

QCAgent: An agentic framework for quality-controllable pathology report generation from whole slide image

Rundong Wang^{1,2†}, Wei Ba^{5†}, Yingtai Li^{1,2}, Ying Zhou^{1,2}, Bowen Liu¹, Baizhi Wang^{1,2}, Yuhao Wang^{1,2}, Zhidong Yang⁶, Kun Zhang^{1,2}, Rui Yan^{1,2,3,4} ✉, and Shaohua Kevin Zhou^{1,2,3,4} ✉

¹ School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China (USTC), Anhui, China

² Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), Suzhou Institute for Advanced Research, USTC, Jiangsu, China

³ Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Jiangsu, China

⁴ Biomedical Basic Research Center (BBRC) of Jiangsu, Jiangsu, China

⁵ Chinese PLA General Hospital Ninth Medical Center, Beijing, China

⁶ Hong Kong University of Science and Technology (HKUST), Hong Kong, China
yanrui@ustc.edu.cn, s.kevin.zhou@gmail.com

Abstract. Recent methods for pathology report generation from whole-slide images (WSIs) are capable of producing slide-level diagnostic descriptions but fail to ground fine-grained statements in localized visual evidence. Furthermore, they lack control over which diagnostic details to include and how to verify them. Inspired by emerging agentic analysis paradigms and the diagnostic workflow of pathologists, who selectively examine multiple fields of view, we propose **QCAgent**, an agentic framework for quality-controllable WSI report generation. The core innovations of this framework are as follows: (i) it incorporates a customized critique mechanism guided by a user-defined checklist specifying required diagnostic details and constraints; (ii) it re-identifies informative regions in the WSI based on the critique feedback and text-patch semantic retrieval, a process that iteratively enriches and reconciles the report. Experiments demonstrate that by making report requirements explicitly prompt-defined, constraint-aware, and verifiable through evidence-grounded refinement, QCAgent enables controllable generation of clinically meaningful and high-coverage pathology reports from WSI. Code is available at <https://github.com/cs-w0r1d/qcagent>.

Keywords: Whole Slide Image · Pathology Report Generation · Agent · Quality-controllability.

[†]Equal contribution. ✉Corresponding authors.

1 Introduction

Pathology foundation models [17,12,21,25,26] have significantly advanced whole slide image (WSI) analysis and its applications. Among its applications, WSI-to-Text generation promises to reduce the pathologist’s burden and enhance diagnostic consistency by generating descriptions [20,4,10,15]. Unlike general image captioning, however, a clinical standard report must not only provide a final diagnosis but also rigorously cover specimen context, microscopic morphological descriptors (e.g., nuclear pleomorphism), and critical predictors such as lympho-vascular invasion (LVI).

Due to the gigapixel scale of WSIs, current generative models (e.g., HistGen [10], SGMT [18] and PRISM [20]) predominantly follow a single-pass paradigm, leading to two intractable issues. First, feature collapse occurs: to maintain computational feasibility, the model compresses thousands of image patches into compact global vectors. This drastic dimensionality reduction, while capturing general lesions, inevitably erases spatially sparse but diagnostically decisive features, such as micro-lesions or evidence of LVI [5]. Second, linguistic prior bias arises: lacking a “visual look-back” mechanism, Large Language Models (LLMs) tend to hallucinate generic clinical commons (e.g., “cellular atypia”) based on diagnostic labels (e.g., “gastric cancer”) when global features are insufficient for detailed synthesis. Such details lack correlation with specific visual evidence from the current slide and do not address missing information required for clinical quality control, severely undermining the clinical credibility of reports [7].

To bridge the gap between global context and local evidence, we draw inspiration from how pathologists iteratively navigate between low-power screening and high-power verification. This workflow aligns with emerging agentic paradigms [13,19,24,23,29]. Although recent work has explored agents in computational pathology [22,9,3,2,11,16], agentic *on-demand back-tracking* for controllable, evidence-grounded report generation remains underexplored.

Motivated by this, we propose QCAgent, an agent-based, quality-controllable pathology report generation framework. First, the system employs a WSI foundation model, PRISM [20], to extract global semantic features, supplemented by representative patch information via a cluster-based random sampling strategy. Subsequently, it invokes a pathology-expert reasoning model, Patho-R1 [28], to generate an initial draft. Crucially, we introduce a quality control module that utilizes advanced LLMs [6,27,1] to simulate a pathologist’s auditing behavior. It identifies critical missing information, converts them into targeted retrieval instructions, and triggers Patho-R1 to revise the report. This process iterates until the report passes the QC criteria. This mechanism enables evidence-based report generation by explicitly aligning visual findings with clinical narratives. The main contributions of this paper are three-fold:

- **Iterative QC architecture.** We reformulate report generation as a multi-round *audit-retrieve-revise* process. A structured QCAgent automatically identifies non-compliant statements and missing clinical information, enabling closed-loop refinement.

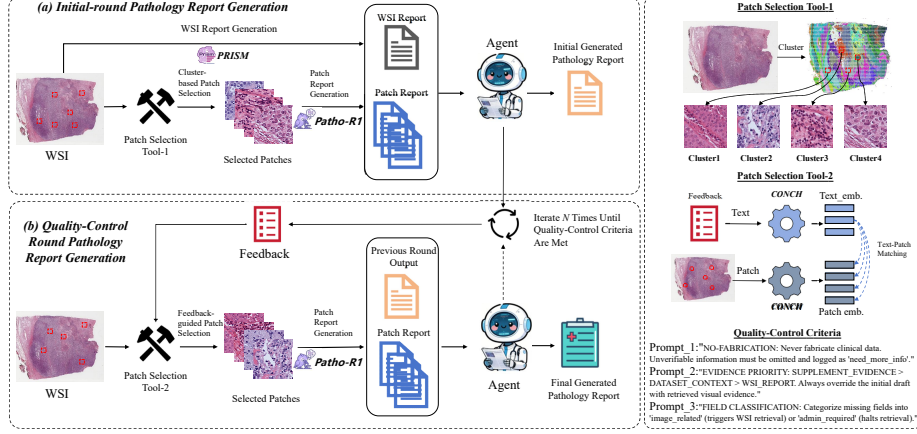


Fig. 1. Pipeline of QCAgent: (a) initial-round report generation using PRISM and cluster-based patch selection (Tool-1) with Patho-R1 patch reports; (b) QC-round iterative refinement using feedback-guided CONCH retrieval (Tool-2) and Patho-R1 evidence until QC criteria are met.

- **Demand-driven evidence retrieval.** We introduce a *text-guided evidence retrieval* mechanism. Instead of random sampling, the QCAgent converts missing fields into morphology-oriented queries to retrieve the most relevant visual patches from the WSI, grounding the report completion in targeted evidence.
- **Strict anti-hallucination design.** We adopt an evidence-priority scheme (supplemental visual evidence > initial draft) to reduce hallucinations. By explicitly separating verifiable from non-verifiable fields, only evidence-supported claims are retained in the final report.

2 Method

This section elaborates on the proposed QCAgent, an agent-based, quality-controllable framework for pathology report generation from whole-slide images (WSIs), as illustrated in Fig. 1. The overall computational pipeline of this framework consists of two core stages: (a) **Initial-round generation**, which aims to rapidly construct a WSI-level draft and extracts local visual evidence with global coverage; and (b) **Quality-Control (QC) rounds of iterative generation**, which translates information gaps within the report into explicit “evidence demands”. Through a closed-loop mechanism of “feedback-guided evidence retrieval” and “evidence-driven revision”, this stage progressively fulfills stringent quality control criteria.

Problem Formulation. Given a whole-slide image (WSI) S and a set of pre-extracted patch features $\mathcal{F} = \{f_1, \dots, f_N\}$ obtained by CONCH [14], together with an initial draft report r_0 produced by a foundation model, our goal is to

generate an informative and accurate pathology report r^* through T rounds of iterative refinement. We formulate the process as a multi-round optimization with a maximum of T iterations:

$$r^* = \text{QC-Iterate}(r_0, \mathcal{F}, S; T), \quad (1)$$

where each iteration comprises four consecutive stages: (i) quality-control (QC) assessment, (ii) evidence retrieval, (iii) visual diagnosis, and (iv) report revision.

We define *quality-controllability* as the capability of the system to explicitly follow and enforce a predefined set of quality control criteria $\mathcal{K}$ during inference. These criteria include, but are not limited to, a report schema, required fields, anti-hallucination rules, evidence-priority principles, and standardized writing templates. At each iteration, the system outputs structured QC results, enabling users to dynamically configure $(\mathcal{K}, T)$ to achieve a controllable trade-off between generation quality and computational cost. Under this formulation, “controllability” is not merely post-hoc scoring; instead, hard constraints are embedded into an iterative “audit–evidence seeking–revision” closed loop, yielding a transparent and auditable reasoning trace.

Concretely, our QC prompts implement three core rules:

- NO-FABRICATION:** Never fabricate clinical data. Any unverifiable information must be omitted from the report and explicitly logged as `need_more_info`.
- EVIDENCE PRIORITY:** When supplemental evidence conflicts with the initial draft, revisions follow the priority order `SUPPLEMENT_EVIDENCE > DATASET_CONTEXT > WSI_REPORT`.
- FIELD CLASSIFICATION:** Missing fields are categorized into `image_related` and `admin_required`.

Initial Round: Global Draft and Coverage-Oriented Evidence Extraction. As illustrated in Fig. 1(a), the initial round consists of two parallel streams:

1. **WSI-level report generation (PRISM).** We first employ a pretrained WSI-level foundation model to produce an initial draft report r_0 , which serves as a global narrative backbone for subsequent refinement.
2. **Cluster-based patch selection (Patch Selection Tool-1).** To provide representative *coverage-oriented* evidence in the initial round, we perform clustering in the feature embedding space over candidate patches and uniformly sample from each cluster to construct a representative patch set $\mathcal{P}^{(1)}$. This strategy ensures comprehensive initial information selection.
3. **Patch-level report generation (Patho-R1).** The selected patches $\mathcal{P}^{(1)}$ are then fed into the pathology vision-language model (VLM) Patho-R1 to generate patch-level descriptions $\mathcal{E}^{(1)}$, which serve as detailed evidence. In practice, Patho-R1 is executed in batches and produces fine-grained diagnostic cues, providing direct visual evidence to complement the WSI-level draft.

Finally, the agent fuses the global draft and local evidence to produce the initial generated report r'_0 , which is the starting point for the subsequent QC iterations.

Quality-Control Rounds: Feedback-Guided Evidence Seeking and Evidence-Driven Revision. As shown in Fig. 1(b), the QC rounds take the *previous-round report* together with accumulated evidence and the quality criteria as inputs, and perform closed-loop refinement with a maximum of T iterations.

At iteration t , the QCAgent conducts a fine-grained, structured audit of the previous report r_{t-1} and outputs: (i) a set of missing key fields $\mathcal{M}_t$, (ii) a set of morphology-driven retrieval queries $\mathcal{Q}_t$, and (iii) a revised draft r'_t . The query set $\mathcal{Q}_t$ follows a demand-driven principle (“retrieve what is missing”). For example, when lymphovascular invasion (LVI) is missing, the agent generates a morphology-oriented query such as “*Histopathological evidence for lymphovascular invasion in gastric adenocarcinoma.*” Meanwhile, the QCAgent enforces the **NO-FABRICATION** rule: for any field that cannot be supported by reliable visual evidence from the WSI, the system refuses to hallucinate and explicitly logs it as `need_more_info`.

Tool-2 performs text-guided retrieval in the CONCH’s embedding space:

1. **Text encoding:** Compute the normalized query embedding $q = \text{CONCH}_{\text{text}}(\mathcal{Q}_t)$;
2. **Similarity search:** Retrieve TopK patches via cosine similarity using a FAISS IndexFlatIP index, i.e., $\text{TopK}(\{q \cdot f_i\}_{i=1}^N)$;
3. **De-duplication:** Exclude patches already retrieved in previous rounds to reduce redundancy;
4. **Patch cropping:** Return the Top- K patch coordinates and similarity scores, and crop the corresponding high-resolution patches from the original WSI.

Compared with random sampling or global clustering alone, this demand-driven retrieval paradigm pinpoints tissue regions most relevant to the current QC feedback, substantially improving the efficiency of evidence acquisition.

The retrieved patches are then re-input to Patho-R1 to generate fine-grained descriptions, yielding a new set of supplemental evidence $\mathcal{E}_t$. This evidence is fed back into the subsequent QC loop as key supporting material for revision. The agent produces the revised report r_t by integrating $(r_{t-1}, \mathcal{E}_t, \mathcal{K})$. The iterative process terminates when any of the following conditions holds: (i) the report satisfies all QC criteria; (ii) the remaining missing fields are all non-image-evidenceable; or (iii) the iteration budget reaches T .

3 Experiments and Results

3.1 Datasets and Evaluation Metrics

Datasets. To assess generalization across clinical settings and languages, we evaluate our framework on two independent, large-scale cohorts of gastric cancer: **TCGA-STAD (public, English)**. This cohort contains high-resolution WSIs from stomach adenocarcinoma in The Cancer Genome Atlas (TCGA). For TCGA-STAD, we use the slide-level pathology reports from TITAN [8] as ground-truth. **In-house dataset (private, Chinese)**. This cohort consists of 524 H&E WSIs. Ground-truth is the corresponding Chinese clinical pathology reports, which are highly structured and concise, emphasizing factual statements.

Algorithm 1 Agent-Based Framework for Quality-Controllable**Input:** WSI $\mathbf{S}$, patch features $\mathcal{F}$, coordinates $\mathcal{C}$, quality criteria $\mathcal{K}$, max rounds T **Output:** Final generated report r^*

```

1:  $r_0 \leftarrow \text{PRISM}(\mathbf{S})$ 
2:  $\mathcal{P}^{(1)} \leftarrow \text{Tool-1\_ClusterSelect}(\mathbf{S})$ 
3:  $\mathcal{E}^{(1)} \leftarrow \text{PathoR1}(\mathcal{P}^{(1)})$ 
4:  $r'_0 \leftarrow \text{AgentFuse}(r_0, \mathcal{E}^{(1)}; \mathcal{K})$ 
5: for  $t = 1, \dots, T$  do
6:    $(\mathcal{M}_t, \mathcal{Q}_t, r'_t) \leftarrow \text{QC}(r_{t-1}, \mathcal{E}_{t-1}; \mathcal{K})$ 
7:   if  $\text{Pass}(\mathcal{M}_t, \mathcal{K})$  then
8:     break
9:   end if
10:   $\mathcal{P}_t^{(2)} \leftarrow \text{Tool-2\_CONCH-Retrieve}(\mathcal{Q}_t, \mathcal{F}, \mathcal{C})$ 
11:   $\mathcal{E}_t \leftarrow \text{PathoR1}(\mathcal{P}_t^{(2)})$ 
12:   $r_t \leftarrow \text{Revise}(r'_t, \mathcal{E}_t; \mathcal{K})$ 
13: end for
14: return  $r^* \leftarrow r_t$ 

```

Evaluation metrics. We report both traditional text overlap metrics (BLEU-1/4, ROUGE-L, METEOR) and a semantic similarity metric (BERTScore). We also introduce a clinically motivated *Field Recall*, which measures the proportion of clinically critical pathology fields successfully covered by the generated report.

Baseline and implementation details. We adopt PRISM, a publicly available WSI-level foundation model, as the initial-report baseline. To validate the effectiveness of iterative quality control, we compare the one-pass draft generated by PRISM (without QC) against the final output of QC-Agent.

Our setup employs the CONCH encoder for text retrieval (Tool-2), PathoR1 for visual generation, and Qwen3-VL-30B-A3B-Thinking as the core QC-Agent. The QC loop runs for up to $T = 3$ iterations. At each iteration, Tool-2 retrieves the top $K = 3$ patches for each QC query, with previously retrieved patches excluded to reduce redundancy.

3.2 Experimental Results

Results on TCGA-STAD (English). Compared with the PRISM baseline, our QC-Agent achieves consistent and substantial improvements across all metrics on TCGA-STAD (Table 1). Notably, Field Recall increases from 32.08% to 63.35%, indicating that the iterative QC process successfully recovers clinically important missing findings. In addition, we observe a clear increase in report completeness: PRISM reports are typically brief, while QC-refined reports provide more detailed and comprehensive descriptions.

Results on In-house Cohort (Chinese). We further evaluate cross-lingual transfer by directly applying our QC-Agent to the private Chinese cohort without any finetuning on Chinese pathology data. As shown in Table 1, the QC-Agent

Table 1. Quantitative results of WSI report generation on two datasets.

Dataset	Method	BLEU-1	BLEU-4	ROUGE-L	METEOR	FieldRecall	BERTScore
TCGA-STAD	PRISM (Baseline)	0.0986	0.0319	0.0947	0.0985	0.3208	0.1245
	QCAgent (Ours)	0.2857	0.0836	0.2784	0.4087	0.6335	0.2738
In-house	PRISM (Baseline)	0.0226	0.0083	0.0484	0.0333	0.1908	-0.0831
	QCAgent (Ours)	0.0924	0.0186	0.2490	0.2307	0.4863	0.2370

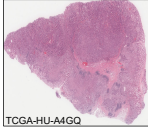
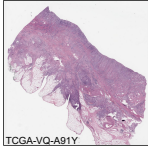
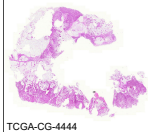
WSI	Ground-Truth	Prism	Ours
 TCGA-HU-MGO	The slide from the gastric antrum shows a moderately differentiated tubular adenocarcinoma (T4b) with infiltrative growth pattern. Lymphovascular and perineural invasion are present. Metastasis observed in 4 out of 49 inferior gastric lymph nodes.	Diagnosis: Moderately differentiated adenocarcinoma of the sigmoid colon with lymph node metastasis.	The slide from the stomach shows a gastric adenocarcinoma with mixed differentiation, including both moderately differentiated (G2) and poorly differentiated (G3) components, demonstrating an infiltrative growth pattern with marked nuclear pleomorphism, architectural disruption, and loss of cellular polarity. The tumor exhibits stromal invasion, desmoplastic stroma, increased mitotic activity, and mucinous differentiation in some areas. Additional features include peritumoral lymphocytic infiltrate, focal necrosis, and tumor budding. The exact tumor size, depth of invasion, lymphovascular and perineural invasion status, surgical margin status, lymph node metastasis counts, and TNM staging remain undetermined.
 TCGA-VQ-A91Y	The slide from the gastric antrum shows a tubular adenocarcinoma, histologic grade III (poorly differentiated), intestinal type (Lauren classification), with a largest tumor size of 9.0 cm. Features include ulceration, local invasion to the serosa, and evidence of neural, lymphatic, and blood vessel invasion.	Diagnosis: High-grade serous carcinoma of the ovary and fallopian tube.	The slide from the stomach shows a poorly differentiated adenocarcinoma with an infiltrative growth pattern, characterized by marked nuclear atypia, frequent mitoses, and stromal desmoplasia. The tumor cells form irregular glandular structures with luminal necrosis and loss of polarity, consistent with a high-grade malignant phenotype. Infiltration into the muscularis propria is confirmed by tumor cells invading between smooth muscle bundles, indicating deep wall invasion. A prominent desmoplastic stromal reaction is observed... Lymphovascular invasion is identified, and perineural invasion is present. The tumor demonstrates mucin production and is consistent with intestinal-type gastric adenocarcinoma (Lauren classification).
 TCGA-CG-4444	The slide from the body of the stomach shows a poorly differentiated adenocarcinoma, partially of the diffuse type (signet-ring carcinoma), with infiltration into the subserosa and multiple regional lymph node metastases (pT2b pN2).	Diagnosis: Metastatic adenocarcinoma with signet ring features in lung tissue.	The slide from the body of the stomach shows a poorly differentiated adenocarcinoma with a mixed mucinous and signet ring cell component, characterized by an infiltrative growth pattern. The tumor exhibits marked nuclear atypia, cytoplasmic mucin vacuoles, and abundant extracellular mucin pools... Lymphovascular invasion is identified, and perineural invasion is present. The tumor invades through the full thickness of the gastric wall, extending into the subserosal adipose tissue. The proximal and distal surgical margins are free of tumor. Metastatic carcinoma is present in multiple lymph nodes, with findings consistent with a mixed mucinous and signet ring cell variant of gastric adenocarcinoma.

Fig. 2. Qualitative comparison of pathology reports on TCGA-STAD. For each case, we show the WSI thumbnail, the ground-truth clinical report, the one-pass draft generated by PRISM, and our QC-refined report.

yields marked improvements over PRISM across overlap-based metrics and semantic similarity. We note that the Chinese ground-truth reports are highly condensed, often in phrase-like formats, while our system generates more detailed narrative paragraphs. This stylistic mismatch can depress n-gram matching metrics. Nevertheless, the improvement in BERTScore-F1 suggests better alignment at the semantic and factual level.

Qualitative comparison. QCAgent improves clinical completeness. Compared with PRISM draft, QCAgent adds morphology-grounded descriptions for previously missing fields and avoids unsupported claims by marking non-image-verifiable information as undetermined (Fig. 2); a representative step-by-step case study is shown in Fig. 3.

3.3 Ablation Study

To evaluate the contribution of key components (Clustering, Retrieval, and Patho-R1), we conducted an ablation study on 100 randomly selected WSIs

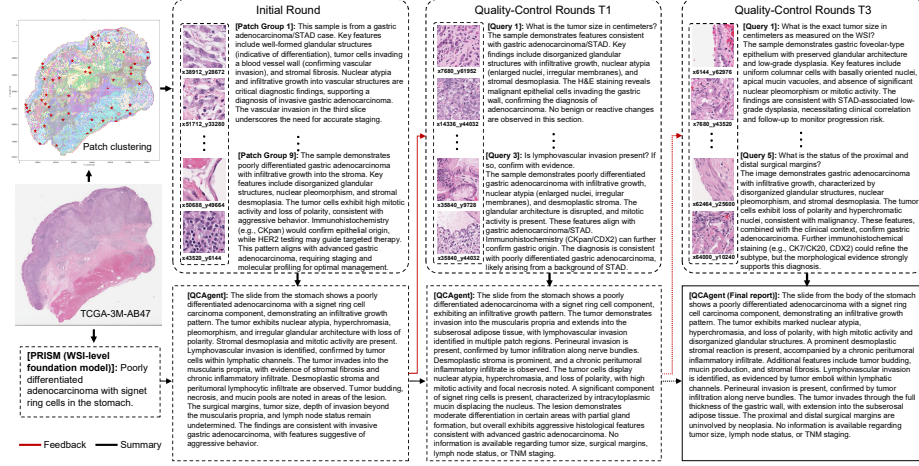


Fig. 3. Case study of the iterative audit-retrieve-revise workflow. Left: WSI thumbnail and patch clustering, with PRISM providing the initial global context. Middle/Right: across QC rounds, QCAgent issues missing-field queries, retrieves supporting patches, and revises the report, improving completeness while marking non-image-verifiable items as *undetermined*.

Table 2. Ablation study results using recall-centric metrics on 100 WSIs.

Configuration	AvgLen	R1-Recall	R2-Recall	RL-Recall	BS-Recall	FieldRecall
Full (Ours)	130	0.5706	0.3061	0.5139	0.3703	0.6690
w/o Retrieval	138	0.5561	0.2652	0.4809	0.3423	0.6440
w/o Patho-R1	98	0.5310	0.2696	0.4495	0.3986	0.7306

from the TCGA-STAD dataset. During preliminary evaluations, we observed a pronounced “length bias”. Specifically, the complete framework (Full) generates highly detailed reports. While clinically informative, this verbosity heavily penalizes precision-based metrics like F_1 -score. Conversely, configurations lacking fine-grained details achieve spuriously high F_1 -scores. To objectively measure the coverage of core pathological features, we employ recall-centric metrics for this analysis.

Ablation experiments confirm that each component is important (Table 2). (a) Removing Patho-R1 (visual feature extraction) module causes the most severe performance degradation alongside a significant reduction in report length. This demonstrates the critical role of Patho-R1 in extracting fine-grained histological features. (b) Removing the CONCH retrieval module reduces ROUGE-L Recall by 6.4%. This validates that dynamically retrieving local WSI regions based on missing fields effectively recovers omitted information.

4 Conclusion

We present an agentic, quality-controllable framework for WSI-based pathology report generation. Through an audit–retrieve–revise loop with cross-modal patch retrieval and evidence-driven revision, the QCAgent improves completeness while enforcing anti-hallucination constraints that distinguish verifiable findings from administrative items. On TCGA-STAD and an in-house gastric cancer cohort, our method consistently outperforms PRISM across overlap, semantic similarity, and Field Recall in a zero-shot, training-free setting. Future work will add expert evaluation and optimize clinical deployment.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62402473, 62502490, 62271465), Basic Research Program of Jiangsu (No. BK20255001), Natural Science Foundation of Jiangsu Province (BK20250496, BK20240464), Beijing Natural Science Foundation (L252175), and Suzhou Basic Research Program (SYG202338).

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

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025), <https://arxiv.org/abs/2511.21631>
2. Buckley, T.A., Weihrauch, K.R., Latham, K., Zhou, A.Z., Manrai, P.A., Manrai, A.K.: Navigating gigapixel pathology images with large multimodal models. arXiv preprint arXiv:2511.19652 (2025), <https://arxiv.org/abs/2511.19652>
3. Chen, J., Cai, L., Wang, Z., Huang, Y., Jiang, S., Huang, S., Wang, H., Zhang, Y.: PathAgent: Toward interpretable analysis of whole-slide pathology images via large language model-based agentic reasoning. arXiv preprint arXiv:2511.17052 (2025), <https://arxiv.org/abs/2511.17052>
4. Chen, P., Li, H., Zhu, C., Zheng, S., Shui, Z., Yang, L.: WsiCaption: Multiple instance generation of pathology reports for gigapixel whole-slide images (2024), <https://arxiv.org/abs/2311.16480>
5. Chen, P., Zhu, C., Zheng, S., Li, H., Yang, L.: WSI-VQA: Interpreting whole slide images by generative visual question answering (2024), <https://arxiv.org/abs/2407.05603>
6. Chen, Q., Hu, Y., Peng, X., Xie, Q., Jin, Q., Gilson, A., Singer, M.B., Ai, X., Lai, P.T., Wang, Z., et al.: Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications* **16**(1), 3280 (2025), <https://doi.org/10.1038/s41467-025-56989-2>
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 (June 2025)
8. 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.: A multimodal whole-slide foundation model for pathology. *Nature Medicine* pp. 1–13 (2025)

9. Ghezloo, F., Seyfioğlu, M.S., Soraki, R., Ikezogwo, W.O., Li, B., Vivekanandan, T., Elmore, J.G., Krishna, R., Shapiro, L.: PathFinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23431–23441 (2025)
10. Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H.: HistGen: Histopathology report generation via local-global feature encoding and cross-modal context interaction (2024), <https://arxiv.org/abs/2403.05396>
11. Huang, G., Chen, W., Yang, J., Lyu, X., Luo, X., Yang, S., Xing, X., Shen, L.: SurvAgent: Hierarchical cot-enhanced case banking and dichotomy-based multi-agent system for multimodal survival prediction (2025), <https://arxiv.org/abs/2511.16635>
12. Jaume, G., Oldenburg, L., Vaidya, A., Chen, R.J., Williamson, D.F., Peeters, T., Song, A.H., Mahmood, F.: Transcriptomics-guided slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9632–9644 (2024)
13. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al.: Agentbench: Evaluating LLMs as agents. In: International Conference on Learning Representations. vol. 2024, pp. 52989–53046 (2024)
14. 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.: A visual-language foundation model for computational pathology. *Nature Medicine* **30**(3), 863–874 (2024)
15. 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.: A multimodal generative AI copilot for human pathology. *Nature* **634**(8033), 466–473 (2024)
16. Lyu, X., Liang, Y., Chen, W., Ding, M., Yang, J., Huang, G., Zhang, D., He, X., Shen, L.: WSI-Agents: A collaborative multi-agent system for multi-modal whole slide image analysis. In: Proceedings of MICCAI 2025. vol. LNCS 15964. Springer Nature Switzerland (September 2025)
17. Ma, J., Guo, Z., Zhou, F., Wang, Y., Xu, Y., Li, J., Yan, F., Cai, Y., Zhu, Z., Jin, C., et al.: A generalizable pathology foundation model using a unified knowledge distillation pretraining framework. *Nature Biomedical Engineering* pp. 1–20 (2025)
18. Qin, W., Xu, R., Huang, P., Wu, X., Zhang, H., Luo, L.: What a whole slide image can tell? subtype-guided masked transformer for pathological image captioning (2023), <https://arxiv.org/abs/2310.20607>
19. Sapkota, R., Roumeliotis, K.I., Karkee, M.: AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Information Fusion* p. 103599 (2025)
20. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., Kunz, J.D., Retamero, J.A., Oakley, G., Klimstra, D., Kanan, C., Hanna, M., Zelechowski, M., Viret, J., Tenenholtz, N., Hall, J., Fusi, N., Yousfi, R., Hamilton, P., Moye, W.A., Vorontsov, E., Liu, S., Fuchs, T.J.: PRISM: A multi-modal generative foundation model for slide-level histopathology (2024), <https://arxiv.org/abs/2405.10254>
21. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
22. Sun, Y., Si, Y., Zhu, C., Zhang, K., Shui, Z., Ding, B., Lin, T., Yang, L.: CPathAgent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists’ diagnostic logic. *Advances in Neural Information Processing Systems* **38**, 101673–101731 (2026)

23. Truhn, D., Azizi, S., Zou, J., Cerda-Alberich, L., Mahmood, F., Kather, J.N.: Artificial intelligence agents in cancer research and oncology. *Nature Reviews Cancer* pp. 1–14 (2026)
24. Wang, H., Zhao, S., Qiang, Z., Xi, N., Qin, B., Liu, T.: Beyond direct diagnosis: LLM-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107* (2024), <https://arxiv.org/abs/2401.16107>
25. Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al.: A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**(8035), 970–978 (2024)
26. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**(8015), 181–188 (2024)
27. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025), <https://arxiv.org/abs/2505.09388>
28. 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 (2025), <https://arxiv.org/abs/2505.11404>
29. Zhao, W., Wu, C., Fan, Y., Qiu, P., Zhang, X., Sun, Y., Zhou, X., Zhang, S., Peng, Y., Wang, Y., et al.: An agentic system for rare disease diagnosis with traceable reasoning. *Nature* pp. 1–10 (2026)