

# Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner

Wenchuan Zhang<sup>1,2\*</sup>, Penghao Zhang<sup>3\*</sup>, Jingru Guo<sup>4\*</sup>, Tao Cheng<sup>5</sup>, Jie Chen<sup>2</sup>,  
Shuwan Zhang<sup>6</sup>, Zhang Zhang<sup>1</sup>, Yuhao Yi<sup>1,2†</sup>, Hong Bu<sup>1,2</sup>

<sup>1</sup>Department of Pathology, West China Hospital, Sichuan University

<sup>2</sup>Institute of Clinical Pathology, West China Hospital, Sichuan University

<sup>3</sup>Independent Researcher

<sup>4</sup>University of Toronto

<sup>5</sup>Business School, Sichuan University

<sup>6</sup>Department of Pathology, Shengjing Hospital of China Medical University

zhangwenchuan@stu.scu.edu.cn, yuhaoyi@scu.edu.cn

## Abstract

Recent advances in vision-language models (VLMs) have enabled broad progress in the general medical field. However, pathology still remains a more challenging sub-domain, with current pathology-specific VLMs exhibiting limitations in both diagnostic accuracy and reasoning plausibility. Such shortcomings are largely attributable to the nature of current pathology datasets, which are primarily composed of image–description pairs that lack the depth and structured diagnostic paradigms employed by real-world pathologists. In this study, we leverage pathology textbooks and real-world pathology experts to construct high-quality, reasoning-oriented datasets. Building on this, we introduce Patho-R1, a multimodal RL-based pathology Reasoner, trained through a three-stage pipeline: (1) continued pretraining on 3.5 million image-text pairs for knowledge infusion; (2) supervised fine-tuning on 500k high-quality Chain-of-Thought samples for reasoning incentivizing; (3) reinforcement learning using Group Relative Policy Optimization and Decoupled Clip and Dynamic sAmpling Policy Optimization strategies for multimodal reasoning quality refinement. To further assess the alignment quality of our dataset, we propose Patho-CLIP, trained on the same figure-caption corpus used for continued pretraining. Comprehensive experimental results demonstrate that both Patho-CLIP and Patho-R1 achieve robust performance across a wide range of pathology-related tasks, including zero-shot classification, cross-modal retrieval, Visual Question Answering, and Multiple Choice Question.

**Code** — <https://github.com/Wenchuan-Zhang/Patho-R1>

**Extended version** — <https://arxiv.org/abs/2505.11404>

## Introduction

In the medical domain, pathology, as the gold standard of modern clinical diagnosis, plays a pivotal role in guiding

\*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

therapeutic decisions (Foersch et al. 2023). However, building robust AI systems for pathology presents significantly greater challenges compared to other medical imaging tasks like MRI (Wu et al. 2023) or CT (Wu, Zhuang, and Chen 2024). This stems from the inherent complexity and fine-grained nature of pathology data, coupled with the scarcity of high-quality, expert-annotated multimodal datasets (Xie et al. 2024). While general vision-language models (VLMs), such as CLIP (Radford et al. 2021) and LLaVA (Liu et al. 2023), have shown promising results across various medical imaging tasks (Wang et al. 2022; Moor et al. 2023), their performance in pathology remains limited, especially in terms of cross-disease generalization and the comprehension of complex diagnostic tasks, highlighting the urgent need for pathology-specific solutions (Moor et al. 2023; Li et al. 2023).

To address this gap, recent efforts introduced pathology-adapted multimodal datasets (Ikezogwo et al. 2023; Huang et al. 2023; Sun et al. 2025). While models trained on these datasets show competitive performance on tasks like Multiple Choice Question (MCQ) and Visual Question Answering (VQA), their decision-making processes remain opaque, limiting their clinical applicability where interpretability and trust are essential. Existing datasets, sourced from educational videos, social media, and institutional archives, tend to favor surface-level descriptions over semantically rich, diagnostically grounded interpretations. “Textbook-quality” data (Gunasekar et al. 2023), encompassing detailed disease mechanisms and organized domain knowledge, could enable models to reason more clinically.

Notably, recent success in reinforcement learning (RL), exemplified by DeepSeek-R1 (Guo et al. 2025) and OpenAI o1 (Jaech et al. 2024), highlights RL’s potential in enhancing reasoning within language models. Among existing methods, the Group Relative Policy Optimization (GRPO) stands out for its simple but effective ground truth-driven reward design, and has been validated in logic-intensive tasks like math and coding, as well as in domains requiring systematic thinking (Huang et al. 2025; Yang et al. 2025).

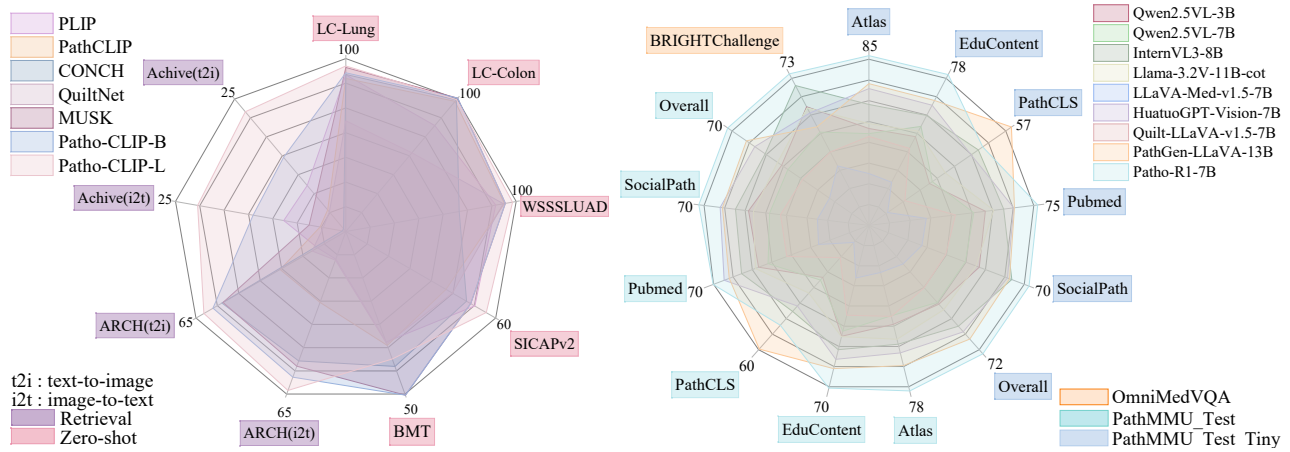


Figure 1: The performance of the proposed Patho-CLIP (left), and the Patho-R1 (right).

More intriguingly, there are works that push it forward: the Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO) (Yu et al. 2025b) leverages a higher clip ratio and dynamically sampled training batches to yield promising performance with fewer training steps, offering an inspiring alternative to domain-aligned reward mechanism design.

Specifically, our key contributions include:

- We propose a comprehensive **data curation pipeline** requiring minimal human effort while ensuring scalable generation of **high-quality SFT data with reasoning**.
- We present **Patho-CLIP**, an open-source pathology adapted CLIP model, which outperforms state-of-the-art models in classification and retrieval tasks.
- We explore the end-to-end training process of domain adaptation of pretrained vision-language models, especially the latest reinforcement learning methods: **GRPO** and **DAPO**. We release the model weights of our vision-language pathology reasoning model: **Patho-R1**, which demonstrates superior performance on various benchmarks as shown in Figure 1.

## Related Work

**Large Vision Language Models in Medical Fields.** The advent of general-purpose VLMs such as CLIP (Radford et al. 2021) has propelled multimodal progress in the medical domain. To better capture domain-specific semantics, various CLIP-based adaptations have been proposed, including PubMedCLIP (Eslami, Meinel, and De Melo 2023), MedCLIP (Wang et al. 2022), UniMed-CLIP (Khattak et al. 2024), and Med-PaLM (Singhal et al. 2023); for computational pathology in particular, specialized variants such as PLIP (Huang et al. 2023), CONCH (Lu et al. 2024a), MUSK (Xiang et al. 2025), TITAN (Ding et al. 2024), and QuiltNet (Ikezogwo et al. 2023) further tailor the image-text pre-training paradigm to histopathology images and reports. To further enhance the utility of medical VLMs, a new line of multimodal medical assistants has emerged, including MedFlamingo (Moor et al. 2023), LLaVA-Med (Li et al. 2023),

RadFM (Wu et al. 2023), and HuatuoGPT-Vision (Chen et al. 2024); their pathology-centric derivatives, including PathAsst (Sun et al. 2024b), Quilt-LLaVA (Seyfioglu et al. 2024), PathChat (Lu et al. 2024b), PathGen-LLaVA (Sun et al. 2025), and PA-LLaVA (Dai et al. 2024). These models use large language models for image-conditioned dialogue and question answering, enhancing interactive and context-aware medical understanding.

## Vision-Language Reinforcement Learning for Post-Training.

Recently, the focus of large language model (LLM) research has been shifting from traditional supervised learning to RL as a means to enhance reasoning capabilities (Guo et al. 2025; Jaech et al. 2024). One line of work focuses on Chain-of-Thought (CoT) prompting, which involves creating structured multimodal reasoning datasets and using instruction tuning to enhance reasoning traceability, as seen in LLaVA-CoT (Xu et al. 2024) and LlamaV-o1 (Thawakar et al. 2025). In contrast, approaches like Visual-RFT (Liu et al. 2025), VLM-R1 (Shen et al. 2025), and Perception-R1 (Yu et al. 2025a) bypass CoT-style tuning and instead apply reinforcement learning to directly improve visual perception, guided by carefully designed reward functions under unsupervised or weakly supervised settings. Building on this, methods such as R1-OneVision (Yang et al. 2025), R1-V (Chen et al. 2025), and Vision-R1 (Huang et al. 2025) introduce a novel cross-modal reasoning pipeline: they first convert images into formalized language representations capturing visual semantics, then construct language-centric visual reasoning datasets for instruction tuning, followed by an RL phase to incentivize deeper multimodal reasoning. While these methods have achieved notable progress in general domains, their application to medical imaging remains nascent. For instance, MedVLM-R1 (Pan et al. 2025) and Med-R1 (Lai et al. 2025) represent only preliminary adaptations, lacking systematic modeling and expert-guided task design.

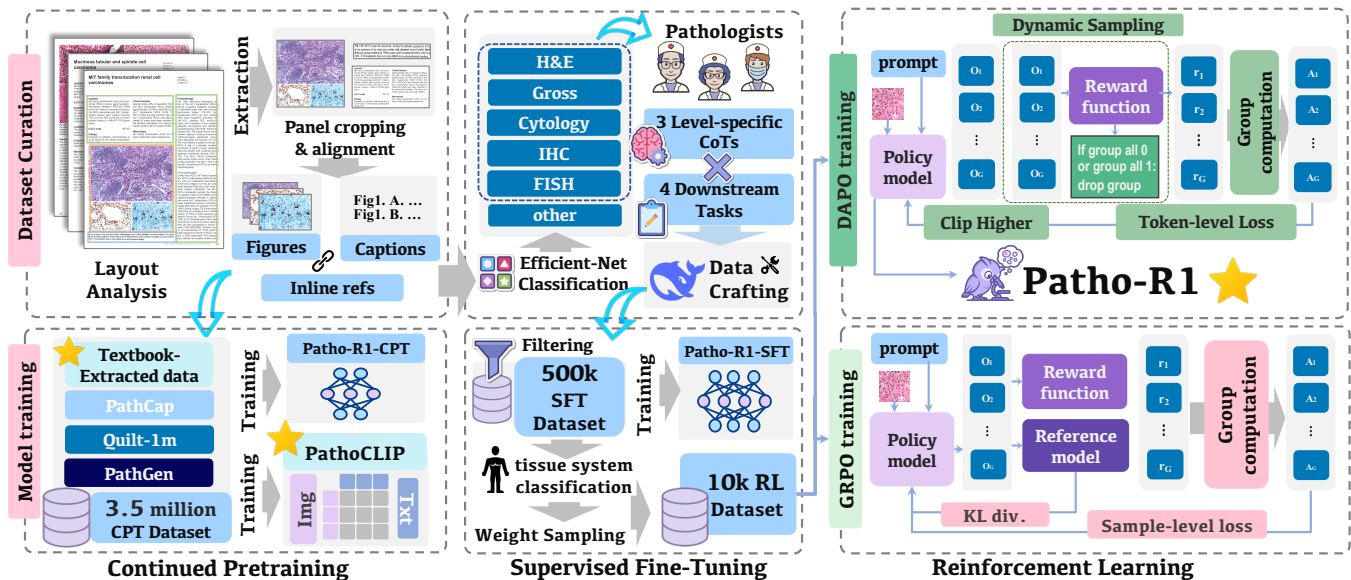


Figure 2: The sequential pipeline for developing Patho-R1. We start by extracting and aligning figures, captions, and in-line references. The image caption pairs, combined with public datasets, will be used for Patho-CLIP training and CPT for Patho-R1 training. Pairs with in-line references will be categorized into pathological sub-fields, clustered by difficulty, applied into tasks with level-specific CoTs to construct SFT data. Data with diagnosis information will be sampled for the final RL.

## Method

### Dataset Overview

As VLMs continue to show promise in domain-specific applications, the construction of high-quality datasets becomes increasingly vital for developing clinically meaningful models. However, existing datasets face two key limitations: (1) *they primarily focus on general medical domains, lacking coverage of pathology-specific data*; and (2) *they often omit detailed, authoritative diagnostic reasoning, emphasizing only brief descriptions and conclusions*. To address these gaps, we curated a large-scale, pathology-specific multimodal corpus comprising three public datasets and 660 authoritative pathology textbooks and educational notes. An overview of our data curation and model training pipeline is presented in Figure 2, with datasets for each training phase detailed below. For further information, refer to Appendix A.

**Continued Pretraining (CPT):** Our CPT dataset comprises a total of **3.5 million image-text pairs**, of which 2.8 million are from PubMed (Sun et al. 2024b), Quilt (Ikezogwo et al. 2023), PathGen (Sun et al. 2025), and 0.7 million from pathology textbooks and notes. This dataset also serves as the training data for our Patho-CLIP.

**Supervised Fine-tuning (SFT):** Our SFT dataset comprises **500k samples** drawn from 5 pathological sub-fields: histopathology, gross examination, immunohistochemistry (IHC), cytology, fluorescence in situ hybridization (FISH); with 3 distinct levels of Chain-of-Thoughts per subfield; spanning 4 downstream task types: descriptive analysis, complex reasoning, multi-turn conversations, multiple-choice questions. This results in a total of 60 data combination types.

**Reinforcement Learning:** Our RL dataset comprises 10k diagnostic-oriented MCQs, aligned with the same five pathological subfields defined in the SFT stage.

### Continued Pretraining and Patho-CLIP Training

**Data Preparation.** We first scanned pathology textbooks into high-resolution images and used DocLayoutYolo (Zhao et al. 2024) to segment each page into body text, figure panels, and captions. Figure-caption pairs were matched based on spatial layout, and captions were extracted using OCR. For multi-panel figures, we applied edge detection and label recognition to split composite images and align sub-figures with corresponding caption segments. To leverage the rich contextual information in textbook narratives, we also extracted body text via OCR and used the Qwen-max model to automatically identify in-line references between figures and surrounding text, reducing the need for manual annotation. Further details are provided in Appendix A.2.

**Patho-CLIP Training.** We adopt OpenAI-CLIP-B and CLIP-L as backbone architectures and design a two-stage progressive training paradigm inspired by (Sun et al. 2025) to build a cross-modal model with fine-grained morphological representation and clinical pathology understanding. In Stage I, the model is initialized via contrastive learning on the PathGen-1.6M dataset, emphasizing tissue-cell morphology and spatial organization to establish high-resolution morphological priors. In Stage II, we retain these representations while integrating PathGen-1.6M with Quilt-1M, PathCap, and our textbook-derived dataset, forming a composite corpus of 3.5 million image-text pairs. Empirical results show that this approach significantly improves the model’s ability to capture tissue heterogeneity and achieves state-

of-the-art performance in semantic understanding and image–text retrieval on multiple pathology benchmarks. See Appendix B.1 for further training details.

**Patho-R1 Continued Pretraining.** Continued pretraining has proven effective for infusing domain-specific knowledge into pretrained language models (Parmar et al. 2024), often yielding performance on par with or exceeding that of larger models (Chen et al. 2024). To ensure broad coverage of pathological knowledge, we combined all publicly available pathology datasets with our curated context-rich image–caption pairs. Given the scale of the dataset (3.5 million samples), we selected Qwen2.5VL-3B and Qwen2.5VL-7B as base models to balance model capacity and data volume.

### Invoking Reasoning Ability via SFT

While continued pretraining on domain-specific corpora imparts domain knowledge, it can compromise a model’s instruction-following ability (Cheng, Huang, and Wei 2024). Our goals in this phase are twofold: (1) *to restore instruction-following capabilities*, and (2) *to elicit reasoning behavior aligned with pathological diagnostics*. Unlike general tasks, pathological diagnosis follows a highly structured, systematic process. Thus, a cost-effective data generation strategy that balances granularity and diagnostic specificity is essential for this training phase.

We began by categorizing textbook-extracted image–caption pairs into pathology subfields. A small subset was manually labeled to train an EfficientNet-based (Tan and Le 2019) classifier. Predictions on unlabeled data were manually reviewed and corrected, then added back into training to iteratively improve classifier accuracy. From 10 initial subfields, we selected 5 with the highest data volume and clinical relevance for SFT: histopathology, gross examination, IHC, cytology, and FISH. However, these subfields still lacked sufficient data diversity. To address this, inspired by (Zhang et al. 2023), we applied three-way K-means clustering within each subfield using joint image–caption embeddings. Interestingly, clusters consistently reflected different difficulty levels, which we used to define three tiers of Chain-of-Thought reasoning: easy, medium, and hard. We then paired these 15 CoTs with 4 downstream task types: MCQ, detailed description, complex reasoning, and multi-turn conversation, resulting in 60 total prompts for SFT data generation. See Appendix A.3 for further details.

We selected DeepSeek-R1 as our SFT data generation model for two main reasons. First, current VLMs are not yet adapted to the pathology domain, and their performance in pathological imagery understanding is not satisfactory. In our experiments, supplying them with image–caption–CoT prompt triplets did not outperform strong text-only models using just captions. Second, DeepSeek-R1 excels in multi-step reasoning and long-context understanding—both essential for our complex CoT-style prompts. However, this choice introduced challenges, such as language mixing in medical terms and repetitive outputs. To address these, we applied a quality control stage involving rule-based filtering and selective manual verification. The final cleaned dataset contains 500k samples. See Appendix B.2 for implementation details.

### Reinforcement Learning

For reinforcement learning data generation, we created a diagnosis-oriented MCQ dataset using a non-overlapping subset of the SFT data. To ensure diversity, we applied proportionate stratified sampling based on subfield distribution. For the two largest subfields, namely histopathology and gross examination, we further stratified by tissue system. The selected samples were then paired with diagnostic information to generate MCQs. See Appendix A.4 for details. For each training iteration, GRPO and DAPO sample  $G$  candidate output  $\{o_i\}_{i=1}^G$  from the old policy  $\pi_{\text{old}}$ , then maximize the following objectives respectively and optimize the model  $\pi_\theta$ :

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) &= \mathbb{E}_{\mathbf{v} \sim P(\mathbf{V}), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|\mathbf{v})} \left[ \right. \\ &\quad \left. \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min(r_{i,t} A_{i,t}, \text{clip}(r_{i,t}, 1 - \epsilon, \right. \\ &\quad \left. \left. 1 + \epsilon) A_{i,t}) - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}}) \right] \\ \mathcal{J}_{\text{DAPO}}(\theta) &= \mathbb{E}_{\mathbf{v} \sim P(\mathbf{V}), \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|\mathbf{v})} \left[ \right. \\ &\quad \left. \frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min(r_{i,t} A_{i,t}, \text{clip}(r_{i,t}, \right. \\ &\quad \left. \left. 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) A_{i,t}) \right] \\ &\quad \text{s.t. } 0 < \left| \{o_i \mid \text{is\_equivalent}(a, o_i)\} \right| < G. \end{aligned}$$

where  $A_i = \frac{r_i - \mu(r)}{\sigma(r)}$  denotes the group-relative advantage, which emphasizes outputs that surpass the group average for prioritized optimization.

**Reward Function Design.** We design reward functions to guide model learning from structured and accurate responses. For the **Format Reward**, we require the model to format its response using `<think>...</think>` for intermediate reasoning steps and `<answer>...</answer>` for the final answer. A reward score of 1 is assigned if both tag pairs appear exactly once and no content exists outside these tags; otherwise, the score is 0. For the **Accuracy Reward**, the content within the `<answer>...</answer>` tags is extracted using regular expressions, and only exact matches to a valid option from the predefined set A, B, C, D, E, F are accepted. A reward score of 1 is assigned if the extracted answer matches the ground-truth label and 0 otherwise. For the **Length-Aware Penalty**, we implement a soft overlong punishment similar to the original work of DAPO. For more details on reward function design, see Appendix B.2. For GRPO, the reward function is:

$$R^{\text{GRPO}}(a_i) = 0.1 R_{\text{fmt}}(a_i) + 0.9 R_{\text{acc}}(a_i)$$

where  $R_{\text{fmt}}$  denotes the format reward and  $R_{\text{acc}}$  denotes the accuracy reward. This reward is applied only when  $R_{\text{fmt}}(a_i) = R_{\text{acc}}(a_i) = 1$ ; otherwise, the reward is 0.

$$R^{\text{DAPO}}(a_i) = 0.5 R_{\text{acc}}(a_i) + 0.5 R_{\text{len}}(a_i)$$

where  $R_{\text{acc}}$  denotes the accuracy reward and  $R_{\text{len}}$  denotes the length-aware reward. This reward is applied only when  $R_{\text{acc}}(a_i) = R_{\text{len}}(a_i) = 1$ ; otherwise, the reward is  $-1$ .

Model	ARCH (@k)					Archive (@k)				
	@1	@5	@10	@20	Aver.	@1	@5	@10	@20	Aver.
OpenAI-CLIP-B	0.41/0.22	1.08/1.34	1.97/2.19	3.57/3.42	1.76/1.79	1.93/1.54	4.71/4.54	6.83/6.76	10.98/11.05	6.11/5.97
OpenAI-CLIP-L	0.86/0.71	2.86/2.75	4.46/4.69	7.33/7.63	3.88/3.95	2.45/2.52	6.31/6.54	9.51/9.58	14.31/14.51	8.15/8.29
PLIP	2.31/3.27	8.07/9.15	13.24/14.21	19.79/21.58	10.85/12.05	2.03/2.25	7.42/6.41	10.72/9.67	15.54/15.69	8.93/8.51
PathCLIP	9.86/10.42	24.48/24.74	32.63/33.26	44.12/43.34	27.77/27.94	1.50/1.41	3.14/3.20	4.31/4.71	5.85/6.70	3.70/4.01
CONCH	24.07/25.63	48.29/50.22	60.04/62.43	70.42/72.62	50.71/52.73	0.03/0.10	0.26/0.36	0.52/0.59	1.21/0.98	0.51/0.51
PathGen-CLIP	14.17/14.99	34.15/32.29	45.20/41.44	56.99/51.75	37.63/35.12	0.62/0.62	1.70/1.67	2.81/2.58	3.69/4.18	2.21/2.26
PathGen-CLIP-L	18.71/17.52	40.55/35.97	52.79/46.13	63.76/56.32	43.95/38.99	1.34/1.11	2.88/2.61	4.28/4.18	6.34/6.14	3.71/3.51
QuiltNet	2.49/3.68	8.59/11.20	12.91/17.26	20.68/25.22	11.17/14.34	0.78/1.08	2.19/2.68	3.89/4.35	6.14/6.41	3.25/3.63
PubmedCLIP	0.15/0.19	0.82/1.00	1.23/1.60	1.97/2.94	1.04/1.43	2.03/2.25	7.42/6.41	10.72/9.67	16.54/15.69	9.18/8.51
MUSK	25.52/23.96	50.04/49.29	62.09/62.20	73.36/73.47	52.75/52.23	1.86/1.90	4.41/5.29	6.18/7.84	8.63/11.18	5.27/6.55
<b>Patho-CLIP-B</b>	<u>27.53/26.71</u>	<u>55.69/55.13</u>	<u>67.60/67.08</u>	<u>77.57/77.08</u>	<u>57.10/56.50</u>	<u>3.63/3.43</u>	<u>10.49/10.78</u>	<u>16.01/16.50</u>	<u>23.24/24.15</u>	<u>13.34/13.72</u>
<b>Patho-CLIP-L</b>	<b>30.73/28.20</b>	<b>61.01/59.45</b>	<b>73.92/71.91</b>	<b>83.44/81.77</b>	<b>62.28/60.33</b>	<b>6.27/6.14</b>	<b>17.29/18.27</b>	<b>25.62/26.93</b>	<b>36.05/37.19</b>	<b>21.31/22.13</b>

Table 1: Cross-modal retrieval results on the ARCH and Archive dataset. In each cell, results are displayed as “i2t / t2i” in the format (%/%). The best-performing results highlighted in bold, with the second-best underlined. Aver. = Average.

## Experiments

### Zero-shot cross-modal retrieval

We compared Patho-CLIP against ten baseline CLIP variants: OpenAI-CLIP-B, OpenAI-CLIP-L, PLIP, PathCLIP, CONCH, PathGen-CLIP, PathGen-CLIP-L, QuiltNet, PubmedCLIP, and MUSK. We evaluated each model on two image–caption pairing datasets: Archive, compiled in-house from diverse pathology texts, and ARCH (Gamper and Rajpoot 2021); performance was measured using Recall@K. As shown in Table 1, on ARCH, Patho-CLIP-L achieves the highest scores across all four Recall@K settings, obtaining mean i2t and t2i recalls of 62.28% and 60.33%, respectively, substantially outperforming the leading baseline, CONCH (50.71% and 52.73%); moreover, Patho-CLIP-B also surpasses CONCH. In the Archive dataset, Patho-CLIP-B attains mean i2t and t2i recalls of 13.34% and 13.72%, respectively, whereas Patho-CLIP-L achieves 21.31% and 22.13%, far exceeding PubmedCLIP’s 9.18% and 8.51%. These results demonstrate that the latent-space alignment strategy of Patho-CLIP confers state-of-the-art accuracy and robustness in cross-modal retrieval.

### Zero-shot image classification

To demonstrate Patho-CLIP’s zero-shot classification performance, we evaluated it on five pathology datasets: LC-Lung and LC-Colon (Borkowski et al. 2019), WSSSLUAD (Han et al. 2022), SICAPv2 (Silva-Rodríguez 2020), and BMT (Welch et al. 2024). For each dataset, we augmented class labels with synonyms (e.g., in LC-Colon dataset, the “colon adenocarcinoma” class was enriched with terms such as “adenocarcinoma of the colon” and “colorectal adenocarcinoma”) and constructed category-specific text templates (e.g., “an H&E image of CLASSNAME.”, “an image of CLASSNAME.”).

As shown in Table 2, Patho-CLIP-L surpasses CONCH by 30.75% on WSSSLUAD dataset and by 7.73% on SICAPv2 dataset, achieving the highest accuracy across all five datasets with a mean accuracy of 76.14%. Notably, on BMT dataset, Patho-CLIP-B attains 49.33% accuracy—exceeding Patho-CLIP-L and substantially outperforming QuiltNet (33.83%) and PathGen-CLIP (37.17%).

The strong performance of Patho-CLIP underscores the diversity of our pathology training data and validates our two-stage progressive training paradigm for enhancing semantic understanding in CLIP models.

Model	Lung	Colon	WSSSL	SICAPv2	BMT	Aver.
OpenAI-CLIP-B	54.71	67.78	81.25	25.68	31.43	52.17
OpenAI-CLIP-L	70.34	70.25	76.25	15.41	26.27	51.70
PLIP	88.79	78.38	86.21	50.61	33.33	67.47
PathCLIP	86.74	96.29	91.25	39.30	34.33	69.58
CONCH	88.93	97.35	65.14	47.36	40.67	67.89
PathGen-CLIP	91.15	96.44	81.06	<u>54.15</u>	37.17	71.99
PathGen-CLIP-L	91.85	98.30	82.50	<u>53.39</u>	<u>47.21</u>	74.65
QuiltNet	62.73	58.15	90.50	41.85	33.83	57.41
PubmedCLIP	86.73	95.66	91.25	39.07	34.83	69.51
MUSK	<u>92.93</u>	<u>98.54</u>	87.50	44.96	45.33	73.25
<b>Patho-CLIP-B</b>	89.90	<b>98.60</b>	<u>91.80</u>	49.01	<b>49.33</b>	<u>75.28</u>
<b>Patho-CLIP-L</b>	<b>93.78</b>	97.48	<b>95.89</b>	<b>55.09</b>	38.44	<b>76.14</b>

Table 2: Comparison of different CLIP models on zero-shot pathology image classification datasets with accuracy (%). The top performance is highlighted in bold, with the second-best underlined. Aver. = Average.

### Few-shot image classification with linear probing

We also discuss the effectiveness of the model in few-shot learning, which is fine-tuned by linear probing of the pre-extracted feature representations of the model. We selected two representative datasets, LC-Lung and BMT, for testing under different training sizes of 2, 8, 16, 32, 64 and 128 samples, each of which implements 10 random samples. For each sample setting, we run 10 independent experiments, and finally the performance of the model under multivariate conditions is visualized by box-and-line plots.

The experimental results show that the Patho-CLIP series demonstrates excellent performance in extreme low-sample scenarios, providing a viable solution for annotation-poor clinical environments. As shown in Figure 3, on the BMT dataset, the Patho-CLIP-L model achieves 73% accuracy with only 2 training samples, which is a significant advantage over the comparative models such as CONCH,

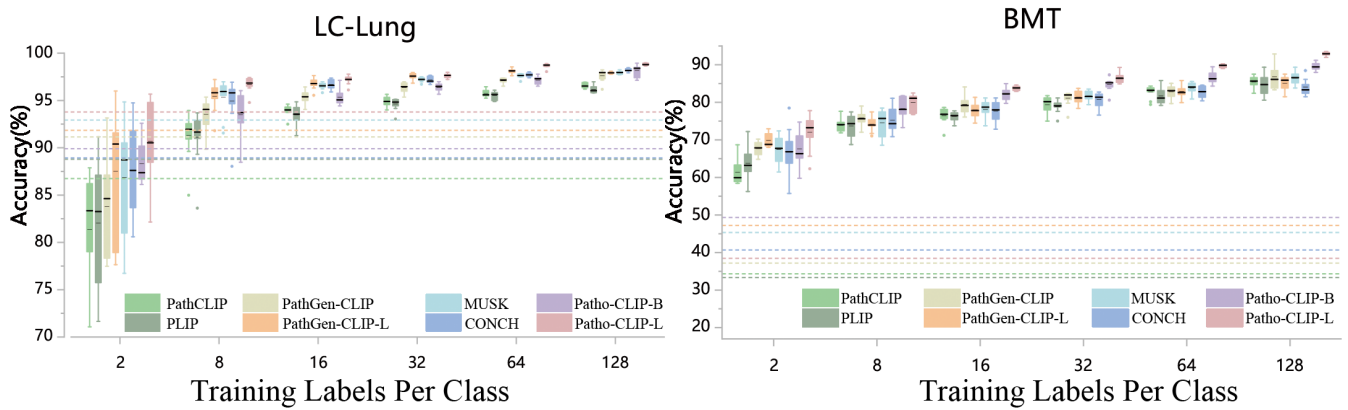


Figure 3: Comparison of few-shot classification accuracy (%) across different CLIP models on various pathology image classification datasets. For reference, the zero-shot performance of each model is shown as a dashed line on the same plot.

Model	PathMMU-val (705) / test-tiny (1139) / test (8454)					YorN (3705)		
	Atlas	EduContent	PathCLS	PubMed	SocialPath	Average	Quilt	Path
<i>Small model without thinking ability</i>								
Paligemma2-3b-pt-224	1.3/1.0/0.5	1.4/2.0/3.2	0.0/0.0/0.1	2.2/2.1/2.0	0.7/0.9/1.6	1.3/1.3/1.6	32.9	44.7
VILA1.5-3B	22.5/28.9/26.5	27.4/29.4/30.6	16.7/16.4/15.1	31.3/34.9/29.1	30.0/23.9/30.7	27.2/27.6/26.7	46.7	52.4
Llama3.2-3B-vl	36.3/30.3/35.8	26.0/27.8/25.0	7.3/7.3/7.0	27.9/27.8/25.4	26.0/30.3/28.9	25.3/25.6/23.4	<b>69.4</b>	55.4
<i>Large model without thinking ability</i>								
LLaVA-Med-v1.5	22.5/25.0/21.4	20.6/21.6/21.5	7.3/7.3/7.7	18.9/24.6/21.3	18.0/22.5/20.5	17.9/20.9/18.6	67.4	55.4
HuatuoGPT-Vision-7B	61.3/65.9/58.1	54.1/60.0/54.7	38.5/40.1/36.6	55.4/61.9/61.4	60.7/58.7/59.4	54.6/58.2/54.6	59.2	<b>65.9</b>
Quilt-LLaVA-v1.5-7B	45.0/42.8/41.4	32.9/38.4/36.7	17.7/14.1/14.7	34.3/37.0/34.8	34.7/32.6/35.3	33.1/34.0/32.0	21.3	20.8
PathGen-LLaVA-13B	67.5/68.3/63.8	61.0/62.4/58.5	<b>50.0/55.9/56.2</b>	59.7/62.3/59.2	53.3/59.6/58.5	58.2/61.9/58.8	47.5	45.8
<i>Small model with thinking ability</i>								
DeepSeek-VL2-Tiny	36.3/38.5/32.2	28.8/33.7/35.4	17.7/19.8/19.2	28.3/38.8/35.4	34.0/36.7/32.1	29.1/34.2/31.4	44.3	46.3
Qwen2.5VL-3B	52.5/46.6/45.1	45.9/44.3/45.3	28.1/23.7/23.7	51.1/49.8/46.8	44.7/46.8/48.1	45.7/43.4/42.1	40.5	43.7
<b>Patho-R1-3B</b>	<b>71.3/74.5/72.2</b>	<b>60.3/67.5/62.5</b>	<b>33.3/36.7/35.0</b>	<b>62.7/67.6/61.1</b>	<b>59.3/66.5/64.1</b>	<b>58.4/63.8/57.9</b>	64.1	49.3
<i>Large model with thinking ability</i>								
InternVL2-8B	41.3/46.6/43.7	41.1/50.6/44.9	24.0/21.5/23.8	42.9/49.1/44.6	45.3/51.4/45.4	40.3/45.1/40.7	63.6	61.4
InternVL2.5-8B	46.3/51.4/50.1	52.1/50.6/50.6	33.3/29.4/32.8	47.6/55.9/50.0	54.0/57.8/50.9	47.8/50.1/47.0	60.1	64.8
InternVL3-8B	52.5/58.2/54.1	45.9/54.9/50.8	40.6/42.9/39.1	52.4/57.7/54.0	54.0/60.6/53.3	49.8/55.4/50.4	33.8	18.6
Llama-3.2-11B-VI	46.3/45.2/41.1	36.3/38.0/37.5	32.3/29.4/26.7	36.1/39.5/38.8	39.3/41.7/39.2	37.5/39.1/36.5	63.3	63.5
Llama-3.2V-11B-cot	46.3/49.0/51.8	42.5/47.1/45.5	35.4/29.9/30.8	47.6/53.4/48.2	52.7/45.4/46.1	45.8/46.0/44.2	54.8	56.4
LLaVA-Onevision-7B	20.0/31.3/21.7	19.2/21.2/21.3	16.7/13.6/12.0	30.0/31.3/27.8	23.3/18.4/21.3	23.4/23.8/23.4	24.2	52.4
Qwen2.5VL-7B	43.8/44.2/41.2	34.9/49.4/43.2	29.2/24.9/24.8	39.9/44.8/42.8	42.7/40.8/39.7	38.4/41.9/38.7	52.2	41.8
<b>Patho-R1-7B</b>	<b>82.5/81.7/75.3</b>	<b>63.0/75.3/66.4</b>	<b>41.7/44.6/45.4</b>	<b>64.0/72.2/66.1</b>	<b>64.7/67.9/67.9</b>	<b>63.0/69.5/63.4</b>	64.7	47.0

Table 3: Comparison of model performance across multiple tasks. Accuracy (%) is reported for each dataset in the following order: PathMMU val, test-tiny, test, Quil-VQA, and Path-VQA

PLIP, MUSK, PathGen-CLIP, etc. (none of the comparative models exceeds 70% accuracy), and the performance of the Patho-CLIP-L model improves faster and converges faster as the number of training samples increases. The model outperforms comparable methods in terms of performance improvement and convergence speed as the number of training samples increases. The narrower distribution interval presented by its box-and-line plot further confirms the leading edge of the model in terms of task adaptability and robustness. Further experiments are provided in Appendix C.1.

### Patho-R1 Evaluation Results

**Closed-ended Benchmarks** Closed-ended questions play a crucial role in pathology-related tasks, particularly in diagnostic classification. To evaluate model performance on

such tasks, we consider two types of close-ended question datasets: (1) Yes/No questions, selected from Path-VQA and Quilt-VQA; and (2) multiple-choice questions, sourced from PathMMU (Sun et al. 2024a), MedXpertQA (Zuo et al. 2025), and OmniMedVQA (Hu et al. 2024). For both types, we report accuracy as the evaluation metric to ensure consistency and comparability across datasets. Experimental results Table 3 demonstrate that Patho-R1 outperforms previous state-of-the-art pathology-focused LMMs. Specifically, it surpasses PathGen-LLaVA-13B by 7.6 % on the PathMMU test-tiny split (PathGen-LLaVA-13B: 61.9%), by approximately 5% on the PathMMU test set and PathMMU validation set. For ablation studies and other experiments, see Appendix C.2 and C.3.

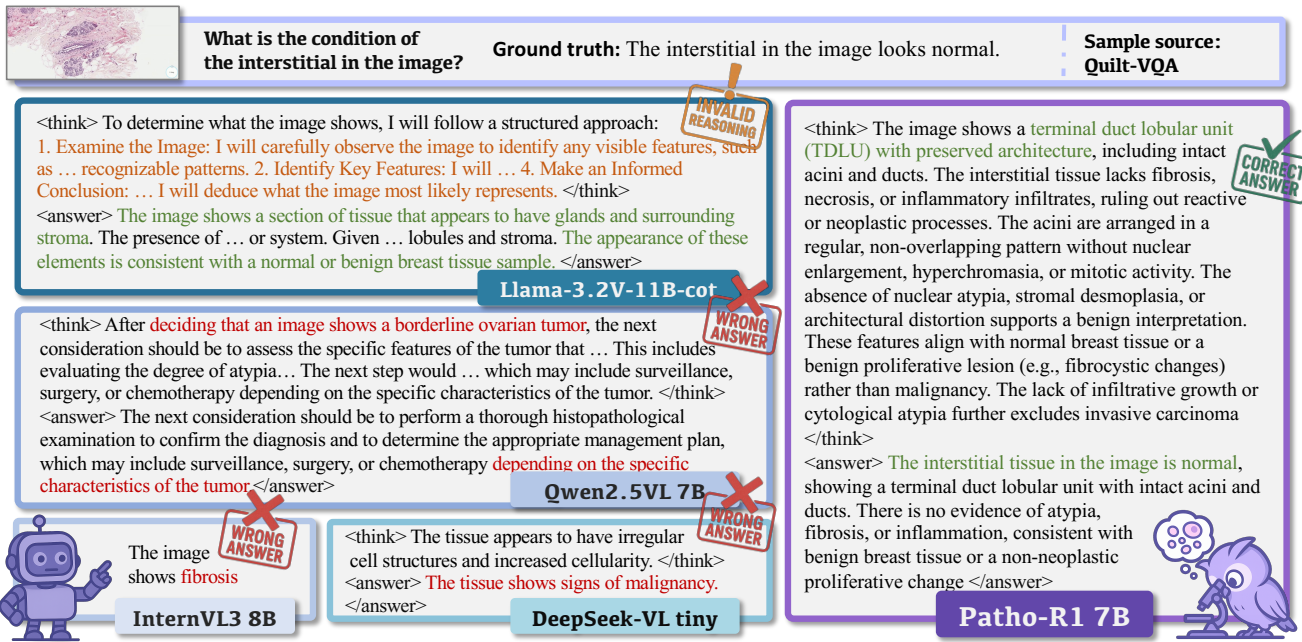


Figure 4: A representative example from the VQA task. While other models exhibit issues in answer formatting, factual accuracy, and reasoning coherence, our model consistently delivers well-structured, accurate, and logically sound responses.

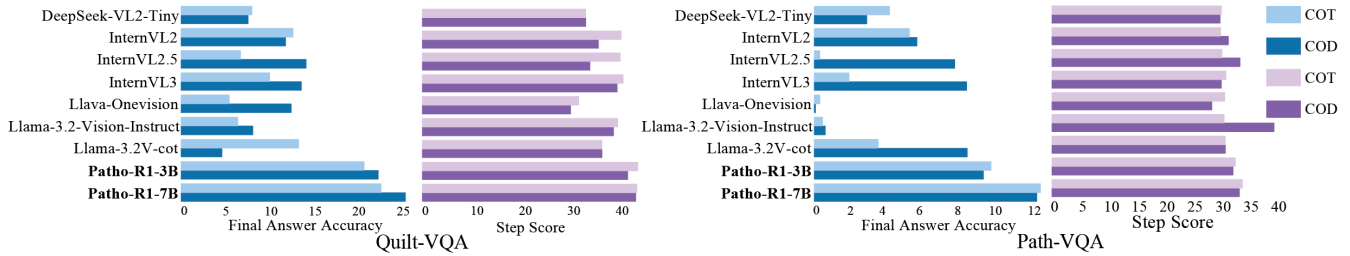


Figure 5: Evaluation of models with reasoning capabilities on final answer accuracy, coherence, and logical consistency of their reasoning steps using Quilt-VQA and Path-VQA.

**Open-ended Benchmarks** Our proposed model, Patho-R1, is designed to perform step-by-step reasoning and produce accurate final answers. As shown in Figure 4, existing general-purpose and medical models do not necessarily possess the reasoning capability required to correctly answer pathology-specific questions. To evaluate its performance, we follow a fuzzy evaluation strategy in which Deepseek-R1 acts as the LLM-judge, scoring generated outputs based on ten well-defined criteria (Thawakar et al. 2025). Evaluations are conducted using the VLMEvalKit (Duan et al. 2024) framework on two publicly available pathology VQA datasets: Quilt-VQA and Path-VQA. This setup ensures consistency, reproducibility, and fair comparison with other models. Our comprehensive evaluation highlights the strength of Patho-R1 in advancing multi-modal reasoning in pathology. In addition to the traditional CoT prompts, we also used the latest Chain-of-Draft (CoD) prompts (Xu et al. 2025) to explore whether simplifying the reasoning chain affects model performance.

Our experimental results in Figure 5 show that Patho-R1 achieves the highest accuracy on both the Quilt VQA and Path VQA open-ended question-answering datasets, while also ranking highly in reasoning quality. An interesting finding is that CoD prompts do help improve answer accuracy, especially on the relatively simpler Quilt VQA dataset, where the overall accuracy is higher.

## Conclusion

In this study, we proposed a data curation pipeline for the three phases of pathology VLM training. Leveraging this pipeline, we constructed a corpus comprising 3.5 million image-text pairs, 500k reasoning-enhanced SFT samples, and 10k diagnosis-oriented MCQs for reinforcement learning. We further developed Patho-R1 and Patho-CLIP, two open-source models adapted via end-to-end domain transfer of pretrained VLMs. These models advance pathology-specific vision-language research and lay a foundation for future development across diverse clinical applications.

## Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62303338).

## References

- Borkowski, A. A.; Bui, M. M.; Thomas, L. B.; Wilson, C. P.; DeLand, L. A.; and Mastorides, S. M. 2019. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142.
- Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; et al. 2024. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7346–7370.
- Chen, L.; Li, L.; Zhao, H.; Song, Y.; and Vinci. 2025. R1-V: Reinforcing Super Generalization Ability in Vision-Language Models with Less Than \$3. <https://github.com/Deep-Agent/R1-V>. Accessed: 2025-02-02.
- Cheng, D.; Huang, S.; and Wei, F. 2024. Adapting Large Language Models to Domains via Reading Comprehension. arXiv:2309.09530.
- Dai, D.; Zhang, Y.; Xu, L.; Yang, Q.; Shen, X.; Xia, S.; and Wang, G. 2024. Pa-llava: A large language-vision assistant for human pathology image understanding. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3138–3143.
- Ding, T.; Wagner, S. J.; Song, A. H.; Chen, R. J.; Lu, M. Y.; Zhang, A.; Vaidya, A. J.; Jaume, G.; Shaban, M.; Kim, A.; et al. 2024. Multimodal Whole Slide Foundation Model for Pathology. arXiv:2411.19666.
- Duan, H.; Yang, J.; Qiao, Y.; Fang, X.; Chen, L.; Liu, Y.; Dong, X.; Zang, Y.; Zhang, P.; Wang, J.; et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 11198–11201.
- Eslami, S.; Meinel, C.; and De Melo, G. 2023. Pubmed-clip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, 1181–1193.
- Foersch, S.; Glasner, C.; Woerl, A.-C.; Eckstein, M.; Wagner, D.-C.; Schulz, S.; Kellers, F.; Fernandez, A.; Tserea, K.; Kloth, M.; et al. 2023. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nature medicine*, 29(2): 430–439.
- Gamper, J.; and Rajpoot, N. 2021. Multiple Instance Captioning: Learning Representations From Histopathology Textbooks and Articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16549–16559.
- Gunasekar, S.; Zhang, Y.; Aneja, J.; Mendes, C. C. T.; Giorno, A. D.; Gopi, S.; Javaheripi, M.; Kauffmann, P.; de Rosa, G.; Saarikivi, O.; Salim, A.; Shah, S.; Behl, H. S.; Wang, X.; Bubeck, S.; Eldan, R.; Kalai, A. T.; Lee, Y. T.; and Li, Y. 2023. Textbooks Are All You Need. arXiv:2306.11644.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.
- Han, C.; Pan, X.; Yan, L.; Lin, H.; Li, B.; Yao, S.; Lv, S.; Shi, Z.; Mai, J.; Lin, J.; et al. 2022. WSSS4LUAD: Grand Challenge on Weakly-Supervised Tissue Semantic Segmentation for Lung Adenocarcinoma. arXiv:2204.06455.
- Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.
- Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-r1: Incentivizing Reasoning Capability in Multimodal Large Language Models. arXiv:2503.06749.
- Huang, Z.; Bianchi, F.; Yuksekogonul, M.; Montine, T. J.; and Zou, J. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9): 2307–2316.
- Ikezogwo, W.; Seyfioglu, S.; Ghezloo, F.; Geva, D.; Sheikh Mohammed, F.; Anand, P. K.; Krishna, R.; and Shapiro, L. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36: 37995–38017.
- Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. OpenAI O1 System Card. arXiv:2412.16720.
- Khattak, M. U.; Kunhimon, S.; Naseer, M.; Khan, S.; and Khan, F. S. 2024. Unimed-clip: Towards a Unified Image-Text Pretraining Paradigm for Diverse Medical Imaging Modalities. arXiv:2412.10372.
- Lai, Y.; Zhong, J.; Li, M.; Zhao, S.; and Yang, X. 2025. Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models. arXiv:2503.13939.
- Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual Reinforcement Fine-Tuning. arXiv:2503.01785.
- Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024a. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3): 863–874.
- Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Zhao, M.; Chow, A. K.; Ikemura, K.; Kim, A.; Pouli, D.; Patel, A.; et al. 2024b. A multimodal generative AI copilot for human pathology. *Nature*, 634(8033): 466–473.

- Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (MLAH)*, 353–367.
- Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. arXiv:2502.19634.
- Parmar, J.; Satheesh, S.; Patwary, M.; Shoeybi, M.; and Catanzaro, B. 2024. Reuse, Don't Retrain: A Recipe for Continued Pretraining of Language Models. arXiv:2407.07263.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.
- Seyfioglu, M. S.; Ikezogwo, W. O.; Ghezloo, F.; Krishna, R.; and Shapiro, L. 2024. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13183–13192.
- Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; et al. 2025. VLM-R1: A Stable and Generalizable R1-Style Large Vision-Language Model. arXiv:2504.07615.
- Silva-Rodríguez, J. 2020. SICAPv2-prostate whole slide images with gleason grades annotations. *Mendeley Data*.
- Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.
- Sun, Y.; Wu, H.; Zhu, C.; Zheng, S.; Chen, Q.; Zhang, K.; Zhang, Y.; Wan, D.; Lan, X.; Zheng, M.; et al. 2024a. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, 56–73.
- Sun, Y.; Zhang, Y.; Si, Y.; Zhu, C.; Zhang, K.; Shui, Z.; Li, J.; Gong, X.; LYU, X.; Lin, T.; and Yang, L. 2025. PathGen-1.6M: 1.6 Million Pathology Image-text Pairs Generation through Multi-agent Collaboration. In *The Thirteenth International Conference on Learning Representations*.
- Sun, Y.; Zhu, C.; Zheng, S.; Zhang, K.; Sun, L.; Shui, Z.; Zhang, Y.; Li, H.; and Yang, L. 2024b. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5034–5042.
- Tan, M.; and Le, Q. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6105–6114.
- Thawakar, O.; Dissanayake, D.; More, K.; Thawkar, R.; Heakl, A.; Ahsan, N.; Li, Y.; Zumri, M.; Lahoud, J.; Anwer, R. M.; et al. 2025. LlamaV-O1: Rethinking Step-by-Step Visual Reasoning in LLMs. arXiv:2501.06186.
- Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive Learning from Unpaired Medical Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3876–3887.
- Welch, E. C.; Lu, C.; Sung, C. J.; Zhang, C.; Tripathi, A.; and Ou, J. 2024. BMT: A Cross-Validated ThinPrep Pap Cervical Cytology Dataset for Machine Learning Model Training and Validation. *Scientific Data*, 11(1): 1444.
- Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards Generalist Foundation Model for Radiology by Leveraging Web-Scale 2D&3D Medical Data. arXiv:2308.02463.
- Wu, L.; Zhuang, J.; and Chen, H. 2024. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22873–22882.
- Xiang, J.; Wang, X.; Zhang, X.; Xi, Y.; Eweje, F.; Chen, Y.; Li, Y.; Bergstrom, C.; Gopaulchan, M.; Kim, T.; et al. 2025. A vision–language foundation model for precision oncology. *Nature*, 638(8051): 769–778.
- Xie, Y.; Zhou, C.; Gao, L.; Wu, J.; Li, X.; Zhou, H.-Y.; Liu, S.; Xing, L.; Zou, J.; and Xie, C. 2024. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. arXiv:2408.02900.
- Xu, G.; Jin, P.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2024. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440.
- Xu, S.; Xie, W.; Zhao, L.; and He, P. 2025. Chain of Draft: Thinking Faster by Writing Less. arXiv:2502.18600.
- Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; et al. 2025. R1-onevision: Advancing Generalized Multimodal Reasoning Through Cross-Modal Formalization. arXiv:2503.10615.
- Yu, E.; Lin, K.; Zhao, L.; Yin, J.; Wei, Y.; Peng, Y.; Wei, H.; Sun, J.; Han, C.; Ge, Z.; et al. 2025a. Perception-R1: Pioneering Perception Policy with Reinforcement Learning. arXiv:2504.07954.
- Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Fan, T.; Liu, G.; Liu, L.; Liu, X.; et al. 2025b. Dapo: An Open-Source LLM Reinforcement Learning System at Scale. arXiv:2503.14476.
- Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2023. Automatic Chain of Thought Prompting in Large Language Models. In *The Eleventh International Conference on Learning Representations*.
- Zhao, Z.; Kang, H.; Wang, B.; and He, C. 2024. DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception. arXiv:2410.12628.
- Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. arXiv:2501.18362.