INFIMED2: A GENERALIST MEDICAL MULTIMODAL FOUNDATION MODEL FROM CONTEXTUAL EVIDENCE AND STABILITY-AWARE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent medical multimodal models have benefited from larger corpora, broader modality coverage, and stronger reasoning-oriented training, yet effective data design across continued pretraining (CPT) and post-training remains challenging. Medical sources vary substantially in structure, granularity, and information density, and their utility shifts as training progresses from broad knowledge acquisition to late-stage consolidation. Meanwhile, post-training is often dominated by short-form visual question answering, providing limited supervision for informative and answer-consistent explanations. We introduce InfiMed2, a family of 4B and 27B generalist medical multimodal foundation models built around stage-aware data design. We curate a 55.68B-token corpus that combines broad clinical knowledge with context-rich biomedical visual evidence through source-specific processing. Our CPT pipeline first adapts the vision encoder, then builds broad medical knowledge, and finally transitions to an evidence-focused data mixture during learning-rate decay. For supervised fine-tuning (SFT), we regenerate visual question-answering responses using answer stability, answer-masked reconstruction, and correctness-constrained selection to produce more informative and answer-consistent supervision. The 4B model is further optimized with reinforcement learning with verifiable rewards (RLVR). Across five medical multimodal benchmarks, InfiMed2-4B achieves 66.73% mean accuracy after RLVR, surpassing the larger Qwen3.5-9B, while InfiMed2-27B reaches 73.72%, the highest among the evaluated open-weight models.

1 INTRODUCTION

Medical multimodal large language models (MLLMs) seek to transfer general multimodal capabilities into a domain where specialized visual evidence must be interpreted together with medical knowledge and clinically appropriate language. Recent progress has followed three complementary directions: expanding domain-specific corpora, adapting visual and multimodal representations, and strengthening instruction following and reasoning through post-training (Li et al., 2023; Chen et al., 2024; Sellergren et al., 2025; Xu et al., 2025; Jiang et al., 2025; Shi et al., 2026). These advances show that capable medical MLLMs depend on both knowledge acquisition and behavioral alignment. They also expose a shared data-design challenge across the two phases: continued pretraining (CPT) must organize heterogeneous evidence into effective learning signals, while post-training must convert the resulting representations into reliable and informative responses.

CPT must reconcile medical sources that differ substantially in structure, granularity, and information density. Existing medical corpora often concentrate on a single source type, while large-scale automatic collection can leave duplicated, noisy, or weakly connected content. Isolated image–caption pairs provide concentrated alignment signals but omit much of the surrounding evidence, whereas long-form and interleaved documents offer richer context at lower and less uniform information density. Using one static mixture throughout training asks the same distribution to support both broad knowledge acquisition and late-stage consolidation. Early updates benefit from diverse coverage, while the final low-learning-rate regime is more sensitive to the quality and task relevance of the remaining evidence. The visual encoder presents a related tension: keeping it fixed limits medical visual adaptation, but unrestricted joint optimization from the outset can disturb previously

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

aligned representations. Prior work has extensively explored data scale and curriculum order, yet the interaction among data mixture, trainable components, and optimization stage remains less studied.

Post-training faces a different limitation. Medical VQA annotations are commonly designed to evaluate answer correctness and therefore often contain only an entity, category, or short judgment (Zhang et al., 2023b). Directly fitting these targets provides little supervision for identifying relevant visual findings or connecting them to medical knowledge. Replacing a short answer with one teacher-generated explanation is also insufficient because a longer response may still be weakly grounded or inconsistent with the verified answer. Moreover, teacher reliability varies across examples. Effective post-training data must therefore improve explanatory content while accounting for the stability and correctness of the generated supervision.

We introduce InfiMed2 to address these data-design problems within a unified training pipeline. CPT begins with vision-encoder adaptation, proceeds to warmup–stable training on a 55.68B-token corpus of complementary medical and general multimodal evidence, and concludes with learning-rate decay on a compact mixture enriched with high-quality medical evidence. For post-training, we estimate answer stability from repeated teacher generations and use it to route response construction. Stable examples undergo answer-masked reconstruction and correctness-constrained selection, less stable examples are assigned to a stronger regeneration model, and low-agreement examples are removed. The retained responses are combined with general multimodal, medical visual, and reasoning supervision for supervised fine-tuning (SFT). The selected 4B model subsequently undergoes reinforcement learning with verifiable rewards (RLVR) using a difficulty-aware mixture derived from pass@8 outcomes.

Under a common evaluation protocol on five medical multimodal benchmarks, InfiMed2-4B reaches 66.73% mean accuracy and surpasses the larger Qwen3.5-9B. InfiMed2-27B achieves 73.72%, the highest mean accuracy among the evaluated open-weight models. Together, these results indicate that coordinated data design across continued pretraining and post-training can support strong medical multimodal performance without depending solely on model scaling.

Our contributions are summarized as follows:

- We construct a 55.68B-token CPT corpus from complementary medical sources using source-specific processing. Our training design separates vision-encoder adaptation, broad medical knowledge acquisition, and learning-rate decay on a rebalanced evidence mixture.
- We develop stability-aware response regeneration for medical post-training. Repeated generations provide an empirical measure of answer stability. This signal guides reconstruction, correctness-constrained selection, selective model escalation, and filtering to produce informative supervision consistent with verified answers.
- We present InfiMed2, a family of 4B and 27B generalist medical multimodal foundation models. The 4B model outperforms a larger general-purpose baseline, and the 27B model attains the strongest mean performance among the evaluated open-weight models.

2 Related Work

2.1 Medical Multimodal Foundation Models

Medical multimodal large language models commonly adapt general-purpose vision-language backbones with domain-specific image–text and instruction data. Early systems such as LLaVA-Med (Li et al., 2023) and HuatuoGPT-Vision (Chen et al., 2024) emphasize biomedical alignment and instruction following, while MedGemma (Sellergren et al., 2025) adapts both visual and language components. Recent generalist systems broaden data, modalities, and reasoning supervision, including Lingshu (Xu et al., 2025), Hulu-Med (Jiang et al., 2025), and MedXiaoHe (Shi et al., 2026). Despite this progress, source-aware evidence curation and verification of synthetic explanations remain less systematically studied. InfiMed2 addresses these data-design problems across continued pretraining and post-training.

2.2 Multimodal Continued Pretraining and Medical Data Curation

Multimodal continued pretraining depends on both coverage and the relation between images and textual evidence. PMC-15M (Zhang et al., 2023a) and BIOMEDICA (Lozano et al., 2025) provide large biomedical figure collections, but literature-derived data often reduce evidence to isolated figure–caption pairs and remain vulnerable to extraction noise. PMC-InterCPT (Zhu et al., 2026) instead reconstructs associations among figures, captions, and their related context. Source diversity, source-specific failure modes, and stage-dependent data value nevertheless remain underexplored. InfiMed2 addresses these limitations by combining complementary medical sources through source-aware processing.

2.3 Medical Post-Training and Reliable Synthetic Supervision

Medical post-training uses supervised question answering, instruction following, and reasoning data, sometimes followed by outcome-based reinforcement learning. PMC-VQA (Zhang et al., 2023b) supplies biomedical VQA supervision, while ReasonMed (Sun et al., 2025) and recent medical models use synthetic reasoning to strengthen complex problem solving (Xu et al., 2025; Shi et al., 2026). Generated explanations, however, may be weakly grounded or inconsistent with verified answers. Self-consistency (Wang et al., 2022) and predictive uncertainty (Kuhn et al., 2023) provide useful reliability signals. Our regeneration pipeline combines these signals with reference-answer constraints to curate explanatory supervision.

3 Data Curation

Figure 1 provides an overview of how source-specific CPT curation and post-training curation supply the successive stages of the InfiMed2 training pipeline.

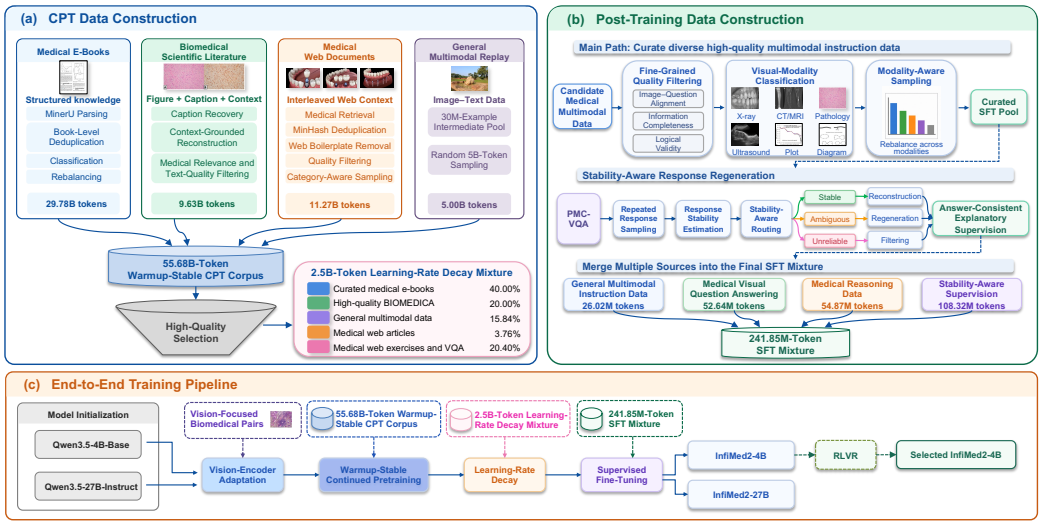

Figure 1: Overview of the InfiMed2 data construction and training pipeline. (a) Source-specific processing forms the warmup–stable CPT corpus and a compact, high-quality mixture for learning-rate decay. (b) Post-training combines fine-grained quality filtering, stability-aware response regeneration, and visual-modality-aware sampling. (c) The resulting data support vision-encoder adaptation, CPT, SFT, and subsequent RLVR for the selected 4B model.

3.1 CPT Data Curation

Curation objective and source complementarity. We construct the CPT corpus by combining complementary sources of medical knowledge and multimodal evidence, rather than applying a

uniform curation strategy to heterogeneous training data: **(1)** medical e-books provide systematic, long-form exposition; **(2)** scientific articles connect biomedical figures to peer-reviewed textual evidence; **(3)** web documents broaden coverage of practical and less formal medical content; and **(4)** general multimodal data serves as replay for capabilities acquired by the backbone. Because each source exhibits distinct types of noise and distributional bias, we develop source-specific cleaning and sampling procedures for each corpus. After source-specific processing, we obtain a 55.68B-token corpus that is used in its entirety for warmup–stable CPT. It comprises 29.78B tokens from medical e-books, 9.63B from biomedical scientific literature, 11.27B from medical web documents, and 5B from general multimodal data.

Medical e-books. We collect approximately 156k medical e-books and convert them into structured Markdown using MinerU (Wang et al., 2024). We perform book-level deduplication and remove parsing failures, empty or repetitive content, and cover-only records. Because the cleaned collection remains imbalanced, we classify each book by medical subject and genre using sampled pages and bibliographic metadata. The subject taxonomy follows a medical-first principle, and the genre classifier selects among medical monograph, medical popular science, medical reference/atlas/quick-reference work, medical textbook, medical research literature/report, medical supplementary teaching material, and medical guideline/standard. We jointly use these labels to increase clinically oriented disciplines and high-value reference material while reducing overrepresented foundational, popular-science, and supplementary teaching content. The final subset contains 60,433 books and 22.41B text tokens, or 29.78B tokens after including visual tokens. Detailed cleaning and sampling statistics are provided in Appendix A.1.

Biomedical scientific literature. For biomedical scientific literature, we use 10.11M examples totaling 9.63B tokens from PMC-InterCPT (Zhu et al., 2026), which reconstructs context-grounded records from BIOMEDICA (Lozano et al., 2025). The pipeline recovers provenance-linked captions, associates each figure with the context that explicitly references it, and places figures sharing the same context in one interleaved sequence. It then repairs incoherent context, applies medical-relevance and text-quality filtering, and excludes PMC-VQA test examples (Zhang et al., 2023b). Unlike isolated image–caption pairs, the resulting corpus preserves the relationships among scientific figures, captions, and figure-linked context, providing supervision for integrating evidence across multiple related elements. Additional construction details are provided in Appendix A.2.

Medical web data. Medical web data complements books and scientific articles with broader practical and long-tail medical content. We train a Qwen3-1.7B (Yang et al., 2025) medical-relevance classifier to retrieve a 32.99B-token candidate pool from OmniCorpus-CC (Li et al., 2025), OBELICS (Laurençon et al., 2023), and MINT-1T (Awadalla et al., 2024), followed by MinHash near-duplicate removal (Broder, 1997). A dedicated model removes navigation, header, and footer text, and we restore interleaving by downloading images referenced by source URLs and rejoining them with their records. We discard documents assigned the lowest score by a three-level quality classifier, then use image categories for balanced resampling. The resulting medical interleaved corpus contains 11.27B tokens.

General multimodal replay. Medical specialization can weaken broad visual and linguistic capabilities if training excludes general-domain evidence. We therefore include 5B tokens from LLaVA-OneVision-1.5-Mid-Training-85M (An et al., 2025) as general multimodal replay during warmup–stable CPT. The source dataset contains 85M curated and concept-balanced examples, and its associated models demonstrate strong performance across a broad benchmark suite. We convert a 30M-example intermediate pool containing 19.89B tokens and randomly sample the 5B-token training subset. We keep this component smaller than the combined medical sources so that it regularizes domain adaptation without diluting the medical focus of the mixture.

From these curated pools, we derive two stage-specific data configurations tailored to the first and final stages of CPT.

Vision-encoder adaptation subset. We construct a visually concentrated 1B-token subset from the curated PMC-InterCPT corpus. Each eligible record is converted into a single-image, single-caption example by retaining its first valid image and the corresponding caption. Deterministic token-budgeted sampling yields 1.71M image–caption pairs, with visual tokens accounting for 73.7% of the subset. This subset provides a vision-focused view of the subsequent corpus rather than introducing an additional data source.

Learning-rate decay mixture. For late-stage consolidation, we assemble a compact, high-quality 2.50B-token mixture from the curated data pools. Medical e-books account for 40% of the tokens, BIOMEDICA for 20%, and general multimodal data for 15.84%. The remaining 24.16% comes from medical web data, including 3.76% medical articles and 20.40% medical exercises and selected VQA examples. This mixture increases the density of systematic medical knowledge and application-oriented supervision while retaining general multimodal replay.

3.2 Post-Training Data Curation

We collect post-training data from heterogeneous medical and general multimodal sources. Directly combining these sources introduces three practical issues. First, some samples contain semantic inconsistencies that metadata cannot reliably identify, including mismatched image modalities, missing anatomical evidence, incomplete problem conditions, or responses not properly grounded in the visual input. Second, many medical VQA datasets provide only short-form answers, which are sufficient for evaluating answer correctness but provide limited supervision for learning informative multimodal responses. Third, the resulting corpus is highly imbalanced in visual content.

We therefore construct the post-training data through a sequential pipeline. We first perform fine-grained sample-level filtering to remove invalid grounded examples. For retained visual question answering (VQA) samples, we then regenerate responses to increase their information density while preserving answer correctness. Finally, we construct the training mixture according to visual content rather than raw dataset size.

3.2.1 Fine-Grained Quality Filtering

We first assess the validity of each candidate sample by jointly considering its visual and textual components. We define a fine-grained annotation scheme along three dimensions: image–question alignment, information completeness, and logical validity. The annotation covers common failure cases such as image-modality mismatch, missing anatomical regions, and inconsistent question–answer pairs.

Based on this annotation scheme, we construct a quality-control dataset and train a 2B parameter evaluator to assess candidate samples. We use a three-level score, where Score 0 indicates a critical defect, Score 1 denotes a limited but non-fatal issue, and Score 2 denotes a valid sample. The evaluator removes Score 0 samples before response construction and mixture formation. The complete quality assessment protocol is provided in Appendix B.1 and Table 5.

3.2.2 Stability-Aware Response Regeneration

After filtering, we refine the response supervision of retained medical VQA data. Existing medical VQA datasets often provide only an entity, category label, or short judgment and therefore provide limited supervision for integrating visual evidence with medical knowledge. Our regeneration procedure produces informative explanations while preserving the reference answer.

Multi-sample response generation and stability estimation. Let $x_i = (I_i, q_i, a_i^*)$ denote a medical VQA instance, where I_i denotes the medical image, q_i the question, and a_i^* the reference answer. We first employ Qwen3-VL-32B-Instruct (Bai et al., 2025) parameterized by θ_T as a teacher model to independently sample $K = 8$ candidate responses:

$$r_i^{(k)} \sim p_{\theta_T}(r \mid I_i, q_i), \qquad k = 1, \ldots, K. \tag{1}$$

Repeated stochastic sampling estimates answer stability through agreement among independently generated reasoning paths, following the intuition of self-consistency decoding (Wang et al., 2022).

Let $\hat{a}_i^{(k)}$ be the final answer extracted from $r_i^{(k)}$, and let $\mathcal{M}$ denote the dataset-specific answer-matching function. We count the correct responses and normalize this count as

$$C_i = \sum_{k=1}^{K} \mathbb{I}\Big[\mathcal{M}(\hat{a}_i^{(k)}, a_i^*) = 1\Big], \qquad S_i = \frac{C_i}{K}. \tag{2}$$

Here, $\mathbb{I}[\cdot]$ is the indicator function, C_i is the number of correct responses, and S_i measures how consistently the teacher recovers the reference answer. As K is fixed, we use C_i for routing.

Answer-masked reconstruction and candidate selection. For high-stability samples with $6 \leq C_i \leq 8$, we retain the explanatory content of correct teacher responses and mask their explicit final answers, yielding $\tilde{r}_i^{(k)} = \mathrm{MaskAnswer}(r_i^{(k)}, \hat{a}_i^{(k)})$. Qwen3-VL-4B-Instruct (Bai et al., 2025), parameterized by θ_R, then reconstructs the response from the image, question, and masked explanation:

$$\bar{r}_i^{(k)} \sim p_{\theta_R}(r \mid I_i, q_i, \tilde{r}_i^{(k)}). \tag{3}$$

Here, $\bar{r}_i^{(k)}$ denotes the reconstructed response. We form $\mathcal{R}_i^+$ from reconstructions whose extracted answers match a_i^* and whose originating teacher responses are also correct.

Entropy-based quantities provide a natural measure of uncertainty in probabilistic language generation (Kuhn et al., 2023). In our setting, reference-answer consistency is first enforced as a hard constraint, after which entropy is used only to rank the remaining valid candidates. For each retained reconstruction, we compute the mean answer-token entropy

$$\mathcal{H}(\bar{r}_i^{(k)}) = \frac{1}{T_k} \sum_{t=1}^{T_k} \left[-\sum_{v \in \mathcal{V}} p_t(v) \log p_t(v) \right]. \tag{4}$$

Here, T_k is the answer length, $\mathcal{V}$ is the output vocabulary, and $p_t(v)$ is the reconstruction model's probability of token v at position t. Among the valid candidates, we select

$$r_i^* = \arg\min_{r \in \mathcal{R}_i^+} \mathcal{H}(r). \tag{5}$$

If $\mathcal{R}_i^+ = \emptyset$, we discard the sample.

Escalated regeneration and routing. For intermediate-stability samples with $3 \leq C_i \leq 5$, Gemini 3.1 Pro (Google DeepMind, 2026) generates a new response conditioned on (I_i, q_i, a_i^*). Samples with $C_i < 3$ are discarded. This routing reserves stronger-model regeneration for ambiguous cases and avoids synthetic supervision when answer stability is too low.

Applied to 176,948 questions, the procedure retains 121,544 with at least one answer-consistent candidate and exports 106,183 final examples. Among 704,046 answer-masked reconstructions, 662,503 (94.10%) recover the reference answer, indicating that masking usually preserves answer-relevant explanatory content while routing removes unrecoverable cases.

3.2.3 Visual-Modality-Aware Mixture Construction

After filtering and regeneration, we classify samples into 27 visual categories spanning medical imaging and broader biomedical content, such as radiology, pathology, plots, and diagrams, and train a classifier to annotate the curated samples. Category-level sampling reduces domination by frequent modalities while preserving clinically relevant long-tail content. The resulting SFT mixture contains 296,725 examples and 241.85M tokens, including 106,183 stability-aware examples totaling 108.32M tokens. Appendix B.2 provides the complete composition.

4 Model Training

The 4B and 27B training runs are initialized from Qwen3.5-4B-Base and the instruction-tuned Qwen3.5-27B checkpoint, respectively. The training pipeline consists of CPT followed by SFT. CPT is organized into three consecutive stages: vision-encoder adaptation, large-scale warmup–stable training, and learning-rate decay. SFT subsequently converts the acquired medical representations into instruction-following and reasoning capabilities using the curated post-training supervision. The 30B-CPT 4B variant additionally undergoes RLVR after SFT.

4.1 Continued Pretraining

Vision-encoder adaptation. We first train the vision encoder and multimodal projector for one epoch on the 1B-token image–caption subset while keeping the language model frozen. This stage establishes medical image–text alignment before training on context-rich interleaved records.

Warmup–stable continued pretraining. We then train all model components for one epoch on the 55.68B-token mixture. This stage provides broad exposure to complementary medical sources while maintaining general multimodal capability.

Learning-rate decay. Finally, we train all model components for one epoch on the high-quality 2.50B-token mixture. Concentrating curated medical and multimodal evidence near convergence strengthens domain adaptation, while general multimodal replay mitigates excessive specialization.

All CPT stages use AdamW with a global batch size of 256. The maximum sequence length is 8,192 for vision-encoder adaptation and 12,288 for the two subsequent stages. Vision adaptation uses 3% warmup to 2×10^{-6} followed by cosine decay to 2×10^{-7}. Warmup–stable CPT reaches 1×10^{-5} after 3% warmup and holds it constant, while the final stage decays it from 1×10^{-5} to 2×10^{-6}.

4.2 Supervised Fine-Tuning

Starting from the post-decay checkpoint, we train all model components for five epochs on the 241.85M-token mixture constructed in Section 3.2. It combines general multimodal instructions, medical VQA, medical reasoning supervision, and stability-aware responses. We use AdamW with a global batch size of 64 and a maximum sequence length of 9,000. After 10% warmup, the learning rates peak at 2×10^{-6} for the language model and projector and 1×10^{-6} for the vision encoder, followed by cosine decay to zero.

4.3 Reinforcement Learning with Verifiable Rewards

Starting from the SFT checkpoint, we perform RLVR on a 21,529-example mixture comprising 19,529 medical VQA examples and 2,000 text-only USMLE questions drawn from public resources and internally curated data (Zhang et al., 2023b; Huang et al., 2026; Li et al., 2026; Hu et al., 2024; Mei et al., 2022; Jin et al., 2021). We train for three epochs with a learning rate of 1×10^{-6} and a global batch size of 128. For each prompt, the policy samples eight rollouts at temperature 1. The reward combines format compliance and answer accuracy with a 1:9 weight ratio.

We estimate difficulty with **pass@8** and prioritize examples with 3–5 correct rollouts, where reward remains attainable but improvement is still possible. Near-unsolved and saturated examples are downsampled, while a smaller easy subset is retained to reduce forgetting after SFT.

Full optimization and implementation details are provided in Appendix C.

5 Experiments

5.1 Experimental Setup

Evaluation benchmarks. We evaluate medical multimodal understanding on five benchmarks: the Health and Medicine test subset of MMMU (Yue et al., 2024), the corresponding 10-option subset of MMMU-Pro (Yue et al., 2025), MedXpertQA-MM (Zuo et al., 2025), PMC-VQA-clean (Zhang et al., 2023b), and OmniMedVQA (Hu et al., 2024). We report accuracy on each benchmark and the unweighted mean across the five benchmarks.

Baselines and evaluation protocol. We compare against general-purpose and medical multimodal models, including the Qwen3.5 backbones and medical models such as MedGemma (Sellergren et al., 2025), Lingshu (Xu et al., 2025), and Hulu-Med (Jiang et al., 2025). We also train SFT-only controls from Qwen3.5-4B-Base and the instruction-tuned Qwen3.5-27B checkpoint, using the same instruction data as their CPT-initialized counterparts. Within each benchmark, all models are evaluated on the same questions with the same prompts and scoring procedure. All scores come from our evaluation rather than heterogeneous published leaderboards.

5.2 Main Results

Table 1 compares the selected InfiMed2 models with general-purpose and medical baselines. After RLVR, InfiMed2-4B reaches 66.73% average accuracy, outperforming Qwen3.5-4B by 6.54 points

Table 1: Medical multimodal accuracy (%) under a common evaluation protocol. Selected InfiMed2 models are shaded in gray. Bold and underlined entries denote the best and second-best scores.

Model	MMMU Med.	MMMU-Pro Med.-10	MedXQA	PMC-VQA clean	OmniMed VQA	Avg.
General-purpose models						
Gemini-3-Pro	81.84	**74.13**	**74.85**	**70.40**	84.55	**77.15**
Gemini-3-Flash	81.51	70.63	68.00	<u>69.75</u>	84.45	<u>74.87</u>
Claude-Opus-4.7	<u>82.47</u>	70.63	69.40	67.15	81.60	74.25
GPT-5	**83.39**	70.90	<u>71.70</u>	67.30	76.40	73.94
GLM-5V-Turbo	73.91	64.69	53.25	66.55	81.95	68.07
Inkling	73.17	60.49	56.00	57.80	69.05	63.30
MiMo-v2.5	73.00	56.65	50.20	53.50	70.55	60.78
Qwen3.5-4B	68.66	54.89	34.75	60.15	82.50	60.19
Qwen3.5-9B	76.31	58.39	40.35	62.60	85.91	64.71
Qwen3.5-27B	74.70	60.48	47.05	64.30	90.60	67.43
Gemma 4-31B	79.16	66.43	54.55	66.00	80.15	69.26
Medical-domain models						
MedGemma-1.5-4B-IT	47.26	30.42	28.40	47.75	69.38	44.64
Lingshu-32B	62.30	41.26	30.90	57.90	83.40	55.15
Hulu-Med-32B	60.84	40.55	34.10	64.55	84.90	56.98
Selected InfiMed2 models (ours)						
InfiMed2-4B	72.43	58.39	46.25	65.80	<u>90.79</u>	66.73
InfiMed2-27B	77.85	<u>70.98</u>	59.40	67.50	**92.93**	73.72

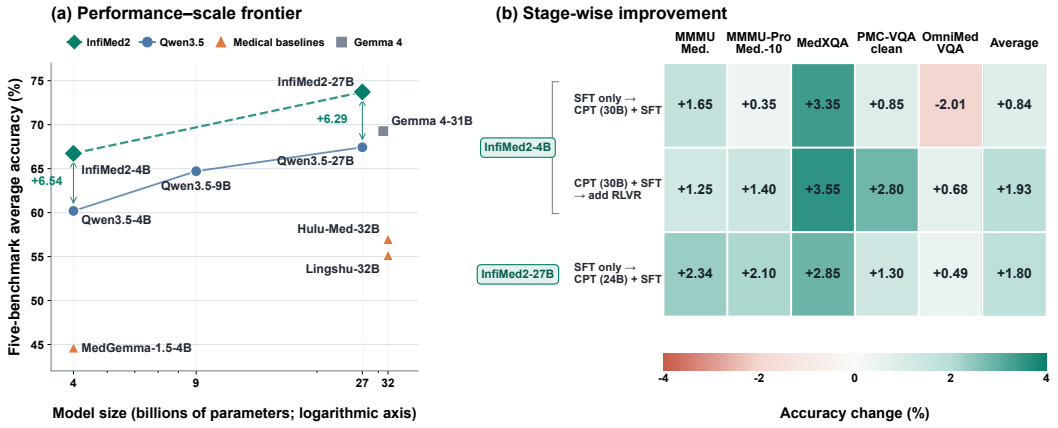

Figure 2: Parameter efficiency and stage-wise gains. (a) Five-benchmark average accuracy versus model size for open-weight models with reported parameter counts. (b) Benchmark-wise score changes after adding CPT or RLVR to the corresponding matched configuration.

and the larger Qwen3.5-9B by 2.02 points. InfiMed2-27B achieves the highest average among the evaluated open-weight models at 73.72%, trailing GPT-5 and Claude-Opus-4.7 by only 0.22 and 0.53 points, respectively. Figure 2(a) further shows that the InfiMed2 scaling curve remains above the corresponding Qwen3.5 curve at both model sizes.

5.3 Ablation Studies

5.3.1 CPT Checkpoint Selection and RLVR

Under matched SFT, average performance peaks after 30B warmup–stable CPT for the 4B model and 24B for the 27B model. RLVR then improves all five benchmarks for the selected 4B checkpoint, with the largest gains on MedXpertQA-MM and PMC-VQA-clean. Figure 2(b) summarizes these matched stage-wise gains, with parenthesized values denoting warmup–stable CPT tokens. Extend-

Table 2: CPT checkpoint selection and RLVR ablation. CPT columns report token budgets.

Parameters	Vision-encoder adaptation	Warmup–stable CPT	Learning-rate decay	Post-training	MMMU Med.	MMMU-Pro Med.-10	PMC-VQA MedXQA	PMC-VQA clean	OmniMedVQA	Avg.
4B	–	–	–	SFT	69.53	56.64	39.35	62.15	**92.12**	63.96
	1B	30B	2.5B	SFT	71.18	56.99	42.70	63.00	90.11	64.80
	1B	55B	2.5B	SFT	70.55	55.94	43.05	63.90	90.04	64.69
	1B	30B	2.5B	SFT+RLVR	**72.43**	**58.39**	**46.25**	**65.80**	90.79	**66.73**
27B	–	–	–	SFT	75.51	68.88	56.55	66.20	92.44	71.92
	1B	8B	2.5B	SFT	78.88	66.78	57.45	66.20	92.69	72.40
	1B	24B	2.5B	SFT	77.85	**70.98**	59.40	67.50	**92.93**	**73.72**
	1B	55B	2.5B	SFT	**79.97**	67.48	**60.05**	**68.60**	91.51	73.51

Table 3: Matched lightweight-SFT ablation of response targets.

SFT target	MMMU Med.	MMMU-Pro Med.-10	PMC-VQA MedXQA	PMC-VQA clean	OmniMedVQA	Avg.
Answer only	58.56	40.21	31.25	**65.70**	**83.19**	55.78
Random generated response	68.04	**52.10**	34.85	59.25	75.60	57.97
Stability-aware selected response	**69.81**	51.75	**35.00**	59.20	77.00	**58.55**

ing the 27B run beyond its selected checkpoint improves several tasks but reduces others. This non-monotonic pattern suggests that additional domain exposure can sharpen capabilities aligned with the CPT mixture without preserving the best downstream balance after a fixed amount of SFT.

5.3.2 Stability-Aware Response Regeneration

Using the same 4B checkpoint and 20k PMC-VQA examples, this ablation changes only the target: the original short answer, a random generated response, or stability-aware selection. Generated explanations improve reasoning-oriented benchmarks, and stability-aware selection achieves the best average and the best scores on MMMU-Medical and MedXpertQA-MM. Answer-only targets remain stronger on direct-answer VQA, revealing a target-format trade-off. The advantage over random selection shows that response expansion alone cannot explain the gain.

5.4 Discussion

The experiments show that neither longer CPT nor more verbose supervision is uniformly beneficial. Under matched SFT, average performance peaks at 30B warmup–stable tokens for the 4B model and 24B for the 27B model, while later checkpoints trade gains on some tasks for losses on others. Additional domain exposure may sharpen capabilities emphasized by the CPT mixture while shifting the balance inherited from the backbone. Checkpoint selection should therefore follow the downstream capability profile rather than token count alone. Likewise, generated explanations improve reasoning-oriented benchmarks, but answer-only targets remain competitive on direct-answer VQA. Stability-aware selection gives the best overall balance, suggesting that post-training should retain some concise targets or adapt response format to the task. Future work can extend this evaluation to open-ended clinical communication and expert assessment of explanation quality.

6 Conclusion

We present InfiMed2, a family of 4B and 27B generalist medical multimodal foundation models built through coordinated data design across continued pretraining and post-training. Source-specific curation and stage-specific mixtures preserve complementary medical evidence and adapt its role from vision alignment to late-stage consolidation. Post-training couples fine-grained quality control with stability-aware response regeneration. Across five benchmarks, InfiMed2-4B surpasses larger open-weight baselines, while InfiMed2-27B achieves the strongest average among the evaluated open-weight models. The ablations show that matching data construction to each training stage offers a practical route to stronger medical multimodal capability.

AI USE STATEMENT

We used generative AI tools to assist with literature search and organization, code development, language polishing of early drafts, and manuscript formatting. Generative models were also used within the post-training data-construction pipeline. The authors reviewed and verified all AI-assisted literature records, code, generated training examples, analyses, and manuscript text, and take full responsibility for the final content.

ETHICS STATEMENT

This study does not collect new data from human participants and does not involve prospective clinical intervention. The training and evaluation resources are publicly available or obtained under their applicable access conditions, and any release will respect source licenses, privacy requirements, and redistribution restrictions. Although our curation pipeline removes low-quality records and obvious mismatches, heterogeneous medical corpora may retain factual errors, demographic or geographic biases, and sensitive content. InfiMed2 may hallucinate, reproduce dataset biases, or provide unsafe medical advice. It is intended for research and is not approved for unverified clinical use.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe the data pipeline, training stages, evaluation protocol, and ablations. Appendix A provides additional corpus statistics and filtering details, while Appendix C records the optimization settings. Upon acceptance, we plan to release the trained models, selected training data and metadata whose licenses and privacy conditions permit redistribution. For restricted or third-party resources, we will provide provenance, preprocessing descriptions, and reconstruction scripts rather than redistributing the underlying data.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *Advances in Neural Information Processing Systems*, 37:36805–36828, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024.

Google DeepMind. Gemini 3.1 pro model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Xiaoke Huang, Ningsen Wang, Hui Liu, Xianfeng Tang, and Yuyin Zhou. Synthesizing high-quality visual question answering from medical documents with generator-verifier lmms. In *International Conference on Learning Representations*, volume 2026, pp. 109770–109806, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

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

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: A unified multimodal corpus of 10 billion-level images interleaved with text. In *International Conference on Learning Representations*, volume 2025, pp. 13647–13689, 2025.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23177–23185, 2026.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19724–19735. IEEE, 2025.

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

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

Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, et al. Medxiaohe: A comprehensive recipe for building medical mllms. *arXiv preprint arXiv:2602.12705*, 2026.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, et al. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26457–26478, 2025.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

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

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

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 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)*, pp. 15134–15186, 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Guanghao Zhu, Zeyu Liu, Zhitian Hou, Pengkai Wang, Zhijie Sang, Yang Yu, Minheng Ni, Wenjun Wang, Yanggan Gu, Shuo Cai, Congkai Xie, Jianmin Wu, and Hongxia Yang. Beyond captions: Context-grounded reconstruction for biomedical multimodal continued pretraining. *arXiv preprint arXiv:2606.01049*, 2026.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

A Additional Data Curation Details

A.1 Medical E-Book Processing and Sampling

We collect 156,498 medical e-books across six source formats, of which 156,051 are successfully converted into structured Markdown with MinerU (Wang et al., 2024), corresponding to a 99.7% file-level conversion rate. Before sampling, we construct a content signature for each parsed book by concatenating its ten longest paragraphs, each containing at least 50 characters. Global exact matching identifies 3,025 duplicate books, or 1.93% of the inputs to this stage. We further remove parsing failures, empty outputs, samples dominated by repeated characters, and records containing only a cover page. After cleaning and deduplication, 152,987 books remain, containing 45.58B text tokens.

For classification, the model receives sampled pages together with a small set of bibliographic metadata and predicts a constrained major subject, an open-ended fine-grained subject, and one of seven medical genres. The subject taxonomy follows a medical-first rule derived from the R categories of the Chinese Library Classification. Content concerning human structure or function, health, disease, diagnosis, treatment, or clinical practice is assigned to the most specific applicable medical category. Biology is reserved for content centered on non-human biological systems, while the non-medical category is used only when the material is unrelated to medicine or biology.

The predicted subject and genre labels jointly determine the sampling mixture. Along the subject axis, rebalancing reduces the text-token share of basic medicine from 21.3% to 9.2% and increases the aggregate share of clinical disciplines from approximately 55% to 75%. Along the genre axis, medical textbooks increase from 13.6% to 27.7%, medical reference/atlas/quick-reference works increase from 11.1% to 17.9%, and medical popular science decreases from 5.6% to 1.3%. We retain all medical textbooks and medical guidelines/standards, 83.9% of medical reference/atlas/quick-reference works, 11.4% of medical popular science, and 7.1% of medical supplementary teaching materials. The sampled corpus contains 60,433 books and 22.41B text tokens, corresponding to 29.78B total tokens after visual tokens are included.

A.2 Biomedical Scientific Literature

PMC-InterCPT (Zhu et al., 2026) first recovers provenance-linked captions from PMC XML and normalizes the source text while preserving figure-reference anchors. It constructs interleaved sequences by attaching figure-linked context only to figures that the context explicitly references. Figures referenced by the same context remain in a shared sequence rather than being paired with duplicated context. A coherence-repair step removes discontinuous context and prunes images that are no longer supported by the retained text. Separate classifiers then filter records by medical relevance and textual quality, and examples overlapping the PMC-VQA test set are excluded (Zhang et al., 2023b). Evidence-aware allocation assigns 45% of the final token budget to biomedical visual evidence, 30% to quantitative and tabular evidence, 20% to mechanism and structure evidence, and 5% to auxiliary evidence.

A.3 Medical Web Curation Diagnostics

The medical-retrieval classifier identifies 22.90M candidate documents containing approximately 33.00B text tokens across MINT-1T, OBELICS, and OmniCorpus-CC. MinHash deduplication reduces this pool to 16.20M documents and 23.29B text tokens. The quality filter removes 2.19M records assigned a score of zero or an invalid prediction. Among 30.02M referenced image URLs, 22.59M are downloaded successfully and joined back to their source records. Subsequent format validation and visual-category-aware sampling produce the 11.27B-token web corpus used for training.

The medical-retrieval classifier is trained on 4,024 annotated examples and reaches 98.5% accuracy on a separate balanced set of 200 documents. The three-level quality classifier is trained on 2,058 examples and obtains 72.5% accuracy and 73.4% macro-F1 on 255 held-out documents. Its precision for the discarded score-0 class is 90.3%. The border-text removal model is trained on 4,619 webpages and achieves 82.1% line-level accuracy, 77.9% precision, and 83.1% recall over 4,923 an-

notated lines from 256 held-out pages. These diagnostics evaluate individual filtering components rather than the end-to-end quality of the resulting corpus.

A.4 Stage-Specific CPT Mixtures

Vision-encoder adaptation uses a 1B-token subset of PMC-InterCPT. Each retained record is reduced to one biomedical image and its associated caption. Deterministic token-budget sampling selects 1,712,897 examples containing 262.50M text tokens and 737.50M visual tokens. The learning-rate decay stage instead uses the compact mixture in Table 4. It emphasizes curated medical sources while retaining a controlled amount of general replay.

Table 4: Composition of the 2.5B-token mixture used during learning-rate decay.

Source	**Tokens**	**Share**
Medical e-books	1.000B	40.00%
Biomedical scientific literature	0.500B	20.00%
General multimodal replay	0.396B	15.84%
Medical web articles	0.094B	3.76%
Medical web exercises and selected VQA	0.510B	20.40%
Total	2.500B	100.00%

B Additional Post-Training Data Details

B.1 Post-Training Quality Assessment

Table 5 presents the complete annotation protocol used to train the sample-quality evaluator. Score 0 marks critical defects that invalidate a sample, Score 1 denotes limited but non-fatal issues, and Score 2 denotes a valid sample without the listed defects.

Table 5: Fine-grained quality assessment criteria used for multimodal data filtering.

Dimension	**Score**	**Criterion**	**Description**
Image–question alignment	0	Non-medical image	The image is unrelated to the medical context required by the question.
	0	Modality mismatch	The visual modality required by the question is inconsistent with the provided image, e.g., a question referring to CT while the input is an X-ray.
	0	Anatomical mismatch	The organ, anatomical region, or laterality specified by the question is absent from the provided image(s).
	0	Missing image	The question refers to an image or panel that is not provided, or the available images are insufficient for answering the question.
	1	Image-independent question	The question can be answered without using the visual input.
Information completeness	0	Invalid answer space	The candidate answers do not contain a valid option or are not mutually compatible with the intended question.
	0	Insufficient conditions	Necessary clinical context, parameters, or problem conditions are missing, making the question under-specified or unsolvable.
	0	Incorrect image description	The textual description or premise is inconsistent with the visual content.
	0	Missed or incorrect visual evidence	Key visual evidence, such as labels or lesions, is incorrectly captured or omitted.
	0	Incorrect answer	The answer is incorrect despite sufficient and valid problem conditions.
	1	Missing measurement scale	The task requires quantitative measurement, but the image does not provide a reliable scale or reference.
Logical validity	0	Ambiguous question	The question is ill-posed, internally inconsistent, or insufficiently precise.
	0	Inconsistent response	The response does not address the question or is logically incompatible with it.
	0	Low-value question	The question does not provide meaningful medical or multimodal supervision.
Overall validity	2	Valid sample	The sample contains sufficient visual and textual evidence, is logically well-formed, and provides a valid target response.

B.2 Supervised Fine-Tuning Mixture Composition

The final mixture contains 296,725 effective training examples and 241.848M tokens. Vision and text contribute 120.459M and 121.389M tokens, respectively.

Table 6: Composition of the supervised fine-tuning mixture.

Category	Examples	Tokens	Share	Vision	Text
General multimodal instruction	50,000	26.021M	10.76%	16.909M	9.112M
Medical visual question answering	90,734	52.644M	21.77%	41.699M	10.945M
Medical reasoning data	49,808	54.868M	22.69%	16.469M	38.399M
Stability-aware supervision	106,183	108.316M	44.79%	45.382M	62.933M
Total	296,725	241.848M	100.00%	120.459M	121.389M

B.3 Stability-Aware Routing and Retention

Table 7 reports the data flow through stability-aware response regeneration. We draw eight candidate responses for each of 176,948 questions. Rule-based answer checking retains 704,046 correct candidates, and 121,544 questions have at least one correct candidate. Questions with no valid candidate are discarded. After candidate selection, removal of single-success cases, and final formatting and cleaning, 106,183 examples enter SFT. The selected supervision contributes 108.316M training tokens, including 45.382M visual tokens and 62.933M text tokens.

Table 7: Data flow through stability-aware response regeneration. Percentages for candidates use all generated candidates as the denominator; percentages for questions use the original question set.

Stage	Count	Retention
Input questions	176,948	100.00%
Generated candidates (8 per question)	1,415,584	100.00%
Candidates passing answer checking	704,046	49.74%
Questions with at least one valid candidate	121,544	68.69%
Long-form examples exported after stability filtering	108,293	61.20%
Final stability-aware SFT examples	106,183	60.01%

Correct candidates contain an average of 418.31 assistant tokens, whereas their answer-masked reconstructions contain 11.62 tokens on average. Of the 704,046 reconstructed answers, 662,503 remain correct, corresponding to 94.10%. The final exported long-form responses average 418.99 assistant tokens. These statistics show that the reconstruction step tests whether a long response preserves the reference answer without replacing the informative long-form target used for SFT.

C Training and Implementation Details

Table 8 consolidates the optimization settings used across continued pretraining and supervised fine-tuning. All CPT and SFT stages use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, and gradient clipping at a maximum norm of 1.0. During warmup–stable CPT and learning-rate decay, we use a weight decay of 0.1 while excluding bias and RMSNorm parameters. Packed sequences preserve sample boundaries through position IDs, and image placeholder and delimiter tokens are excluded from the language-model loss.

Table 8: Optimization settings for the main training stages. Batch sizes are reported as global and per-device micro-batch sizes.

Stage	Trainable components	Epochs	Max. length	Batch	Learning-rate schedule
Vision-encoder adaptation	Vision encoder, projector	1	8,192	256 / 2	3% warmup to 2×10^{-6}, cosine decay to 2×10^{-7}
Warmup–stable CPT	All components	1	12,288	256 / 1	3% warmup to 1×10^{-5}, then constant
Learning-rate decay	All components	1	12,288	256 / 1	Cosine decay from 1×10^{-5} to 2×10^{-6}
Supervised fine-tuning	All components	5	9,000	64 / 2	10% warmup, then cosine decay to zero. Peak 2×10^{-6} for language model and projector and 1×10^{-6} for vision encoder

C.1 Reinforcement Learning with Verifiable Rewards

C.1.1 Optimization Settings

RLVR is run for three epochs with a learning rate of 1×10^{-6} and a global batch size of 128. For each prompt, the policy samples eight rollouts at temperature 1. We retain the default KL regularization setting of the training framework. The reward combines format compliance and answer accuracy with a weight ratio of 1:9, placing most of the optimization signal on verifiable correctness while preserving the required response format. Table 9 summarizes these settings.

Table 9: Optimization settings for reinforcement learning with verifiable rewards.

Hyperparameter	Value
Learning rate	1×10^{-6}
Global batch size	128
Training epochs	3
Rollouts per prompt	8
Sampling temperature	1
Format-to-accuracy reward ratio	1:9

C.1.2 Data Composition

The RLVR resources combine image-grounded clinical questions with text-only medical reasoning. Table 10 reports the provided train and test partitions for each source.

Table 10: Composition of the medical reasoning resources prepared for RLVR.

Category	Source	Train	Test	Total
Medical VQA	PMC-VQA	7,210	790	8,000
	MedSynVQA	5,371	629	6,000
	GMAI-VL-5.5M	2,721	279	3,000
	Internally curated medical data	1,848	206	2,054
	OmniMedVQA + RadImageNet	426	49	475
Text QA	MedQA (English USMLE)	1,800	200	2,000

D Evaluation Details

D.1 Benchmark Overview

MMMU-Medical-test. MMMU evaluates expert-level multimodal perception and reasoning with questions collected from examinations, quizzes, and textbooks across six broad disciplines, 30 sub-

jects, and 183 subfields (Yue et al., 2024). We report the test questions belonging to its Health and Medicine discipline.

MMMU-Pro-Medical-10. MMMU-Pro strengthens MMMU by removing questions that can be solved reliably without images, expanding the answer space, and introducing a vision-only setting (Yue et al., 2025). We use its 10-option Health and Medicine subset, which reduces the influence of option-level shortcuts and random guessing.

MedXpertQA-MM. MedXpertQA contains 4,460 expert-level medical questions spanning 17 specialties and 11 body systems (Zuo et al., 2025). Its multimodal subset combines diverse medical images with clinical information such as patient records and examination findings, emphasizing medical knowledge and multi-step reasoning rather than caption-level recognition.

PMC-VQA-clean. PMC-VQA is constructed from biomedical articles and contains approximately 227k visual question-answer pairs associated with 149k images across diverse modalities and diseases (Zhang et al., 2023b). We evaluate on the cleaned PMC-VQA split used throughout our experiments.

OmniMedVQA. OmniMedVQA aggregates authentic medical images from 73 datasets, covering 12 imaging modalities and more than 20 anatomical regions (Hu et al., 2024). Its broad visual coverage complements the knowledge- and reasoning-oriented benchmarks above.

D.2 Inference and Prompting Protocol

All models are evaluated with native thinking disabled (`enable_thinking=False`) and a maximum generation length of 4,096 tokens (`max_new_tokens=4096`). MMMU-Medical-test, MMMU-Pro-Medical-10, MedXpertQA-MM, and PMC-VQA-clean use the following reasoning prompt:

```
You should think step by step and answer with the option's letter from
the given choices and put the letter within <answer> and </answer>.
```

OmniMedVQA uses a direct-answer prompt:

```
Answer the question directly and put the option's letter from the
given choices within <answer> and </answer>.
```

The evaluator extracts the option letter enclosed by the answer tags and computes exact-match accuracy. The same benchmark examples, prompts, generation limits, and answer-extraction procedure are used for every evaluated model.

D.3 Qualitative Case Studies

Figures 3–6 compare representative responses from the backbone models and InfiMed2 variants. Red text marks an incorrect claim or answer, whereas green text marks correctly identified evidence, reasoning, or conclusions.

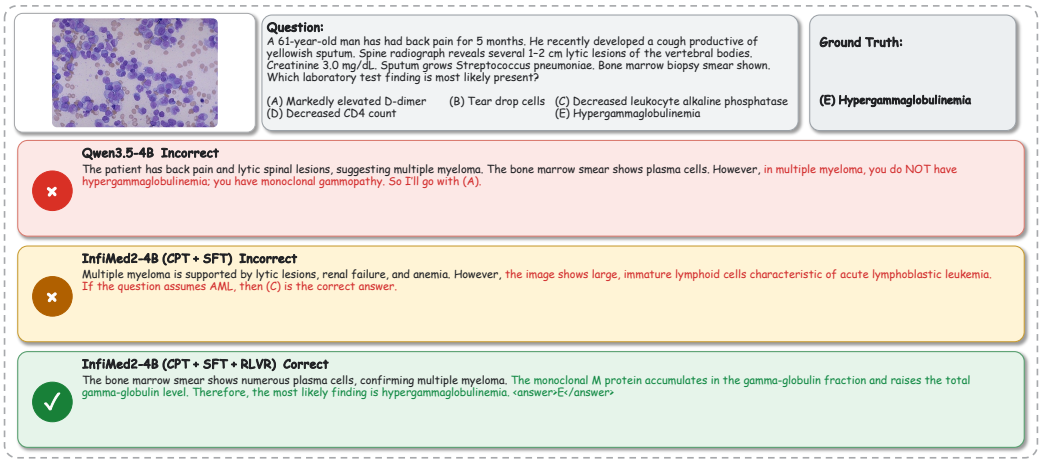

Figure 3: Qualitative comparison on a multiple-myeloma question from MMMU-Medical-test.

Case 1: integrating visual and clinical evidence. All three responses recover at least part of the clinical pattern, but they differ in how the marrow image is integrated with the accompanying findings. Qwen3.5-4B recognizes multiple myeloma yet incorrectly treats monoclonal gammopathy as incompatible with hypergammaglobulinemia. The CPT+SFT model instead misidentifies plasma cells as lymphoblasts and changes the diagnosis. After RLVR, InfiMed2-4B jointly uses the plasma-cell morphology, osteolytic lesions, renal impairment, and monoclonal protein to select the correct answer. The example illustrates that a correct disease hypothesis is insufficient when the final laboratory implication remains inconsistent with it.

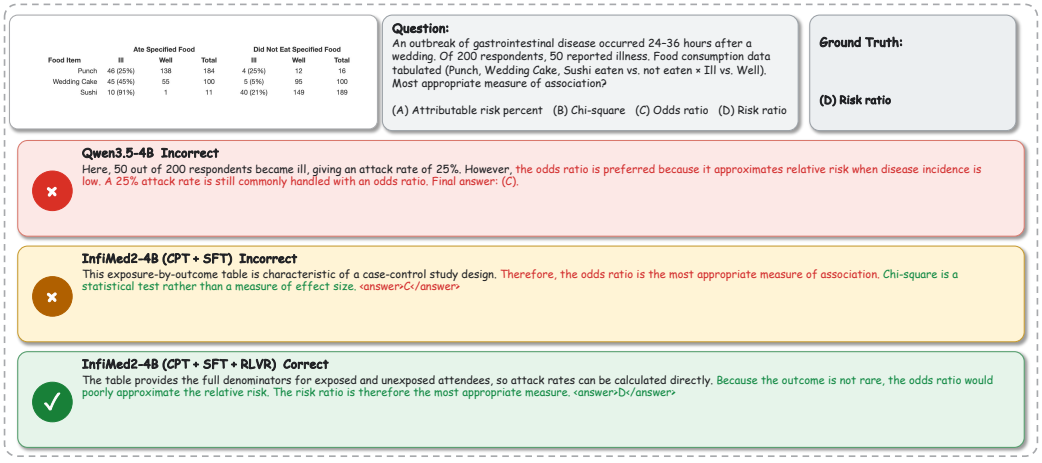

Figure 4: Qualitative comparison on an outbreak-analysis question from MMMU-Medical-test.

Case 2: identifying the study design. The two incorrect responses focus on the odds ratio, despite the table providing denominators for both exposed and unexposed wedding attendees. One response notices the 25% attack rate but still applies the rare-disease approximation, while the other misclassifies the retrospective cohort as a case-control study. The RLVR model uses the observable attack rates to identify the risk ratio as the direct measure of association. This case highlights the

importance of selecting a statistical quantity from the data-generating design rather than from the superficial appearance of a contingency table.

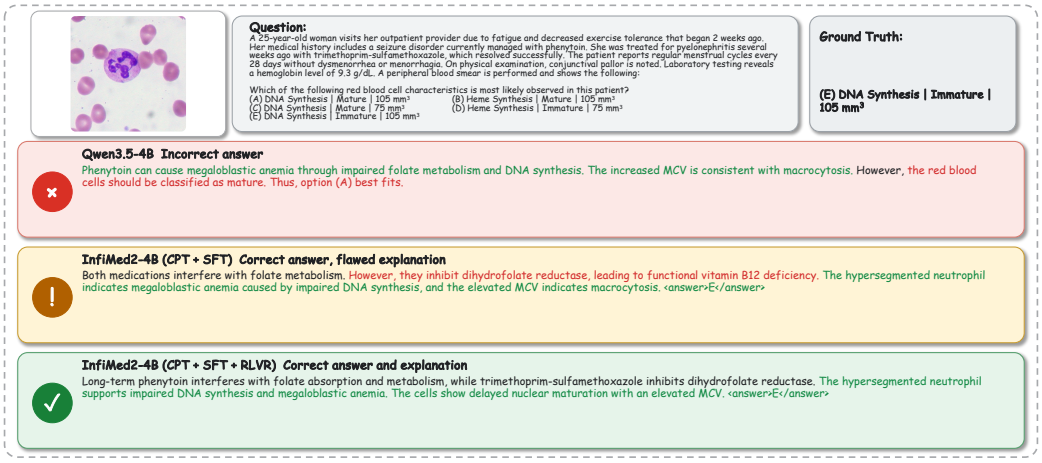

Figure 5: Qualitative comparison on a drug-induced megaloblastic-anemia question from MedXpertQA-MM.

Case 3: separating answer correctness from explanation quality. Qwen3.5-4B correctly associates phenytoin with impaired folate metabolism and macrocytosis but selects the wrong maturation state. The CPT+SFT model reaches the correct option while incorrectly attributing the mechanism to functional vitamin B12 deficiency and treating both drugs as dihydrofolate-reductase inhibitors. In contrast, the RLVR response distinguishes phenytoin-related folate disruption from the action of trimethoprim-sulfamethoxazole and aligns the mechanism, smear finding, and selected option. This example motivates evaluating the factual consistency of an explanation even when its final answer is correct.

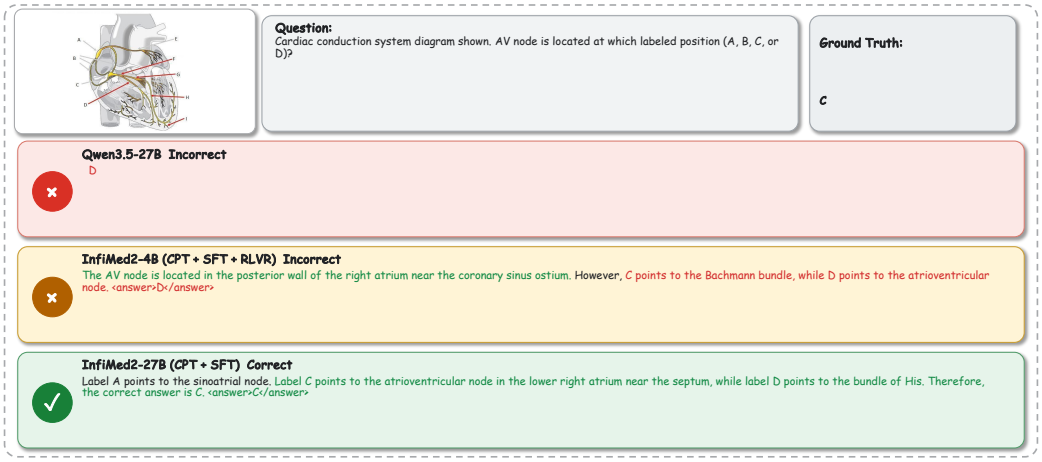

Figure 6: Qualitative comparison on cardiac-conduction anatomy from MMMU-Medical-test.

Case 4: grounding anatomical knowledge in diagram labels. Qwen3.5-27B provides only the incorrect label. InfiMed2-4B with RLVR accurately describes the anatomical location of the atri-

oventricular node but maps that description to the wrong marker in the diagram. InfiMed2-27B correctly distinguishes the sinoatrial node, atrioventricular node, and bundle of His and maps the atrioventricular node to label C. The comparison exposes a residual failure mode in which verbal anatomical knowledge is correct but its spatial grounding is not, while also showing that the larger InfiMed2 variant resolves the visual-label correspondence in this example.