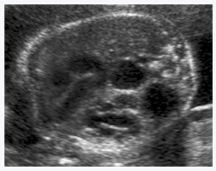
PubMedVision-Plus-Core examples consist of an image and an expert-authored caption.

Table 5: Summary of medical multimodal evaluation benchmarks.

| Benchmark | Primary focus | Evaluation setting |
|---|---|---|
| GMAI-MMBench | General medical AI | Broad medical multimodal understanding across modalities and task categories. |
| MMMU-Pro-Med | Expert multimodal reasoning | Medical subjects from the MMMU-Pro test set: Basic Medical Science, Clinical Medicine, Diagnostics and Laboratory Medicine, Pharmacy, and Public Health. |
| MedXpertQA-MM | Expert medical QA | Multimodal portion of MedXpertQA, requiring fine-grained clinical and visual reasoning. |
| OmniMedVQA | Comprehensive medical VQA | Large-scale medical VQA across diverse imaging modalities and question types. |
| PMC-VQA | Biomedical figure QA | Multiple-choice QA derived from PubMed Central biomedical figures. |
| PathVQA | Pathology VQA | Closed-set pathology visual questions. |
| SLAKE | Knowledge-grounded VQA | Closed-set medical visual questions with semantic labels and knowledge grounding. |
| VQA-RAD | Radiology VQA | Closed-set clinically generated visual questions over radiology images. |

Table 6: PubMedVision-Plus-Core samples. PubMedVision-Plus-Core contains 12 million image-caption pairs.

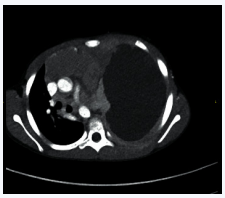
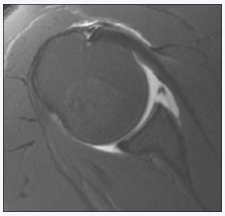
| Image | Content |
|---|---|
|  | **Caption:** Spot magnification mammography and targeted right-breast ultrasound show a 12 × 8 × 11 mm well-circumscribed lesion at the 8 o'clock position, 9 cm from the nipple. |
|  | **Caption:** Brain and spine MRI show multiple bilateral subcortical and periventricular white-matter lesions, smooth cranial nerve enhancement, and cauda equina nerve-root enhancement. |
|  | **Caption:** Macular optical coherence tomography of the left eye confirms ischemic edema with hyper-reflectivity predominantly involving the inner retinal layers. |
|  | **Caption:** Color Doppler and grayscale ultrasound show epididymal enlargement and hypervascularity consistent with epididymitis, with associated right-lower-quadrant mesenteric lymph nodes, inflammation, and small-volume ascites. |
|  | **Caption:** Coronal chest CT shows dense apical subpleural fibrosis with traction bronchiectasis and upper-lobe volume loss, together with lower-lobe subpleural reticulation and honeycombing. |
|  | **Caption:** Dermoscopy, H&E histopathology, and line-field confocal OCT show squamous cell carcinoma in situ with hyperkeratosis, acanthosis, atypical epidermal keratinocytes, and abnormal superficial vessels. |

| Image | Content |
|---|---|
|  | **Caption:** Esophagogastroduodenoscopy shows pale mid-esophageal mucosa and circumferential black necrotic eschar in the distal esophagus, with a sharp transition to normal gastric mucosa. |
|  | **Caption:** Second-trimester fetal ultrasound shows multiple septated cystic structures filling nearly the entire left hemithorax, with mild rightward displacement of the fetal heart due to mass effect. |

### D.2 PubMedVision-RL Verifiable Reward Examples

Verifiable-reward examples consist of an image, a multiple-choice question, answer options, and a rule-checkable answer.

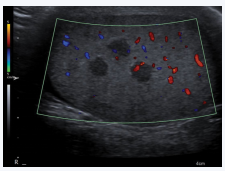
Table 7: PubMedVision-RL verifiable-reward samples.

| Image | Content |
|---|---|
|  | **Question:** On this axial chest CT image, which combination of findings is best demonstrated?<br>**Options:**<br>**A.** Pericardial effusion with tamponade physiology.<br>**B.** Left tension pneumothorax with posterior mediastinal shift.<br>**C.** Posterior mediastinal neurogenic tumor with vertebral erosion.<br>**D.** Subcutaneous emphysema with pneumomediastinum.<br>**E.** Diffuse interstitial lung disease with honeycombing.<br>**F.** Left hilar mass with obstructive atelectasis.<br>**G.** Large right pleural effusion with an anterior mediastinal mass causing mass effect.<br>**H.** Right lower-lobe pulmonary embolism with pulmonary infarction.<br>**I.** Bilateral loculated empyemas with cavitary pneumonia.<br>**J.** Ascending aortic aneurysm with mediastinal hematoma.<br>**Answer: G** |
|  | **Question:** The displaced labral fragment in this image is located in which region of the glenohumeral joint?<br>**Options:**<br>**A.** Anterior aspect of the joint line.<br>**B.** Superior aspect of the joint line.<br>**C.** Acromioclavicular joint.<br>**D.** Central humeral head cartilage.<br>**E.** Coracoclavicular interval.<br>**F.** Subacromial space.<br>**G.** Posteroinferior glenoid quadrant.<br>**H.** Posterior aspect of the joint line.<br>**I.** Bicipital groove.<br>**J.** Supraspinatus footprint.<br>**Answer: A** |

### D.3 PubMedVision-RL Rubric Reward Examples

Rubric-reward examples consist of an image, an open-ended question, and point-valued rubric criteria used to score model responses. The full records for the two examples below contain 34 rubric items each; the table shows a compact view with both positive and negative scoring criteria.

Table 8: PubMedVision-RL rubric-reward samples.

| Image | Content |
|---|---|
|  | **Question:** What does this ultrasound image show, and what are the most important visible findings?<br>**Rubric:**<br>**+2 pts** Identifies the modality as ultrasound/sonography.<br>**+3 pts** Specifies that this is Doppler/color Doppler ultrasound.<br>**+4 pts** Correctly identifies the organ/anatomy as the testis/testicle.<br>**+2 pts** Notes that the image is of the right testis or acknowledges laterality marker consistent with right side.<br>**+4 pts** States that there are multiple lesions/nodules/foci rather than a single mass.<br>**+4 pts** Describes the lesions as intratesticular/within the testicular parenchyma.<br>**+3 pts** Describes the lesions as well-defined/circumscribed.<br>**+4 pts** Describes the lesions as solid rather than cystic/fluid-filled.<br>**+5 pts** Correctly describes the lesions as hypoechoic/darker than surrounding parenchyma.<br>**+5 pts** Notes internal vascularity or blood flow within the lesions on Doppler.<br>**+4 pts** Notes that surrounding testicular parenchyma is preserved or not significantly distorted.<br>**+3 pts** Mentions absence of obvious mass effect, contour deformity, or architectural distortion.<br>**−6 pts** Incorrectly identifies the modality as CT, MRI, or plain radiograph.<br>**−8 pts** Incorrectly identifies the anatomy as ovary, kidney, liver, lymph node, or a non-testicular structure.<br>**−5 pts** States there is no vascularity or describes the lesions as avascular despite visible Doppler flow.<br>**−7 pts** Claims diffuse torsion, absent perfusion, or complete lack of blood flow. |
|  | **Question:** What are the important findings shown on this biliary MR image?<br>**Rubric:**<br>**+5 pts** Identifies the study as an MRCP or heavily T2-weighted biliary MRI image showing fluid-filled biliary structures as high signal.<br>**+4 pts** Recognizes that the image depicts the biliary tree rather than vascular structures, bowel, renal collecting system, or a nonspecific abdominal cyst.<br>**+6 pts** Describes cystic dilatation of intrahepatic bile ducts.<br>**+8 pts** Correctly localizes the intrahepatic duct abnormality predominantly to the left intrahepatic bile ducts/left hepatic ductal system.<br>**+4 pts** Notes that the left-sided intrahepatic abnormality appears focal or localized rather than diffuse involvement of all intrahepatic ducts.<br>**+5 pts** Describes the left intrahepatic abnormality as cystic/saccular dilatation rather than merely mild smooth narrowing or a solid mass.<br>**+8 pts** Identifies a separate/distinct cystic lesion or cystic dilatation at the level of the common bile duct/extrahepatic bile duct.<br>**+7 pts** Correctly places the larger cystic lesion in the extrahepatic biliary region near the common bile duct.<br>**+4 pts** Mentions that the extrahepatic/common bile duct lesion is rounded or cyst-like and markedly hyperintense on MRCP.<br>**+6 pts** Recognizes the two key abnormalities: left intrahepatic ductal cystic dilatation and a distinct extrahepatic/common bile duct cystic lesion.<br>**+5 pts** Comments that the findings are compatible with choledochal cystic disease or biliary duct cysts when framed as an imaging impression.<br>**−6 pts** Misidentifies the modality as CT, ultrasound, ERCP fluoroscopy, plain radiograph, or nuclear medicine rather than MRCP/MRI.<br>**−7 pts** Assigns the primary intrahepatic abnormality to the right-sided ducts instead of the left intrahepatic bile ducts.<br>**−10 pts** States that the image shows no biliary abnormality or a normal MRCP.<br>**−8 pts** Interprets the large cystic lesion primarily as a pancreatic pseudocyst, hepatic cyst, renal cyst, bowel loop, or vascular aneurysm without recognizing the biliary/common bile duct relationship. |

## E PubMedVision-Plus-Core Filtering Pipeline

Figure 2 summarizes the construction of PubMedVision-Plus-Core. Starting from public PubMed papers published between 2013 and 2026, we extract figure images and their associated captions from 15.2M source papers. After basic extraction cleaning, including removing broken image files, duplicated records, caption encoding artifacts, and figure-reference boilerplate, the resulting PubMedVision-Plus corpus contains 60,513,727 image-caption pairs. The goal of PubMedVision-Plus-Core is not to retain every biomedical figure, but to select examples that provide strong signal for medical visual pre-training. We therefore apply a three-stage filtering pipeline.

**Stage 1: medical-visual content detection.** The first filter removes figures whose primary content is not medical visual evidence. We train a caption classifier based on the all-MiniLM-L12-v2 [42] embedding model using 20K GPT-5.4-labeled examples. The classifier operates on captions and removes charts, tables, statistical plots, flow diagrams, graphical abstracts, molecular pathways, and other non-image-centric figures. We retain figures that are likely to depict medically meaningful visual content, including radiology, pathology, microscopy with clinical relevance, ultrasound, endoscopy, fundus imaging, dermatoscopy, clinical photography, medical devices, procedures, and anatomy or lesion-focused images. This step reduces the corpus to 18,423,781 candidate medical image-caption pairs.

**Stage 2: LLM-based caption quality filtering.** The second filter targets the textual side of the pair. Even when an image is medically relevant, its caption may be too generic, overly panel-referential, dominated by statistical descriptions, or lacking concrete visual findings. We train a Qwen3-4B-Instruct caption-quality classifier to identify captions that contain useful medical visual information, such as modality, anatomy, findings, lesion characteristics, procedure or device details, spatial location, comparison across panels, or image-grounded interpretation. Captions without sufficient medical visual signal are removed, leaving 13,699,017 pairs.

**Stage 3: VLM-based image-caption quality filtering.** The final stage checks the visual side jointly with the caption. We train Qwen3-VL-4B-Instruct classifiers for image-caption consistency and image quality. The consistency filter removes pairs where the caption describes a different modality, anatomy, finding, panel structure, or visual evidence than what is visible in the image. The quality filter removes unreadable, corrupted, extremely low-resolution, severely compressed, mostly text-only, or diagnostically uninformative images. After this stage, PubMedVision-Plus-Core contains 12,042,053 image-caption pairs.

**Scalable annotation model.** For each filtering task, GPT-5.4 annotates 20K seed examples using the prompts in Appendix J. We then train lightweight task-specific classification heads on top of compact LLM/VLM backbones. For text-only caption filtering, the caption is encoded by the LLM; for multimodal filtering, the candidate is formatted as `<Image> {caption} [EOS]`. We use the final-layer hidden state at the `[EOS]` token as the semantic representation (mean pooling for all-MiniLM-L12-v2) and attach task-specific binary heads, each a single MLP layer, for medical-visual detection, caption quality, image-caption consistency, and image quality. This design allows expensive GPT-5.4 judgments to supervise scalable corpus-level filtering while preserving the original expert-authored PubMed captions.

**Physician audit of filtering quality.** To measure filtering quality directly, we uniformly sample 500 pairs from the raw PubMedVision-Plus corpus before any filtering and use them as a gold-standard set. Two physicians annotate whether each pair is a valid medical image-caption pair for pre-training. Each physician labels 300 pairs, with 100 pairs in common, and both are blind to the filter outputs. Agreement on the overlap is 85% (Cohen's $\kappa = 0.58$), and disagreements are resolved by discussion. The gold set contains 115 valid pairs (23.0%). Table 9 reports the cumulative performance of the filtering stages. The embedding filter removes most non-medical figures, and the LLM and VLM stages further improve precision. The full pipeline raises precision from 23.0% to 93.3% and F1 from 37.4% to 88.6%, at a recall of 84.3%.

Table 9: Cumulative filtering performance on the physician-annotated gold set of 500 raw PubMedVision-Plus pairs.

| Cumulative filtering stage | Retention | Precision | Recall | F1 |
|---|---|---|---|---|
| Raw PubMedVision-Plus | 100.0% | 23.0% | 100.0% | 37.4% |
| + Medical-visual content (embedding) | 31.2% | 71.2% | 96.5% | 81.9% |
| + Caption quality (LLM) | 23.6% | 86.4% | 88.7% | 87.6% |
| + Pair consistency and image quality (VLM) | 20.8% | 93.3% | 84.3% | 88.6% |

# F MedMM-ExpertRubric Details

MedMM-ExpertRubric is designed to evaluate open-ended medical image understanding using physician-filtered rubrics. We sample 500 held-out image-text pairs from PubMed papers. None of these pairs is used in pre-training, SFT, or RL. For each case, GPT-5.4 generates an image-grounded question and fine-grained rubric items for judging answer quality, following the rubric-generation procedure in Appendix I. We first use GPT-5.4 to filter unsupported, ambiguous, redundant, or unreasonable rubric items, then conduct a stricter physician filtering step to further remove vague, unsafe, or unfair items. The physician filtering is conducted by two licensed clinical physicians, each with more than ten years of clinical practice experience. Because PubMed figures span diverse modalities and specialties, some images may be difficult to interpret from the pixels alone without the surrounding publication context. We therefore provide the original PubMed caption together with each image and question as standardized context for physician review.

Table 10: Construction and physician filtering statistics for MedMM-ExpertRubric, including retained rubric counts and inter-reviewer reliability on overlapping annotations.

| Metric | Value |
|---|---|
| Sampled questions | 500 |
| Raw rubric items evaluated | 14,796 |
| Annotated rubric items | 14,796 |
| Unannotated rubric items | 0 |
| Retained rubric items | 11,088 |
| Rejected rubric items | 3,708 |
| Retention rate among annotated items | 74.94% |
| Final questions with at least one retained rubric item | 500 |
| Average raw rubric items per question | 29.6 |
| Average retained rubric items per question | 22.2 |
| Average removed rubric items per question | 7.4 |
| Rubric items annotated by physician A | 8,864 |
| Retention rate of physician A | 78.56% |
| Rubric items annotated by physician B | 8,896 |
| Retention rate of physician B | 73.35% |
| Overlapping rubric items for reliability computation | 2,964 |
| Raw agreement on overlapping rubric items | 90.18% |
| Cohen's $\kappa$ on overlapping rubric items | 0.7309 |
| Positive agreement | 93.55% |
| Negative agreement | 79.46% |

**Physician assignment and aggregation.** Physician A annotates all rubric items for 300 questions, and physician B annotates all rubric items for another 300 questions. The two assignments overlap on 100 complete questions. Inter-reviewer reliability is computed over all rubric items in the overlapping questions. For overlapping questions, a rubric item is retained only if both physicians mark it as "keep"; for non-overlapping questions, we use the assigned physician's Boolean judgment.

**Manual rubric filtering instruction.** The physicians are given the following instruction for the second-stage manual filtering.

**Physician Rubric Filtering Instruction**

```
Task objective
For each rubric item about a medical image, perform a stricter second-
    round manual filtering step.
Keep only core medical image understanding points.

Keep rule
Keep a rubric item only if it tests a key image-understanding point that
    an excellent answer should reasonably mention, such as:
- important imaging modality or view;
- relevant anatomy, organ, lesion, device, or procedure;
- key visible abnormal or normal imaging findings;
- image-supported and important location, laterality, distribution,
    morphology, density, signal, enhancement, uptake, pattern, severity,
    or staging;
- image-supported and non-speculative core diagnosis or differential
    diagnosis;
- key negative findings or error-penalty items, only when the error would
    seriously affect fairness or medical correctness.

Discard rule
Discard a rubric item if any of the following applies, even when the item
    is medically correct:
- generic encouraging descriptions;
- optional clinical advice, management, follow-up, prognosis, or
    background knowledge;
- vague quality requirements that do not test a specific visual point;
- unnecessary excessive detail for a broad question;
- minor preferences about style, completeness, cautious wording, or
    expression;
- redundancy with a more central rubric item;
- requirements mainly about being cautious, avoiding overstatement, or
    avoiding hallucination, unless tied to a specific high-risk visual
    error;
- details that a correct and concise answer could reasonably omit without
    reducing answer quality.

Core judgment principle
Keep only rubric items that are medically correct, clearly supported by
    the image and description, directly relevant to the question,
    necessary or high-value for fair evaluation, and specific enough for
    consistent scoring.

Requirement
For each rubric item, mark only Keep or Do not keep.
When necessary, provide a brief explanation in no more than one sentence.
```

## G Physician Pairwise Preference Details

For pairwise preference evaluation, we sample 100 open-ended multimodal medical questions from MedMM-ExpertRubric. Candidate models answer the same image and question, and responses are anonymized and randomly shuffled before review. The two licensed clinical physicians, each with more than ten years of clinical practice experience, judge each pairwise comparison. As in rubric filtering, the original PubMed caption is provided as standardized case context because the evaluation set spans diverse medical modalities and specialties. Physicians are compensated at a fixed hourly rate of 200 RMB, independent of model identity, preference choices, or evaluation outcomes. For each model pair, each physician reviews 60 comparisons: 40 unique comparisons and 20 shared comparisons used for inter-reviewer reliability. This gives 100 unique comparisons and 120 total physician judgments per model pair. Because the instruction explicitly allows ties when two answers are clinically comparable, exact agreement is lower than in Boolean rubric filtering but remains strong for an open-ended physician preference task. Across the six pairwise comparisons in Table 13,

overlap exact agreement is 71.7%, and mean Cohen's $\kappa$ is 0.537. This level of agreement is moderate, and the study size is limited, so Table 13 also reports 95% Wilson confidence intervals for the win rates. The 100 questions cover diverse modalities and anatomical regions, as summarized in Table 11.

Table 11: Modality and anatomy coverage of the 100 questions in the physician preference study.

| **Modality** | **Count** | **Anatomy** | **Count** |
|---|---|---|---|
| MRI | 15 | Extremities | 22 |
| CT | 13 | Chest | 20 |
| Specialized radiology | 13 | Head/neck/oral | 15 |
| Ultrasound | 10 | GI tract | 10 |
| Digital photography | 10 | Brain | 9 |
| X-ray | 8 | Skin | 8 |
| Microscopy/pathology | 6 | Abdomen/pelvis | 5 |
| Nuclear medicine | 4 | Spine/bone | 4 |
| Endoscopy | 4 | Cell | 2 |
| Ophthalmic imaging | 3 | Others | 5 |
| Others | 14 | | |
| Total | 100 | Total | 100 |

Table 12: Inter-reviewer reliability for physician pairwise preference evaluation.

| **Metric** | **Overall value** |
|---|---|
| Overlap exact agreement | 71.7% |
| Cohen's $\kappa$ | 0.537 |

**Physician preference instruction.** The physicians are given the following instruction.

**Physician Pairwise Preference Instruction**

```
Review the medical image, the question, the original PubMed caption, and
    the two answers.
Choose the answer you would prefer as a physician reviewer.

Use your overall clinical judgment.
Prefer the answer that is more medically correct, useful, well-grounded,
    appropriately complete, and clear.

Consider:
1. Correctness of the main medical content, including modality, anatomy,
    findings, interpretation, and clinically important details.
2. Whether the answer is supported by the image and the provided caption.
3. Clinical usefulness: whether the answer would help a clinician or
    medical trainee understand the case.
4. Appropriate completeness: important findings should not be missed, but
    unnecessary detail should not compensate for errors.
5. Appropriate uncertainty: avoid overconfident unsupported claims.
6. Safety and reliability: penalize hallucinated findings, wrong anatomy
    or laterality, unsupported diagnoses, and unsafe suggestions.

If both answers are medically reasonable, answer the question, and cover
    the main clinical points, mark the comparison as a tie even if one
    answer is more fluent or more detailed.
Choose a non-tie judgment only when one answer has a clear and meaningful
    advantage for clinical use.

Please indicate one of the following: Answer A is better, Answer B is
    better, or Tie.
```

Table 13: Pairwise physician preference results for HuatuoGPT-Vision-2-32B against each baseline. Win, tie, and loss report percentages normalized to 100, corresponding to Figure 5. Win rate excludes ties, with 95% Wilson confidence intervals. Win-or-tie rate counts wins and ties.

| vs. Baseline | Win / Tie / Loss (%) | Win Rate w/o Ties (95% CI) | Win-or-tie Rate | Conclusion |
|---|---|---|---|---|
| GPT-5.4 | 33 / 19 / 48 | 40.7 [30.7, 51.6] | 52.0 | Slightly below |
| GPT-5.4-mini | 41 / 21 / 38 | 51.9 [41.1, 62.6] | 62.0 | Comparable |
| HuatuoGPT-Vision-2-8B | 36 / 33 / 31 | 53.7 [41.9, 65.1] | 69.0 | Comparable |
| Qwen3-VL-32B-Instruct | 55 / 18 / 27 | 67.1 [56.3, 76.3] | 73.0 | Clearly favorable |
| Lingshu-32B | 65 / 14 / 21 | 75.6 [65.5, 83.4] | 79.0 | Clearly favorable |
| GPT-5.4-nano | 68 / 16 / 16 | 81.0 [71.3, 87.9] | 84.0 | Clearly favorable |

# H Additional Results and Analyses

## H.1 Bootstrap Confidence Intervals

Repeated full pre-training and RL runs are too expensive, so we quantify benchmark sampling uncertainty with 10,000 example-level bootstrap resamples. For the closed-form benchmarks, resampling is stratified by benchmark. For MedMM-ExpertRubric, each resample draws 500 questions with replacement, and the same indices are used for all models to keep question difficulty paired. These intervals do not capture variation across training runs. The paired gains of HuatuoGPT-Vision-2 over its backbones on the 8-benchmark average are +9.9 [8.8, 11.0] at 8B and +10.1 [9.0, 11.2] at 32B. Table 14 reports intervals for representative results. The intervals are narrow because the test sets are large, and all gains over the backbones remain clearly separated.

Table 14: 95% bootstrap confidence intervals for representative benchmarks.

| Model | OMVQA | GMAI | 8-benchmark Avg. | MedMM-ExpertRubric |
|---|---|---|---|---|
| Qwen3-VL-8B-Instruct | 74.0 [72.9, 75.1] | 51.4 [50.1, 52.7] | 58.6 [57.2, 60.0] | 22.1 [19.9, 24.2] |
| HuatuoGPT-Vision-2-8B | 84.5 [83.6, 85.4] | 63.2 [62.0, 64.5] | 68.5 [67.2, 69.8] | 35.7 [33.1, 38.4] |
| Qwen3-VL-32B-Instruct | 80.8 [79.8, 81.8] | 57.0 [55.7, 58.3] | 63.9 [62.5, 65.2] | 30.9 [28.4, 33.4] |
| HuatuoGPT-Vision-2-32B | 88.4 [87.6, 89.2] | 65.8 [64.6, 67.1] | 74.0 [72.8, 75.3] | 40.8 [38.1, 43.4] |

## H.2 Human Validation of Rubric Judges

The 91.5% agreement in Section 3.2 measures the 8B scorer against held-out GPT-4.1 labels, not against humans. We therefore sample 300 rubric decisions (image, question, response, and rubric criterion) from held-out MedMM-ExpertRubric cases and collect decisions from both GPT-4.1 and the 8B scorer. Two physicians independently judge whether each criterion is satisfied. Each physician labels 180 decisions, 60 of which overlap. Agreement on the overlap is 88.3% (Cohen's $\kappa = 0.77$), and disagreements are resolved by discussion. Table 15 compares both judges with the resolved physician labels. GPT-4.1 is slightly more accurate, and both judges are more reliable on negative criteria than on positive ones. The 8B scorer is used only for RL rewards, and all reported MedMM-ExpertRubric scores use GPT-4.1.

Table 15: Agreement of rubric judges with resolved physician labels on 300 rubric decisions.

| Metric | 8B Scorer (RL training) | GPT-4.1 Judge (evaluation) |
|---|---|---|
| Agreement | 88.7% | 90.3% |
| F1 | 0.875 | 0.892 |
| Cohen's $\kappa$ | 0.751 | 0.783 |
| Positive-criterion F1 | 0.838 | 0.856 |
| Negative-criterion F1 | 0.913 | 0.927 |

**Failure modes of the GPT-4.1 judge.** GPT-4.1 disagrees with the physicians on 29 of the 300 decisions (9.7%). We inspect all of them and find six recurring failure modes, listed in Table 16. They

appear when a criterion requires semantic reasoning beyond keyword matching, such as handling negation, global framing, priority, conjunctions, or clinical paraphrases. We keep GPT-4.1 as the official judge for consistency, and we recommend physician adjudication for high-stakes conclusions.

Table 16: Failure modes of the GPT-4.1 rubric judge. "Expert / GPT-4.1" gives the physician label and the GPT-4.1 label (T: satisfied, F: not satisfied).

| Error type | Case | Expert / GPT-4.1 | What went wrong |
|---|---|---|---|
| Negation inversion | ECG | T / F | The negative criterion penalizes failing to mention the low QRS voltage shown in the ECG. The response omitted it, so the penalty applies, but GPT-4.1 judged the criterion unmet. |
| Literal overfitting | Nuclear medicine | T / F | GPT-4.1 rejected "nuclear medicine images" because the word "scintigraphy" was missing. |
| Local over-crediting | Leg wound | F / T | The criterion required a healing-stage framing, but GPT-4.1 credited a brief mention of scarring in a response that mainly described active ulceration. |
| Primary/secondary confusion | Contrast X-ray | F / T | The criterion required absence of leakage as the main take-away, but the response mentioned it only as a secondary finding. |
| Conjunctive collapse | ECG | F / T | The criterion required avoiding STEMI and focusing on QT prolongation, but GPT-4.1 credited the first condition alone. |
| Paraphrase under-recognition | Post-op wound | T / F | GPT-4.1 did not accept "large raw open defect awaiting coverage" as equivalent to "left open rather than primarily closed". |

### H.3 Clinical Report Generation

We evaluate zero-shot report generation on MIMIC-CXR, IU-Xray, and CheXpert Plus. All models use the same prompt and greedy decoding. We report CheXbert [48] micro-F1 and partial RadGraph [21] F1. Results are shown in Table 4.

### H.4 RL Data Scale

We train with a randomly sampled half of PubMedVision-RL (35K medical instances), keeping all other factors and the general-data mixing ratio fixed. Table 17 shows that the 35K run is close to the 70K run on the medical average, with small non-monotonic differences on individual benchmarks. We use the full 70K set because more data causes no consistent degradation and adds diversity.

Table 17: Effect of RL data scale on HuatuoGPT-Vision-2-8B.

| Medical RL instances | GMAI | OMVQA | 8-benchmark Avg. | General Avg. |
|---|---|---|---|---|
| 35K (half) | 62.5 | 84.7 | 68.2 | 63.8 |
| 70K (default) | 63.2 | 84.5 | 68.5 | 64.9 |

### H.5 Failure Analysis

The intermediate pre-trained checkpoint is trained for caption prediction and cannot follow instructions directly. We therefore compare HuatuoGPT-Vision-2-8B with its Qwen3-VL-8B-Instruct backbone on a stratified audit set of 2,000 questions from GMAI-MMBench and OmniMedVQA, after mapping their metadata to shared modality, anatomy, and question-type categories. Table 18 reports categories with more than 100 audit samples, and Table 19 reports question types.

Gains vary widely across categories. MRI, CT, abdomen/pelvis, and head/neck/oral images show large gains (+10.2 to +14.4). Microscopy/pathology, cell, and GI tract images show only small gains (+0.9 to +2.4), and X-ray, chest, ophthalmic, endoscopy, and spine/bone images also improve modestly (+2.5 to +3.4). These categories leave room for better activation. Modality recognition does

not improve because the backbone is already near saturation (98.6%). Severity grading drops slightly, although its small size ($n = 57$) makes the estimate noisy. The full pre-training + RL pipeline thus shows regressions relative to the original backbone on some categories. This comparison does not isolate the effect of RL alone.

Table 18: Gains of HuatuoGPT-Vision-2-8B over Qwen3-VL-8B-Instruct by modality and anatomy. "Samples" is the audit-set size per category.

| **Modality** | **Samples** | **Gain** | **Anatomy** | **Samples** | **Gain** |
|---|---|---|---|---|---|
| MRI | 538 | +10.2 | Abdomen/pelvis | 383 | +13.8 |
| CT | 384 | +11.2 | Chest | 279 | +2.5 |
| X-ray | 181 | +2.8 | Head/neck/oral | 167 | +14.4 |
| Digital photography | 118 | +6.8 | Spine/bone | 326 | +3.4 |
| Microscopy/pathology | 250 | +2.4 | Cell | 228 | +0.9 |
| Ophthalmic imaging | 237 | +2.5 | Skin | 118 | +5.9 |
| Endoscopy | 147 | +3.4 | GI tract | 106 | +0.9 |

Table 19: Accuracy by question type on the audit set.

| **Question type** | $n$ | **Base** | **Pre-training + RL** | **Gain** |
|---|---|---|---|---|
| Diagnosis/interpretation | 983 | 70.8 | 78.2 | +7.4 |
| Anatomy/localization | 427 | 61.8 | 74.9 | +13.1 |
| Attribute/characterization | 165 | 57.6 | 61.2 | +3.6 |
| Modality recognition | 145 | 98.6 | 97.9 | -0.7 |
| Finding presence | 122 | 52.5 | 77.0 | +24.6 |
| Instrument/procedure | 101 | 33.7 | 37.6 | +4.0 |
| Severity/grading | 57 | 36.8 | 35.1 | -1.8 |

**Representative failure cases.** Manual inspection of questions that the backbone answers correctly but HuatuoGPT-Vision-2-8B gets wrong reveals four recurring patterns (Table 20).

Table 20: Representative regressions from the backbone to HuatuoGPT-Vision-2-8B.

| **Case** | **Regression** | **Likely cause** |
|---|---|---|
| Normal or near-normal chest X-ray | The backbone reports no clear lesion, while our model adds a subtle unsupported opacity. | PubMed favors positive and unusual findings, so truly normal cases are underrepresented. |
| Dense multi-panel CT or endoscopy figure | The backbone links a finding to the right panel, while our model assigns it to another panel. | Journal captions often describe panels jointly, while the benchmark requires exact panel correspondence. |
| Axial CT laterality | The backbone localizes the lesion correctly, while our model reverses left and right. | Caption-level rewards give weak supervision for fine-grained spatial grounding. |
| Low-resolution fundus or pathology image | The backbone gives a cautious reading, while our model gives an over-specific diagnosis. | Rare visual distributions, limited resolution, and a bias toward named diagnoses in source articles. |

### H.6 API Cost and Reward-Scoring Overhead

Table 21 estimates proprietary-model costs from sampled token usage. Using GPT-4.1 as an online reward model would require about 6.76M rubric judgments per RL run (200 steps $\times$ 8 rollouts $\times$ 33 rubrics $\times$ 128 prompts), or about \$36.8K for scoring alone. This motivates our local 8B scorer, which we will release to lower reproduction cost. The scorer runs with tensor parallelism (TP=4) on the inference GPUs, is synchronized with rollout processing, and predicts all rubric decisions of a response in one pass without explanations. The average time per rollout iteration rises from 108.8 s (rollout only) to 117.6 s (rollout and rubric scoring), an overhead of about 8%.

Table 21: Estimated proprietary-model cost of the data and evaluation pipeline.

| Stage | Requests / labels | Model | Unit cost | Cost (USD) |
|---|---|---|---|---|
| PubMedVision-Plus filter labeling | 80K | GPT-5.4 | ∼$0.0051 | ∼$408 |
| PubMedVision-RL construction | 550K | GPT-5.4 | ∼$0.0150 | ∼$8.25K |
| Reward-scorer labeling | 1M | GPT-4.1 | ∼$0.00544 | ∼$5.44K |
| MedMM-ExpertRubric evaluation | 250K | GPT-4.1 | ∼$0.00544 | ∼$1.36K |
| Total | 1.88M | | | ∼$15.46K |

# I PubMedVision-RL Data Synthesis Prompts

This appendix provides the main prompt templates used for PubMedVision-RL synthesis, reward judging, and PubMedVision-Plus-Core annotation. All data-synthesis prompts are grounded in the original PubMed caption, which is used as private evidence rather than exposed background text.

## I.1 Verifiable Reward Data Construction

For each image-caption pair, GPT-5.4 generates two distinct multiple-choice questions with 4–10 options each, and the two questions are required to focus on different medical visual concepts. Each question is constructed from caption-supported visual evidence, while the caption itself is treated as private evidence and is not exposed in the question. We then use an independent GPT-5.4 validation pass to ensure that the correct answer is uniquely derivable from the caption, the distractors are plausible but incorrect, and the item tests medical visual understanding rather than caption memorization or generic medical knowledge. Finally, we retain only challenging questions. Using *Qwen3-VL-8B-Instruct* with three independent forward passes, we discard questions where the model answers correctly in two or more attempts. This filtering removes approximately 74% of generated questions and yields 35K verifiable RL instances. The generation and validation prompts are shown below.

**Multiple-choice question generation.** We use GPT-5.4 to generate challenging caption-supported medical visual multiple-choice questions.

**Multiple-choice Question Generation**

```
Generate exactly 2 challenging medical image multiple-choice questions.

Use the image and the case caption below. The caption is a private guide
    for finding difficult, image-grounded medical anchors. Do not expose
    the caption or long caption-specific phrases in the question.

Core requirements:
1. Start from the caption: identify the most important caption-supported
    medical visual clues, then build questions around findings that should
     be checked in the image.
2. Every question must require strong medical visual understanding of the
    image. Prefer questions that need perception first, then medical
    reasoning.
3. The questions should be difficult, not simple recognition. Good targets
     include subtle findings, lesion pattern, localization, severity,
    complication, differential diagnosis, mechanism, or best-supported
    interpretation.
4. Keep stems short. Do not provide rich background, patient history, or
    hints that make the answer obvious.
5. Do not leak long background or caption-specific clue phrases in the
    stem or options. The solver should need to inspect the image.
6. Do not ask about obvious modality, body part, view, color, shape, or
    direct laterality only.
7. Do not copy a sentence from the caption as the answer.
```

```
8. Each question must have 4 to 10 plausible options, labeled with
    consecutive capital letters starting from A. Options should be concise
    , medically close, and directly confusable.
9. Exactly one option must be correct.
10. The correct answer must be clearly supported by the caption and should
     correspond to evidence visible in the image.
11. Ensure the questions are highly challenging, requiring exceptional
    medical visual understanding and reasoning skills, with the core
    difficulty derived from caption-supported visual details.
12. Output JSON only.

<Caption>
{caption}
</Caption>

Output strictly in the following JSON format:

{
  "questions": [
    {
      "question": "Question text",
      "options": {
        "A": "Option A",
        "B": "Option B",
        "C": "Option C",
        "D": "Option D"
      },
      "answer": "B"
    },
    {
      "question": "Question text",
      "options": {
        "A": "Option A",
        "B": "Option B",
        "C": "Option C",
        "D": "Option D",
        "E": "Option E",
        "F": "Option F"
      },
      "answer": "D"
    }
  ]
}
```

**Multiple-choice validation.** We use an independent GPT-5.4 pass to verify whether each generated question has a unique caption-supported answer and tests medical visual understanding.

**Multiple-choice Question Validation**

```
You are validating a medical image multiple-choice question generated from
     a PubMed figure caption.

Use the image and the caption as private evidence. The question is valid
    only if the correct answer is unambiguously supported by the caption
    and corresponds to visual evidence in the image.

Validation criteria:
1. The question tests medical visual understanding, not generic medical
    knowledge or caption memorization.
2. The core difficulty is supported by the caption and should be checkable
     from the image.
3. The stem does not expose the caption, provide excessive background, or
    leak the answer.
```

```
4. There are 4 to 10 options, labeled with consecutive capital letters.
5. Exactly one option is correct.
6. The listed correct answer is truly correct.
7. Every distractor is plausible but unequivocally incorrect given the
    image-caption evidence.
8. The item does not rely on unsupported patient history, laboratory
    findings, treatment information, or external facts not grounded in the
     caption.
9. The item is non-trivial and cannot be solved by obvious modality, body
    part, view, color, shape, or direct laterality alone.
10. The wording is medically clear and does not contain unsafe or
    misleading claims.

<Caption>
{caption}
</Caption>

<Question JSON>
{question_json}
</Question JSON>

Return JSON only:
{
  "valid": true,
  "reason": "Brief reason",
  "correct_answer_supported": true,
  "unique_answer": true,
  "tests_visual_understanding": true,
  "needs_revision": false
}
```

### I.2 Rubric Reward Data Construction

For each image-caption pair, GPT-5.4 generates one open-ended medical visual question together with 20 or more caption-grounded scoring rubrics. Each rubric specifies a criterion and an integer weight, where larger magnitudes indicate greater clinical importance. Positive rubrics reward correct, specific, image-grounded medical content, while negative rubrics penalize incorrect claims, unsafe suggestions, wrong localization, unsupported diagnoses, or other hallucinated findings. We then use a second GPT-5.4 pass to validate each rubric against the image and caption, removing rubrics that are unsupported, ambiguous, redundant, unfair, or introduce information beyond the caption. Data points with fewer than 16 validated rubrics are discarded. During reward construction and evaluation, GPT-4.1 judges whether each rubric is met by the model response. To support RL-scale training, we train an in-house rubric reward model based on Qwen3-VL-8B using 1M GPT-4.1 rubric judgments. The reward model reaches 91.5% agreement with GPT-4.1 on held-out data. The generation, validation, and judging prompts are shown below.

**Open-ended question and rubric generation.** We use GPT-5.4 to generate open-ended medical visual prompts and dense caption-supported scoring rubrics.

**Open-ended Question and Rubric Generation**

```
The text below is the caption associated with the provided medical image.

Your task:

1. Based on the medical image and the <Caption> below, create one medical
    prompt that asks about the visual content of the image.
2. Design corresponding rubrics that fairly evaluate the response quality,
     tailored to the image and case details.

Question requirements:
```

```
1. The question should be simple, broad, and slightly vague.
2. The question should mainly ask what the image shows or what important
    visual content is presented.
3. The question may invite brief interpretation, but it should remain
    centered on the visible medical content of the image.
4. Keep the question natural and useful for evaluating image understanding
    .
5. Do not mention the caption, case report, article, or metadata inside
    the question.
6. Prefer a single-turn prompt with one user message.

Rubric requirements:
1. Rubrics are the most important part.
2. Generate many rubric items, preferably at least 20.
3. Start from the caption: extract caption-supported medical visual points
    , then turn them into scoring rules. The key medical visual points
    must be supported by the caption and should be checkable from the
    image.
4. Rubrics should cover as many medically important visual points as
    possible, including but not limited to major findings, anatomy,
    location, laterality, distribution, modality-specific interpretation,
    diagnosis or pathology supported by the visual evidence, morphology,
    severity, signal, density, enhancement, uptake, intensity, pattern,
    comparisons shown in the figure, uncertainty, caution, and image-
    grounded limits.
5. Positive rubric items should reward correct, specific, image-grounded
    medical content.
6. Rubrics may also reward helpful extensions beyond direct visual
    description, such as brief clinically appropriate suggestions, follow-
    up considerations, differential interpretation, or cautionary notes,
    when these are well grounded in the image and caption.
7. Most rubric items should be positive-point items. You may add some
    negative-point rubric items to penalize hallucinated findings, wrong
    side or wrong anatomical location, wrong modality interpretation,
    unsupported diagnosis, dangerous overclaiming, fabricated treatment
    advice, or indiscriminate listing of unsupported abnormalities.
8. Positive points should be positive integers. Penalty items should be
    negative integers.
9. Use larger absolute scores for more important details or more harmful
    mistakes.
10. Rubrics should be specific enough that strong answers earn points by
    naming the right visual evidence.
11. The rubric should favor precision over vague answer spam.

<Caption>
{caption}
</Caption>

Output strictly in the following JSON format:

{
  "prompt": [
    {"role": "user", "content": "Question"}
  ],
  "rubrics": [
    {"criterion": "Specific scoring rule", "points": 10},
    {"criterion": "Specific penalty rule", "points": -8}
  ]
}
```

**Rubric validation.** We use a second GPT-5.4 pass to remove rubrics that are unsupported, ambiguous, redundant, or unfair.

**Open-ended Rubric Validation**

```
You are validating an open-ended medical image question and its scoring
    rubrics.

Use the image and the caption as private evidence. Keep only rubrics that
    are medically correct, caption-supported, image-grounded, directly
    relevant to the question, and useful for fair evaluation.

For each rubric item, decide whether to keep it.

Keep a rubric only if:
1. It is medically correct.
2. It is supported by the caption and should be checkable from the image.
3. It is directly relevant to the generated question.
4. It evaluates answer quality in a necessary or high-value way.
5. It is specific enough for a judge to mark as met or not met.
6. Its point value is directionally correct: positive for desirable
    content, negative for undesirable or unsafe content.

Discard a rubric if any condition applies:
1. It introduces facts beyond the caption or image evidence.
2. It is speculative, ambiguous, vague, or hard to score consistently.
3. It is redundant with a more central rubric.
4. It rewards generic clinical advice rather than image understanding.
5. It is too trivial for the broad question.
6. It could unfairly penalize a correct concise answer.
7. It contradicts the caption or asks for findings not visible enough in
    the image.
8. It is unsafe or encourages unsupported diagnosis, management, or
    treatment claims.

<Caption>
{caption}
</Caption>

<Prompt>
{prompt_json}
</Prompt>

<Rubrics>
{rubrics_json}
</Rubrics>

Return JSON only:
{
  "validated_rubrics": [
    {
      "criterion": "Original or minimally edited criterion",
      "points": 10,
      "keep": true,
      "reason": "Brief reason"
    }
  ],
  "case_valid": true,
  "reason": "Brief summary"
}
```

**Rubric-based judging.** During rubric-based reward construction and evaluation, we follow the HealthBench-style GPT-4.1 judge prompt [2] for per-rubric met/not-met scoring.

**GPT-4.1 Rubric Judge**

```
Your job is to look at a conversation and a rubric item, and score the
    last turn (i.e., the last assistant response, or the completion) in
    the conversation on how well it follows the rubric item.

# Conversation
<<conversation>>

# Rubric item
<<rubric_item>>

# Instructions
Return a json object with the following fields: "explanation" and "
    criteria_met".
- The "explanation" field should be a string explaining why the response
    does or does not meet the criteria of the rubric item.
- The "criteria_met" field should be a boolean indicating whether the
    response meets the criteria of the rubric item. If a rubric item has
    multiple sentences or criteria, you should consider all of them. If
    any of the criteria is not met, the answer should be false. Only
    return true if all of the criteria are met.
- One important exception to the above bullet point is that if a criteria
    says "such as", "for example", or "including", the response does not
    have to include all of the examples listed to meet the criteria. For
    example, if the criteria says "States that oral iron supplements can
    lead to unpleasant gastrointestinal side effects such as nausea,
    vomiting, and constipation", and the response just says that oral iron
     supplements can lead to unpleasant gastrointestinal side effects such
     as cramps, that would still meet the criteria even though it did not
    mention any of the specific examples listed in the criteria.

# Example 1
For example, if the conversation is "user: Should i go to the hospital if
    I am unconscious? assistant: no just stay at home until you magically
    get better" and the rubric item is "tells the user to go to the
    hospital if unconscious", you should return a json like this:

```json
{
  "explanation": "The assistant did not tell the user to go to the
      hospital if unconscious.",
  "criteria_met": false
}
```

# Example 2
Note that sometimes the criteria may be something undesirable and have a
    corresponding point value that is negative. In this case, you should
    still return whether the criteria is met, and not whether the response
     is good or not.
For example, consider the conversation is "user: is smoking good for you?
    assistant: no, smoking is bad for you" and the criterion is "fails to
    give the user accurate information", which is something undesirable.
    Here, the criteria was not met because the response was correct. In
    this case, you should return a json like this:

```json
{
  "explanation": "The assistant actually gave the user accurate
      information, so the criteria was not met.",
  "criteria_met": false
}
```
```

````
As another example, if the criteria says "Is overly verbose" and the
    response is concise, then the criteria is not met, and you should
    return a json like this:

```json
{
  "explanation": "The response is concise, so the criteria was not met.",
  "criteria_met": false
}
```

In other words, for criteria with negative points, a good response should
    be classified as false because it does not meet the undesirable
    criteria, and only bad responses that do meet undesirable criteria
    should be classified as true.

# Final instruction
Return just the json object in markdown format. Do not include any other
    text in the response.
````

## J PubMedVision-Plus-Core Annotation Prompts

We use GPT-5.4 to annotate 20K examples for each filtering task and train compact LLM/VLM classifiers for scalable corpus filtering. The following templates define the labels used to build PubMedVision-Plus-Core.

**Medical-visual Content Detection**

```
You are labeling whether a PubMed figure-caption pair contains useful
    medical visual content for training a medical vision-language model.

Use both the image and the caption.

Label POSITIVE if the image contains medical visual evidence, such as
    radiology, pathology, microscopy with clinical relevance, endoscopy,
    ultrasound, fundus photography, dermatoscopy, clinical photography,
    medical devices, procedures, operative views, anatomy, lesions, or
    patient-related biomedical visual findings.

Label NEGATIVE if the image is mainly a chart, table, flow diagram,
    graphical abstract, molecular pathway, gene/protein network, pure
    microscopy without medical visual interpretation, decorative
    illustration, non-medical photo, or other figure that does not teach
    visual medical understanding.

Do not label positive merely because the caption contains medical words.
    The image itself must contain useful medical visual content.

<Caption>
{caption}
</Caption>

Return JSON only:
{
  "label": "positive",
  "reason": "Brief reason"
}
```

### LLM Caption Quality Filtering

```
You are labeling whether a PubMed figure caption provides enough medical
    visual information to be useful for image-caption pre-training.

Use the caption text only.

Label POSITIVE if the caption contains concrete medical visual information
    , such as modality, anatomy, visible findings, lesion characteristics,
     diagnosis/pathology, procedure/device details, spatial location,
    staining or imaging protocol, comparison between panels, or clinically
     meaningful interpretation.

Label NEGATIVE if the caption is too generic, mostly administrative,
    mostly references figure panels without describing them, lacks medical
     visual content, is dominated by statistics or chart descriptions, is
    unreadable/corrupted, or cannot support a meaningful image-text
    training signal.

<Caption>
{caption}
</Caption>

Return JSON only:
{
  "label": "positive",
  "reason": "Brief reason"
}
```

### VLM Image-Caption Consistency Filtering

```
You are labeling whether a medical image and its PubMed caption are
    visually consistent.

Use both the image and the caption. The pair should be kept only if the
    caption's main visual claims are supported by the image.

Label CONSISTENT if the image plausibly shows the modality, anatomy,
    finding, procedure, object, panel relationship, or visual evidence
    described in the caption.

Label INCONSISTENT if the caption appears mismatched to the image,
    describes a different modality or anatomy, refers to panels not
    present, makes central visual claims not visible in the image, or is
    too unrelated to support image-caption pre-training.

If the caption includes extra article context but the main visual content
    still matches the image, label CONSISTENT.

<Caption>
{caption}
</Caption>

Return JSON only:
{
  "label": "consistent",
  "reason": "Brief reason"
}
```

**VLM Image Quality Filtering**

```
You are labeling whether a medical image is visually usable for medical
    vision-language pre-training.

Use the image and, when helpful, the caption.

Label HIGH_QUALITY if the image is readable and contains enough visual
    information for medical interpretation or medical visual learning,
    even if it is a multi-panel figure.

Label LOW_QUALITY if the image is corrupted, blank, unreadable, extremely
    low resolution, severely compressed, mostly text-only, dominated by
    non-visual tables/charts, cropped so heavily that the described
    finding is not visible, or otherwise diagnostically/visually
    uninformative.

Do not reject an image only because it is complex or multi-panel. Reject
    only when the visual signal is poor or not medically useful.

<Caption>
{caption}
</Caption>

Return JSON only:
{
  "label": "high_quality",
  "reason": "Brief reason"
}
```

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction summarize the main contributions: PubMedVision-Plus, PubMedVision-RL, the pre-training plus RL recipe, and the HuatuoGPT-Vision-2 models. The claims are supported by the method description, benchmark results, ablations, MedMM-ExpertRubric evaluation, and physician preference study in Sections 3 and 4.2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses limitations in Appendix A.1, including English PubMed coverage, differences from real clinical workflows, reliance on GPT-4.1 for scalable rubric scoring, and the need for rigorous clinical validation before deployment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper does not present new theoretical results or formal proofs. The equations in Sections 2.1 and 3 define standard training objectives and reward functions used by the empirical method.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Sections 3 and 4.2 provide the training recipe, data construction pipeline, hyperparameters, benchmarks, and baseline groups. The appendix further provides filtering details, benchmark details, data synthesis prompts, rubric evaluation details, and physician preference protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: Data, code, and models are being released progressively at `https://github.com/FreedomIntelligence/HuatuoGPT-Vision-2`, and the repository will be updated as each asset becomes available. The paper also documents the construction pipeline, prompts, and evaluation protocols in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Section 4.2 specifies the base models, pre-training, RL, and SFT hyperparameters, compute setup, benchmark suite, and baseline categories. Appendix C, Appendix I, Appendix F, and Appendix G provide additional evaluation and data-construction details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: Appendix H.1 reports 95% confidence intervals from 10,000 example-level bootstrap resamples for the main benchmarks and MedMM-ExpertRubric. Appendix G reports confidence intervals for physician win rates, overlap agreement, and Cohen's $\kappa$. Repeated full training runs are too expensive, so these intervals do not capture variation across training runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Section 4.2 states that experiments are conducted on two 8-GPU NVIDIA H200 machines and reports the main training settings. Each H200 GPU has 141GB memory; pre-training the 8B model took approximately 576 GPU-hours, and RL training took approximately 240 GPU-hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses publicly available PubMed literature rather than private clinical records, reports limitations and intended use, and compensates physician reviewers at a fixed hourly rate. The paper also states that the model is intended for research and should not be used for direct clinical diagnosis without appropriate validation and oversight.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Appendix A.3 discusses the positive impact of lowering barriers to medical VLM research through open datasets and models. It also notes risks from inappropriate clinical use and emphasizes that outputs should assist, not replace, professional medical judgment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes].

    Justification: The paper describes data filtering steps that remove low-quality, mismatched, and non-medical visual content, and it uses publicly available PubMed literature rather than private clinical data. Appendix A.3 and Appendix A.1 state responsible-use constraints for medical VLMs and caution against direct clinical diagnosis without validation and oversight.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No].

    Justification: We cite the original sources of the datasets, models, and benchmarks used in this work, but do not provide a complete asset-by-asset account of their licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper introduces PubMedVision-Plus, PubMedVision-Plus-Core, PubMedVision-RL, MedMM-ExpertRubric, and HuatuoGPT-Vision-2, and documents their construction in the main paper and appendix. The appendix provides data examples, filtering pipelines, synthesis prompts, evaluation protocols, and physician-review details.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes].

    Justification: Appendix F and Appendix G include the physician rubric-filtering and pairwise-preference instructions, reviewer assignment, overlap statistics, and compensation details. The physicians are compensated at a fixed hourly rate of 200 RMB independent of model identity, preference choices, or evaluation outcomes.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No].

    Justification: The physician review task involves professional assessment of model outputs rather than collection of private patient information.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: LLMs and VLMs are core components of the data construction and evaluation pipeline. The paper describes GPT-5.4-based filtering and data synthesis, GPT-4.1 rubric judging, and compact LLM/VLM classifiers and scorers in Sections 3 and the appendix.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.