Enhancing Pathological VLMs with Cross-scale Reasoning

Chi Phan[1⋆], Tianyi Zhang[1⋆], Qiaochu Xue[1], Yufeng Wu[2], Dan Hu[3], Zeyu Liu[2], Sudong Wang[2⋆⋆], and Yueming Jin[1⋆⋆]

[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{chiphan,zhangtianyi,e1352520}@u.nus.edu, ymjin@nus.edu.sg
[2] PuzzleLogic Pte Ltd, Singapore
{yufengwu,zeyuliu,sudongwang}@puzzlelogic.com
[3] Department of Pathology, Fujian Medical University Cancer Hospital & Fujian Cancer Hospital, Fuzhou, China
hudan@fjmu.edu.cn

Abstract. Pathological images are inherently multi-scale, requiring pathologists to integrate evidence from global tissue architecture at low magnification to cellular morphology at higher magnification for accurate diagnosis. While existing pathological datasets for vision-language models (VLMs) include various scales, they often lack explicit cross-scale reasoning objectives. This limitation prevents VLMs from capturing essential cross-scale representations and learning evidence-based reasoning. To bridge this gap, we introduce the first cross-scale training and evaluation paradigm that formulates pathology interpretation as multi-magnification reasoning. However, creating such a task reveals a critical challenge: multi-image visual question answering (VQA) is prone to text-only shortcuts, which allow models to guess answers using magnification-dependent artifacts rather than visual evidence. To address this, we propose a leakage-aware curation pipeline that combines adversarial text-only screening with constraint-guided question design. Using this pipeline, we construct Scale-VQA, a high-quality benchmark with 4,685 multiple-choice questions grounded in 2,537 pathology images across multiple magnification levels. Finally, we present ScaleReasoner-R1, a model trained via reinforcement learning to optimize performance on cross-scale VQA tasks. ScaleReasoner-R1 achieves state-of-the-art performance on our cross-scale reasoning benchmark and generalizes to SOTA performance on established single-scale benchmarks. Findings suggest that even the limited cross-scale supervision can significantly improve pathological understanding. Code is available at https://github.com/iMVR-PL/ScaleReasoner-R1.

Keywords: VLM · Computational Pathology · Deep Learning.

⋆ Chi Phan and Tianyi Zhang contributed equally to this work.
⋆⋆ Sudong Wang and Yueming Jin are co-corresponding authors.

1 Introduction

Pathological image analysis is a core component of cancer diagnosis. With the increasing availability of digitized whole-slide images (WSIs), deep learning methods for computational pathology have advanced rapidly. Recent vision-language models (VLMs) have achieved strong performance on visual question answering (VQA) tasks in medical imaging [14, 23, 25] and pathology [4, 18, 26], enabling more interpretable and interactive computer-assisted workflows.

In routine practice, pathologists reason across magnifications [1, 12, 27]. At low magnification, they assess global tissue architecture (e.g., histologic subtypes and growth patterns). At high magnification, they examine cellular morphology (e.g., nuclear atypia and mitotic activity). Observations at one magnification often determine what to verify at another, and the final diagnosis emerges from integrating evidence across scales.

However, most existing pathology VLMs [4, 18, 24, 26, 7, 17] are developed and evaluated on single-image tasks at a fixed magnification (Fig. 1a). Correspondingly, many benchmarks treat isolated regions of interest (ROIs) as independent diagnostic units [13, 21], or represent WSI as thumbnails or feature embeddings at one specific scale [5, 6, 15]. This single-scale framing creates a mismatch with clinical practice: models may learn scale-specific recognition without acquiring the hierarchical rationale that links low-magnification architecture to high-magnification cellular evidence. Consequently, they often struggle with tasks that require cross-scale verification, such as localizing anatomical compartments (e.g., determining whether malignant cells invade specific tissue layers [7]) or confirming suspected architectural patterns (e.g., validating a low-power suspicion of malignancy with high-power cellular findings [17]). Bridging this gap requires moving beyond isolated recognition toward coordinated evidence integration across magnifications.

Formulating cross-scale VQA tasks, however, encounters a key challenge: text-only shortcut solutions [2, 10]. We refer to a shortcut solution as a case where the correct option can be inferred from the question and options without using the image evidence from all required magnification views [2, 10, 19]. In pathology, this leakage can arise because some findings are strongly associated with specific magnifications, and option sets can contain systematic linguistic or biomedical priors (Fig. 1b). Consequently, a naively constructed multi-image VQA benchmark may allow models to obtain high accuracy while ignoring one or more magnification views, leading to inflated results [2] that do not reflect genuine cross-scale evidence integration. Therefore, reliable cross-scale training and evaluation requires leakage-aware benchmark construction, where questions are designed so that answering correctly depends on evidence across magnifications and text-driven shortcuts are explicitly suppressed [19].

In this work, we introduce a cross-scale training and evaluation paradigm that frames pathology interpretation as coordinated reasoning across magnifications, rather than isolated single-scale recognition. First, we develop a rigorous annotation pipeline: fifteen pathologists (>5 years of experience) and five senior pathologists (>20 years of experience) annotate 177 publicly available TCGA

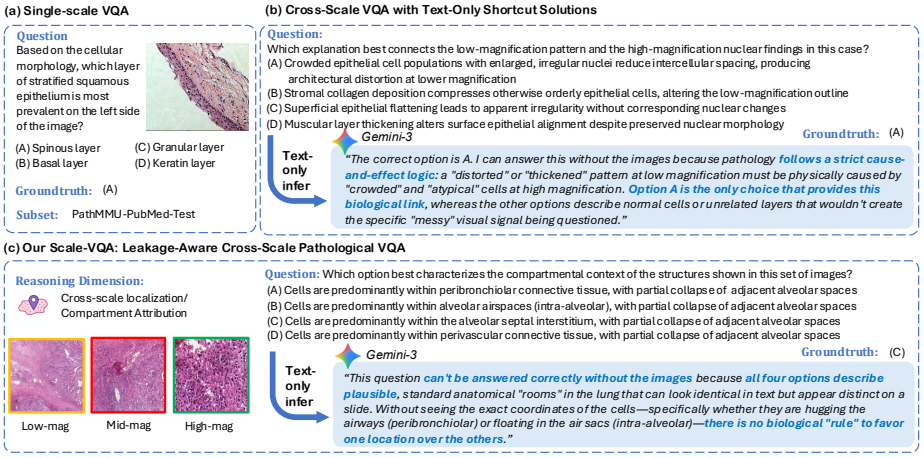

Fig. 1. A comparison of (a) single-scale VQA, (b) naïve cross-scale VQA with text-only shortcut solutions, and (c) our leakage-aware cross-scale VQA.

WSIs [9]. We curate 2,537 ROI captions at 10×, 40×, and 200×, and collect 937 cross-scale captions that synthesize visual patterns across magnifications. Second, we design a leakage-aware VQA curation pipeline (Fig. 2a) that combines adversarial text-only screening with constraint-guided question design to suppress shortcut solutions and ensure each question depends on visual evidence across views. This produces Scale-VQA, a dataset of 4,685 multiple-choice questions that require visually grounded cross-scale reasoning. Finally, we present ScaleReasoner-R1, a model trained via curated reinforcement learning tasks. We enable the cross-scale reasoning insights to improve single-scale performance: ScaleReasoner-R1 achieves state-of-the-art (SOTA) performance on cross-scale multi-image benchmarks as well as existing single-image pathology VQA benchmarks.

In summary, we (i) introduce the first cross-scale paradigm for pathological reasoning, (ii) develop a high-quality cross-scale benchmark Scale-VQA with a leakage-aware VQA curation pipeline, and (iii) release ScaleReasoner-R1, a high-performing RL-trained model optimized for cross-scale VQA. Built on small-scale (177 WSIs) but high-quality data, we demonstrate that curated RL tasks can achieve state-of-the-art results on cross-scale VQA while transferring strongly to single-scale pathology VQA benchmarks. Findings suggest that cross-scale supervision promotes more robust and interpretable pathological understanding even for single-scale perception. Underscoring its practical value for pathological image analysis, we open-source the high-quality Scale-VQA and the curation pipeline to encourage broad community participation.

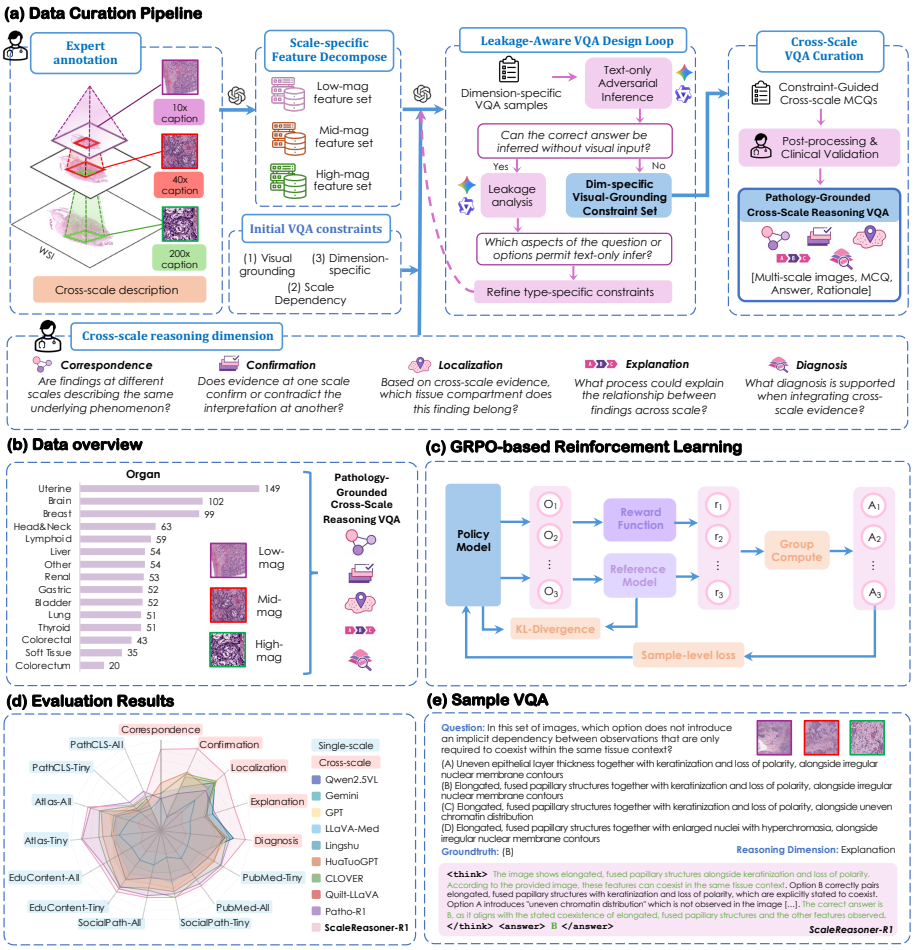

Fig. 2. Overview of Scale-VQA and ScaleReasoner-R1. (a) Leakage-aware curation pipeline. (b) Dataset overview. (c) GRPO-based RL training. (d) Cross-scale reasoning results. (e) Example cross-scale VQA.

2 Methods

2.1 Clinical Annotation for Cross-scale Reasoning

We construct a clinically verified foundation for cross-scale reasoning by annotating 177 publicly available TCGA WSIs [9]. Unlike prior benchmarks [21, 15, 6] that typically only involve experts at the final evaluation stage, our design integrates clinical expertise at the earliest stages of data curation to provide high-fidelity visual grounding and professional captioning. To maximize diagnostic rigor, we utilize a collaboration workflow: each case was independently annotated by three junior pathologists and subsequently reviewed and validated by a

senior pathologist. For every diagnostic case, annotators first select representative ROIs at low magnification (10×) and further zoom in to pick corresponding regions at intermediate (40×) and high (200×) magnifications. This builds a visual search path across scales, which is then confirmed by senior pathologists. Next, pathologists provide the rationale for each zoom-in decision based on the preceding field of view and provide detailed captions of pathological features for every ROI. Finally, we use the GPT-5.2 API to construct a cross-scale clinical synthesis from the zoom-in rationales and ROI captions. Pathologists then refine and confirm the fine-grained visual findings that explicitly link broad tissue architecture with localized cellular morphology. These expert annotations serve as high-quality evidence anchors of the cross-scale search process, enabling clinically meaningful supervision for subsequent VQA curation and model training.

2.2 Scoping Cross-scale Instruction with Leakage-Aware VQA

Building on these annotations, we construct Scale-VQA via a leakage-aware cross-scale curation pipeline that suppresses text-only shortcut solutions (Fig. 2a). The pipeline consists of three steps:

Step 1: Scale-specific Feature Decomposition. We decompose expert annotations into scale-specific evidence sets to reduce semantic leakage and linguistic shortcuts that can arise from full captions. To ensure clinical relevance, we define five cross-scale reasoning dimensions, namely *Correspondence*, *Confirmation*, *Localization*, *Explanation*, and *Diagnosis*. These dimensions were designed and validated with pathologists to assess the multifaceted cross-scale knowledge used in practical clinical assessment. We also establish a set of initial constraints that guide the VQA generation process: (i) visual-grounding constraint that ties each question to the provided images, (ii) scale-dependency constraint requiring evidence from at least two magnification levels, and (iii) dimension-specific constraints to ensure diverse questions across reasoning dimensions for the same image set. These scale-specific feature sets and constraints are then provided as input to GPT-5.2 to generate the candidate MCQs.

Step 2: Leakage-aware Screening via Text-Only Adversarial Inference. To detect and eliminate text-only shortcuts, we introduce an iterative screening loop. For each candidate MCQ, we use Gemini 3 Pro and Qwen3-Max as text-only adversaries that receive only the question and options. If either model answers correctly, we analyze its rationale to diagnose leakage (e.g., linguistic cues, label priors, or unbalanced distractors) to update the constraints. We repeat this process until the adversaries can no longer reliably infer the correct answer without visual input.

Step 3: Cross-scale MCQ Construction and Clinical Validation. Using the refined constraints, we generate the final MCQs, each paired with multi-scale images (low/mid/high), a correct option, and a concise ground-truth rationale. We further post-process the dataset by merging questions across the five dimensions and balancing answer positions to reduce positional bias. Finally, pathologists validate the questions to ensure that correct answers are visually supported and distractors remain clinically plausible.

Together, this pipeline results in SCALE-VQA, a high-quality dataset that reliably measures cross-scale reasoning with 4,685 MCQs across 15 organs and 5 reasoning dimensions (Fig. 2b). By requiring coordinated multi-image, multi-magnification evidence integration, SCALE-VQA provides a principled foundation for training and evaluating pathology VLMs in a setting that better reflects real-world diagnostic workflows.

2.3 Cross-scale Reasoning Optimization via RL

We train SCALEREASONER-R1 on SCALE-VQA using Group Relative Policy Optimization (GRPO) [11] to improve answer selection for cross-scale pathology VQA. While standard supervised fine-tuning (SFT) can overfit to demonstration-style traces or surface correlations, RL provides outcome-driven feedback that encourages SCALEREASONER-R1 to learn decision rules that integrate evidence across magnifications, without requiring costly step-by-step human rationales. As illustrated in Fig. 2(c), given a multi-scale image set $I = \{I^{10\times}, I^{40\times}, I^{200\times}\}$, a question q, and options $\mathcal{O}$, the policy π_θ generates a structured response with reasoning (`<think>`) and a final answer (`<answer>`). We optimize π_θ with a sparse reward $R = \lambda R_{\text{acc}} + (1-\lambda)R_{\text{form}}$, where $R_{\text{acc}} = 1$ if the selected option matches the ground truth (and 0 otherwise), and R_{form} enforces valid output formatting; we set $\lambda = 0.8$. GRPO updates the policy by sampling multiple responses per question and computing group-relative advantages, eliminating the need for an explicit critic and encouraging responses that leverage cross-scale visual evidence to reach correct decisions.

3 Experiments

3.1 Implementation Details

Training Setup. We patient-wise split SCALE-VQA into train/validation/test with 3,230/450/1,005 samples. All multi-scale views and VQA samples derived from the same WSI are kept within the same split to avoid leakage. SCALEREASONER-R1 is initialized from Patho-R1-7B [26], a pathology VLM trained on large-scale pathology data but focused only on single-image analysis. This provides strong domain knowledge before adapting to our cross-scale setting. Based on UnPuzzle Pipeline [16], we train with GRPO for 500 steps on 8×H100 GPUs.
Benchmarks. We evaluate our model on two VQA settings: (1) *Cross-scale multi-image:* SCALE-VQA-Test containing 1,005 VQA samples that require reasoning across magnifications. Performance is reported as accuracy across 5 clinically-aligned reasoning dimensions; (2) *Single-image evaluation:* To assess model performance within the conventional single-image diagnostic paradigm, we utilize PathMMU [21], an authoritative and large-scale public pathological VQA benchmark. We conduct a comprehensive evaluation across all PathMMU partitions (val/test-tiny/test), with a total of 10,387 VQA samples across 5 subsets, namely, Atlas, EduContent, PathCLS, PubMed, and SocialPath.

Table 1. Accuracy on the cross-scale multi-image Scale-VQA-Test benchmark. Results are reported across five cross-scale reasoning dimensions (Correspondence, Confirmation, Localization, Explanation, Diagnosis) and averaged (AVG). The best results are highlighted in **bold**.

	Corresp.	*Confirm.*	*Localiz.*	*Explan.*	*Diagno.*	AVG
General VLM						
Qwen2.5VL-7B	41.79	47.26	71.14	44.28	63.68	53.63
Gemini 3 Flash	48.26	58.71	71.64	53.73	72.14	60.90
GPT-5.2	47.76	59.70	74.13	45.77	65.17	58.51
Medical VLM						
LLaVA-Med-7B	25.87	16.92	28.36	20.90	24.88	23.38
HuatuoGPT-7B	37.81	45.77	65.67	46.77	55.72	50.35
Lingshu-7B	47.26	63.68	62.69	48.26	61.19	56.62
Pathological VLM						
Quilt-LLaVA	32.34	14.43	45.27	30.35	29.35	30.35
CLOVER	37.31	61.69	73.13	46.27	65.77	56.82
Patho-R1	31.84	40.30	59.70	56.22	68.66	51.34
ScaleReasoner-R1	**80.60**	**89.05**	**84.58**	**76.12**	**84.08**	**82.89**
Ablation Study						
SFT	69.65	83.58	78.11	55.22	64.68	70.25
SFT + RL	72.64	86.57	79.10	57.71	68.16	72.84

Compared models. We compare against representative models across domains: general VLMs (Qwen2.5-VL-7B [3], GPT-5.2 [20], Gemini 3 Flash [22]), medical VLMs (LLaVA-Med-7B [14], HuatuoGPT-7B [25], Lingshu-7B [23]), and pathology VLMs (Quilt-LLaVA [18], CLOVER [4], Patho-R1-7B [26]).

3.2 Comparison Results

As shown in Table 1 and Fig. 2(d), ScaleReasoner-R1 establishes a new state-of-the-art for cross-scale reasoning, achieving an average accuracy of 82.89% on Scale-VQA-Test. Compared to its base model, our approach yields a remarkable improvement across all five reasoning dimensions despite the relatively small-scale training data. Notably, although trained only on cross-scale multi-image questions, ScaleReasoner-R1 also improves performance on the single-image PathMMU benchmark (Table 2). Our model's performance even surpasses or approaches the expert performance reported in the original PathMMU paper for most of the subsets, such as in EduContent and Atlas. We also observe a modest drop on PathCLS, which is closer to label-centric single-image classification than morphology-grounded reasoning, suggesting a possible trade-off between cross-scale evidence integration and category-recognition priors. These results demonstrate that cross-scale reasoning supervision strengthens the model's pathological understanding and transfers beyond the training setting to improve on conventional single-image pathology VQA.

Table 2. Overall results of models on the PathMMU validation and test set. The best model performance in each column is highlighted in **bold**. Results marked with † are taken directly from the original PathMMU paper.

	Val	**Test overall**		***PubMed***		***SocialPath***		***EduContent***		***Atlas***		***PathCLS***	
	overall	Tiny	All	Tiny	ALL	Tiny	All	Tiny	All	Tiny	All	Tiny	All
Random Choice†	24.6	22.1	23.7	22.1	25.1	25.5	26.5	25.5	26.0	19.7	23.0	15.3	16.3
Frequent Choice†	27.5	27.7	25.5	28.8	26.1	27.7	26.7	29.8	26.5	28.4	27.5	22.0	21.0
Expert performance†	-	71.8	-	72.9	-	71.5	-	69.0	-	68.3	-	78.9	-
General domain VLM													
Qwen2.5-VL-7B	44.3	48.2	44.4	53.4	50.5	50.0	47.7	59.6	48.9	48.6	46.8	29.4	28.0
Gemini Pro Vision†	41.9	42.8	42.7	43.8	44.9	42.4	42.0	43.5	43.7	49.5	49.4	32.8	34.7
GPT-4V†	49.3	53.9	49.8	59.4	53.5	58.7	53.9	60.4	53.6	48.1	52.8	36.2	33.8
Medical VLM													
LLaVA-Med-7B	17.5	22.2	22.7	24.6	25.1	23.4	21.1	25.1	23.6	17.8	24.5	20.3	19.2
HuatuoGPT-V-7B	43.2	43.8	41.2	49.1	47.7	52.3	46.0	52.9	46.1	44.7	46.9	19.8	19.2
Lingshu-7B	51.9	56.0	53.7	60.5	57.6	61.6	58.1	69.4	58.7	58.1	62.5	30.4	31.6
Pathological VLM													
Quilt-LLaVA	33.8	32.3	31.9	30.3	34.2	26.8	34.6	35.7	33.3	41.8	33.3	27.1	24.0
CLOVER	53.1	60.6	56.6	68.0	59.8	67.6	60.5	68.6	62.2	62.0	64.1	36.7	36.6
Patho-R1	62.3	64.8	62.6	68.7	64.9	63.9	65.7	71.0	66.1	78.4	73.5	**41.8**	**42.7**
ScaleReasoner-R1	**62.4**	**66.2**	**63.8**	**71.2**	**67.3**	**67.6**	**67.6**	**74.9**	**69.2**	**79.3**	**76.2**	37.9	38.5
Ablation Study													
SFT	47.8	44.0	43.3	59.1	52.1	57.9	53.3	56.9	51.6	43.3	47.7	22.0	19.6
SFT + RL	48.3	45.2	44.1	57.7	52.7	56.0	52.8	57.6	52.9	46.6	48.1	23.7	19.7

3.3 Ablation Study: Activating Generalized Pathological Reasoning via RL

To gain deeper insights into the training dynamics and the necessity of our outcome-driven optimization, we compare our GRPO-based RL against SFT-only and SFT+RL paradigms. To enable SFT, we constructed an instruction-tuning set comprising 4,475 cross-scale VQA samples derived from 93 WSIs, pairing each question's rationale with its cross-scale caption. All variants share the Patho-R1-7B backbone and are evaluated uniformly. Results on cross-scale (Table 1) and single-scale benchmarks (Table 2) reveal a clear dynamic: RL-only consistently outperforms both SFT-only and SFT+RL. We observe that SFT, while improving cross-scale accuracy compared to baseline, suffers a noticeable drop on single-scale benchmarks. This indicates a strict trade-off where forcing the model to mimic instruction-style rationales leads to catastrophic forgetting of its pre-trained single-scale representations. SFT+RL mitigates this degradation but still falls short of pure RL. This demonstrates that pure RL avoids the bottleneck of linguistic imitation: rather than training the model to reproduce human-written rationales, RL directly optimizes answer selection from cross-scale visual evidence via outcome rewards. Our observation aligns with *SFT Memorizes, RL Generalizes* [8], which similarly reports that SFT can overfit to demonstration traces while outcome-based RL yields stronger generalization. Together, these results demonstrate the strength of our RL-centric approach as

a more effective and annotation-efficient strategy for cross-scale pathology VQA in the low-data setting.

4 Conclusion

In conclusion, we present the first cross-scale training and evaluation paradigm for pathological VLMs, framing diagnosis as coordinated reasoning across magnifications. We release a high-quality dataset Scale-VQA, constructed with cross-scale captions and a leakage-aware VQA curation pipeline that suppresses text-only shortcuts and enforces multi-view evidence integration. We further introduce ScaleReasoner-R1, a data-efficient RL-trained model for cross-scale reasoning, achieving state-of-the-art performance on both multi-magnification and standard single-image pathology VQA benchmarks. These gains improve diagnostic robustness at both single and multiple scales, underscoring the practical value of cross-scale supervision.

Acknowledgments. This work was supported by the Ministry of Education Tier 1 grant, Singapore (24-1250-P0001), and the Ministry of Education Tier 2 grant, Singapore (T2EP20224-0028). This work was powered by the UnPuzzle & PuzzleCloud Platform (https://puzzlelogic.com/unpuzzle) and supported by PuzzleLogic Pte Ltd, Singapore.

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

References

1. Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M.D., Van der Laak, J., Bui, M.M., Vemuri, V.N., Parwani, A.V., Gibbs, J., Agosto-Arroyo, E., et al.: Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. The Journal of pathology **249**(3), 286–294 (2019)
2. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4971–4980 (2018). https://doi.org/10.1109/CVPR.2018.00522
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, K., Liu, M., Yan, F., et al.: Cost-effective instruction learning for pathology vision and language analysis. Nature Computational Science (2025). https://doi.org/10.1038/s43588-025-00818-5
5. Chen, P., Zhu, C., Zheng, S., Li, H., Yang, L.: Wsi-vqa: Interpreting whole slide images by generative visual question answering. In: European Conference on Computer Vision. pp. 401–417. Springer (2025)
6. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., , Ming, H., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. arXiv preprint arXiv:2410.11761 (2024)

7. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Hu, M., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5134–5143 (Jun 2025)
8. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
9. Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., et al.: Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. Nucleic acids research **44**(8), e71–e71 (2016)
10. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017). https://doi.org/10.1109/CVPR.2017.670
11. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
12. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3852–3861 (2020)
13. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
15. Liang, Y., Lyu, X., Chen, W., Ding, M., Zhang, J., He, X., Wu, S., Xing, X., Yang, S., Wang, X., et al.: Wsi-llava: A multimodal large language model for whole slide image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22718–22727 (2025)
16. Liao, D., Chen, S., Xi, N., Xue, Q., Li, J., Hou, L., Liu, Z., Low, C.H., Wu, Y., Liu, Y., Jiang, Y., Li, D., Lyu, S.: Unpuzzle: A unified framework for pathology image analysis (2025), https://arxiv.org/abs/2503.03152
17. Lu, M.Y., Chen, B., Williamson, D.F.K., Chen, R.J., Zhao, M., Chow, A.K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J.J., Gerber, G., Liang, I., Le, L.P., Parwani, A.V., Weishaupt, L.L., Mahmood, F.: A multimodal generative ai copilot for human pathology. Nature **634**(8033), 466–473 (Oct 2024). https://doi.org/10.1038/s41586-024-07618-3
18. Saygin Seyfioglu, M., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. arXiv e-prints pp. arXiv–2312 (2023)
19. Shrestha, R., Kafle, K., Kanan, C.: A negative case analysis of visual grounding methods for VQA. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 8172–8181 (2020). https://doi.org/10.18653/v1/2020.acl-main.727
20. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)

21. Sun, Y., Wu, H., Zhu, C., Zheng, S., Chen, Q., Zhang, K., Zhang, Y., Wan, D., Lan, X., Zheng, M., et al.: Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In: European Conference on Computer Vision. pp. 56–73. Springer (2024)
22. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
23. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
24. Xu, Z., Liu, Z., Hou, J., Ma, J., Jin, C., Wang, Y., Chen, Z., Zhang, Z., Huang, F., Guo, Z., et al.: A versatile pathology co-pilot via reasoning enhanced multimodal large language model. arXiv preprint arXiv:2507.17303 (2025)
25. Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Li, J., Chen, G., Wu, X., Zhang, Z., Xiao, Q., Wan, X., Wang, B., Li, H.: Huatuogpt, towards taming language models to be a doctor. arXiv preprint arXiv:2305.15075 (2023)
26. Zhang, W., Zhang, P., Guo, J., Cheng, T., Chen, J., Zhang, S., Zhang, Z., Yi, Y., Bu, H.: Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. arXiv preprint arXiv:2505.11404 (2025)
27. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., et al.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**(5), 236–245 (2019)