Toward Interpretable Analysis of Whole-slide Pathology Images via Large Language Model-based Agentic Reasoning

Jingyun Chen[1⋆], Linghan Cai[1⋆], Zhikang Wang[2⋆], Yi Huang[2], Songhan Jiang[1], Shenjin Huang[1], Hongpeng Wang[1†], and Yongbing Zhang[1†]

[1] Harbin Institute of Technology, Shenzhen
[2] Fudan University
{jychen,cailh}@stu.hit.edu.cn

Abstract. Analyzing whole-slide pathology images (WSIs) requires an iterative, evidence-driven reasoning process that parallels how pathologists dynamically zoom, refocus, and self-correct while collecting evidence. However, existing computational pipelines often lack this reasoning trajectory, resulting in opaque and unjustifiable predictions. To bridge this gap, we present PathAgent, a training-free, large language model (LLM)-based agent framework that emulates the reflective, stepwise analytical approach of human experts. PathAgent can autonomously explore WSIs, iteratively and precisely locating significant micro-regions using the Navigator module, extracting morphological visual cues using the Perceptor, and integrating these findings into the continuously evolving natural language trajectories in the Executor. The entire sequence of observations and decisions forms a structured evidence trajectory, yielding traceable and evidence-grounded analyses. Evaluated across five challenging datasets spanning both whole-slide and patch-level settings, PathAgent exhibits strong zero-shot generalization and achieves performance comparable to human experts, surpassing task-specific baselines across diverse pathology tasks, including molecular subtype classification, histological grading and tumor type diagnosis. Human collaborative evaluations demonstrate that PathAgent produces diagnostic decisions highly consistent with expert-selected regions. These results demonstrate that PathAgent enables interpretable and clinically grounded WSI analysis. The code is available at `https://github.com/G14nTDo4/PathAgent`.

Keywords: Computational Pathology · Large Language Model · Visual Question Answering · Agentic System

1 Introduction

Analyzing whole-slide pathology images (WSIs), which span up to $100k \times 100k$ pixels, poses a significant challenge in computational pathology (CPath). Ideal

⋆Equal contribution. †Corresponding authors.

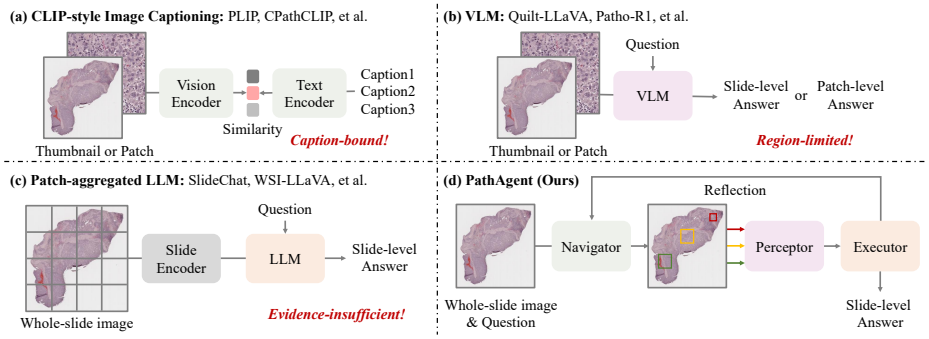

Fig. 1: Illustration of current multi-modal CPath models. (a) CLIP-style models are bound by pre-defined captioning templates. (b) VLMs are constrained by a limited receptive field. (c) Patch-aggregated LLMs lack traceable intermediate evidence. (d) PathAgent emulates the reflective, stepwise thinking of pathologists while performing analysis.

models must demonstrate three key abilities: identifying sparse diagnostic clues, navigating gigapixel-scale images, and reasoning across a continuously shifting field of view. Despite the emergence of numerous studies, these methods [4, 6, 7, 22, 25, 38, 42–44, 48] struggle to handle all three aspects simultaneously (Fig. 1). Specifically, CLIP-style visual-text alignment models [27, 33], such as PLIP [15], CPath-CLIP [30], demonstrate proficiency in retrieving relevant slide regions based on text queries; however, their dependency on static, pre-defined templates inherently limits their scalability. Visual-language models (VLMs), such as Patho-R1 [49] and PathReasoner-R1 [16], advance the field by generating open-ended descriptions. Yet, the vast pixel volume of WSIs rapidly saturates VLM context windows, necessitating compromises such as low-resolution thumbnails or patches that sacrifice crucial global context [1, 12, 17, 21, 28, 46]. Large language models (LLMs) possess exceptional capabilities for long-form reasoning and report generation. Nevertheless, in the field of pathology, their effectiveness is limited by an inherent lack of visual grounding, i.e., without direct access to morphological features from images. Although systems like SlideChat [8] and WSI-LLaVA [18] introduce a slide encoder to bridge visual inputs with LLM-driven text generation, they suffer from the lack of traceable evidence to validate their diagnostic reasoning. Moreover, despite exhibiting impressive task-specific capabilities, the incompatible architectures and inference processes of these models hinder the synthesis of their respective advantages.

Pathological analysis is fundamentally a goal-driven, iterative reasoning process. This workflow begins with a low-magnification survey of the WSI to retrieve initial regions of interest (RoIs). Subsequent steps involve progressively refined focus and targeted evidence retrieval at appropriate magnifications, guided by the specific diagnostic query. This cycle continues until diagnostic sufficiency is attained, allowing for the final biomedical conclusion. The distinctiveness of this

workflow from existing monolithic computational models [6, 8, 18] underscores the need for designing more scientifically rigorous and rational architectures.

Inspired by the above analysis, we propose a training-free, LLM-based agent framework, termed **PathAgent**, which is motivated by the iterative inspection workflow of pathologists. Specifically, we design a Navigator for RoI selection, a Perceptor for extracting morphological characteristics, and an Executor, the core of PathAgent, to provide guidelines and analytic logic in each iteration with Multi-Step Reasoning. The collaborative operation of these components generates traceable decisions and interpretable results for WSI analysis. Our contributions can be summarized in three aspects. (1) Dynamic Analytic Logic: We replace single-step reasoning with Multi-Step Reasoning in the Executor. This mechanism can construct analytic logic and dynamically provide guidelines to retrieve task-relevant information. (2) Adaptive Magnification: PathAgent can adaptively select an appropriate scale based on the analytic state, generating more refined visual evidence. (3) Enhanced Evidence Retrieval: We improve the accuracy of evidence capture by simplifying the query strategy of the Navigator.

We rigorously evaluated PathAgent on open-ended and closed-ended questions using five datasets. The experimental results demonstrate its superiority over existing methods. Case studies further reveal that the Multi-Step Reasoning framework provides PathAgent with a logical and robust thought process during analysis. Critically, by including pathologists in the reasoning process, PathAgent achieves improved accuracy and efficiency, further highlighting its scalability. In summary, PathAgent is the first training-free interactive agent specifically designed for WSI analysis. By coordinating off-the-shelf pathology models through an agent, it yields traceable decisions and competitive accuracy, suggesting a pragmatic route of CPath.

2 Related Work

2.1 Multi-Modal Large Language Models in CPath

Recent advances in multi-modal large language models for CPath have demonstrated significant potential in WSI analysis. CLIP-style models such as PLIP [15], MUSK [42], CONCH [22], CPath-CLIP [30] and PathFLIP [20] excel at retrieving patches via visual-text alignment but lack generative capabilities for traceable analytic logic. VLMs like Quilt-LLaVA [28] and Patho-R1 [49] address this by integrating visual evidence into open-ended descriptions. However, constrained by limited context windows, these models can only process thumbnails or local patches. To address this, SlideChat [8], WSI-LLaVA [18], and TITAN [9] attempt patch-aggregated LLMs but still fall short in providing sufficient visually grounded reasoning and comprehensive diagnostic planning. Consequently, while LLMs are increasingly adopted for holistic WSI analysis, their reasoning remains grounded solely in textual data without direct visual evidence, limiting trust and interpretability.

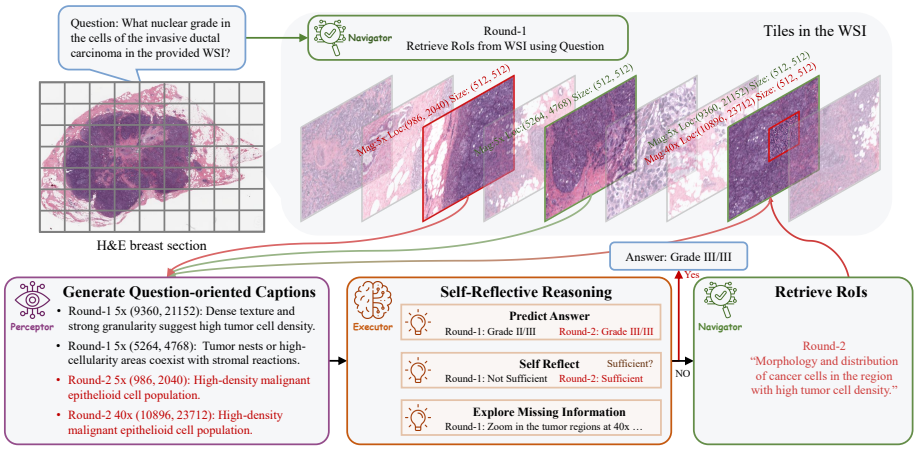

Fig. 2: Overview of the proposed PathAgent. Given an input WSI, PathAgent iteratively collects visual evidence and aggregates key information to generate traceable and evidence-grounded results. The process is completed by Navigator, Perceptor, and Executor, where the Executor serves as the central module orchestrating all analytic actions. "Mag" means magnification and "Loc" is the abbreviation for location.

2.2 Agentic Systems in CPath

To bridge this visual-grounding gap, the computer vision community has begun exploring LLM-based agent systems that actively plan, observe, and reason over visual inputs [10, 34, 37, 39–41, 50]. In CPath, early attempts integrate LLMs with external tools for WSI analysis [11, 14, 23, 25, 29, 36, 44]. For instance, the paradigm of Slideseek [5] and PathFinder [11] coordinates a Supervisor to plan exploration, Explorers to scan suspicious regions, and a Reporter to synthesize findings into diagnostic reports, thereby enabling autonomous WSI analysis. CPathAgent [31] mimics pathologists' multi-scale analytic logic through a three-stage training pipeline on navigation-reasoning instructions. Despite great progress, existing CPath agents mostly rely on extensively curated training data that is prohibitively expensive to acquire. Moreover, the majority of these approaches have not been open-sourced, which limits reproducibility and broader adoption. PathAgent addresses these challenges as a training-free agent that constructs a structured evidence trajectory, enabling traceable and evidence-grounded WSI analysis.

3 Methodology

We model the analytic process as a sequential, agentic workflow that mimics pathologists. As shown in Fig. 2, the workflow involves four steps implemented with three key components: (1) Navigator, a CLIP-based model, retrieves RoIs; (2) Perceptor, a CPath VLM, generates morphological descriptions; (3) Executor, as the core of the system, uses an LLM to provide guidelines, i.e., adjusting

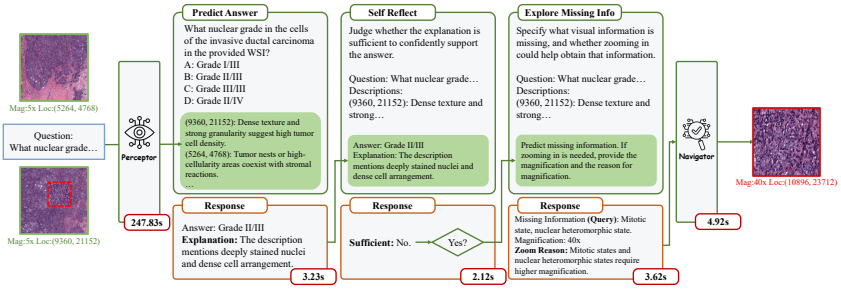

Fig. 3: Inference details of the Executor in PathAgent. The Executor executes answer prediction, self-reflection, and missing information exploration (if necessary) sequentially, providing guidelines for the entire analysis process. The mean runtime for each stage is shown in the red box below.

magnification for details or retrieving new RoIs using text queries (if necessary); and (4) the Executor integrates accumulated evidence to produce traceable and evidence-grounded analyses. Formally, the process is defined as a sequential trajectory represented by the analytic states (S_t), actions (A_t), and question-related visual findings (X_t): $\{(S_t, A_t, X_t)|1 \leq t \leq T\}$. Here, T denotes the maximum iteration budget, and the actual number of iterations is determined by the Executor through self-reflection. The following parts introduce each step in detail.

3.1 Obtaining Initial Analytic State

Given an input WSI and the targeted question q, the image is first decomposed into a set of downsampled patch instances $X = \{x_i\}_{i=1}^{N}$, where N denotes the total number of patches. For a single-query analysis, providing the textual descriptions of all the patches to the Executor at once is computationally and functionally infeasible. This would exceed the maximum context length of the Executor and bring redundant or interfering information. To address this, we design a Navigator module to filter and prioritize regions. Specifically, this module calculates the relevance score, r_i^1, between the visual embedding of each patch x_i and the question q, enabling the selection of the most relevant regions in the first analytic iteration. Based on the relevance scores given by the Navigator, we can select the top k_1 most relevant patches and define the initial question-related visual findings X_1:

$$X_1 = \{x_i | i \in \mathrm{Top}_{k_1}(\{r_i^1\}_{i=1}^{N})\}. \tag{1}$$

Next, each patch in X_1 together with a prompt "Please describe the pathology features in this image." is fed into the Perceptor to generate an initial morphological description $Des(x_i)$. Formally, the initial analytic state S_1 can be represented as follows:

$$S_1 = \{(M_1, Loc(x_i), Des(x_i)) \mid x_i \in X_1\}, \tag{2}$$

where M_1 represents the magnification scale of patches in the initial analytic iteration, $Loc(x_i)$ and $Des(x_i)$ indicate the coordinates (upper left corner) and textual description of the patch x_i, respectively. In addition, the Perceptor generates question-guided supplementary descriptions for the top 5 patches with the highest relevance scores. Specifically, the prompt: "Please describe the pathology features related to the question: [QUESTION] in this image." is fed into the Perceptor, where [QUESTION] is substituted with the targeted question. This procedure highlights desired morphological clues when constructing the analytic state.

3.2 Deciding the Next Analytic Action

After initializing the analytic state S_1, the analytic process proceeds iteratively, where each iteration follows the same reasoning structure. We therefore illustrate the general iteration process starting from iteration t. Given the current analytic state S_t that stores information from all previously examined patches, three possible actions A_t are available for the next iteration:

- **Action 1: Exploring for Additional Evidence (Sec. 3.3).** If the current analytic state S_t is insufficient to answer the question, and a reflection suggests that merely increasing the magnification level will not resolve the missing information, the Executor identifies what additional evidence is required to address the issue and explore for visual findings in other unchecked patches.
- **Action 2: Zooming in for Finer Details (Sec. 3.4).** If the current analytic state S_t indicates that higher magnification is required to answer the question, PathAgent conducts further examination at a higher magnification level to obtain supplementary evidence before making the final prediction.
- **Action 3: Concluding the Analysis (Sec. 3.5).** If the current analytic state S_t is sufficient to answer the question, PathAgent answers the question, provides an evidence-grounded rationale, and terminates the iteration.

To decide among the three actions, the Executor performs Multi-Step Reasoning. It first estimates the answer from the analytic state S_t and the question q, while generating reasoning logic as supporting evidence for its decision. Next, the Executor reflects on the question to evaluate whether the evidence is sufficient to support the estimated answer. Finally, we choose the action based on the response. This process is illustrated in Fig. 3.

3.3 Exploring for Additional Evidence

When Action 1 is triggered, PathAgent collects new visual evidence to update its analytic state S_t. The Executor is asked to provide guidelines on the required information I_t. Since the evidence from previously analyzed regions is preserved for the final analysis, the Navigator then retrieves missing information by computing a relevance score r_i^t for each patch x_i in the unexamined regions, where

Algorithm 1 PathAgent workflow

Require: Whole-slide image X, question q, Executor F_E, Perceptor F_P, Navigator F_N, max iteration T
Ensure: Prediction $\hat{Y}$, evidence-grounded rationale R, state-action-finding sequence $\{(S_t, A_t, X_t)|1 \le t \le T\}$

$X_1 \leftarrow \text{GuidedSample}(F_N, X, q)$
$S_1 \leftarrow \text{Describe}(F_P, X_1)$
for $t = 1$ **to** T **do**
  $\hat{Y}_t, \hat{R}_t \leftarrow \text{PredictAnswer}(F_E, S_t, q)$
  $C \leftarrow \text{SelfReflect}(F_E, S_t, q, \hat{Y}_t, \hat{R}_t)$
  if $C = \text{Yes}$ **then**
    $A_t = Action3$
    break
  else
    $I_t, Z, M_t \leftarrow \text{ExploreMissingInfo}(F_E, S_t, q)$
    if $Z = \text{Yes}$ **then**
      $A_t = Action2$
      $X_{mag} \leftarrow \text{Magnify}(X_t, M_t)$
      $x_{mag} \leftarrow \text{GuidedSample}(F_N, X_{mag}, I_t)$
      $S_t \leftarrow \text{Merge}(S_t, \text{Describe}(F_P, x_{mag}))$
      break
    else
      $A_t = Action1$
      $X_{t+1} \leftarrow \text{GuidedSample}(F_N, X, I_t)$
      $S_{t+1} \leftarrow \text{Describe}(F_P, X_{t+1})$
    end if
  end if
end for
$\hat{Y}, R \leftarrow \text{PredictAnswer}(F_E, \text{Merge}(S_1, \ldots, S_t), q)$
return $\hat{Y}, R, \{(S_t, A_t, X_t)|1 \le t \le T\}$

the relevance is measured between the visual embeddings and the instruction I_t. After obtaining the relevance score r_i^t, we can use the top k_t relevant patches to update the question-related visual findings X_{t+1}:

$$X_{t+1} = \{x_i | i \in \text{Top}_{k_t}(\{r_i^t\}_{x_i \in X \setminus \bigcup_{k=1}^{t} X_k})\}. \tag{3}$$

The coordinates of the selected patches and their corresponding descriptions are utilized to update the analytic state:

$$S_{t+1} = \{(M_{t+1}, Loc(x_i), Des(x_i)) | x_i \in X_{t+1}\}. \tag{4}$$

3.4 Zooming in for Finer Details

When Action 2 is selected in iteration t, the patches X_t are further divided according to the magnification level M_t specified by the Executor. Each region is partitioned into non-overlapping finer patches, forming a new image set X_{mag}.

Considering that generating descriptions for all magnified patches at higher magnification levels may introduce redundant information that interferes with

Table 1: Zero-shot visual question answering comparisons between PathAgent and current state-of-the-art methods on SlideBench-VQA (BCNB), WSI-VQA, and PathVQA datasets. The best performance is in **bold**, while the second-best performance is underlined. Differences from previous methods are in red by default.

Method (Input)	SlideBench-VQA (BCNB)						WSI-VQA	PathVQA
	Tumor Type	Receptor Status	HER2 Expression	Histological Grading	Molecular Subtype	Accuracy	Accuracy	Accuracy
General domain VLMs								
Qwen3-VL (Thumbnail)	41.48	50.63	19.37	25.95	11.03	30.70	21.84	-
Qwen3-VL (Patch)	43.96	53.26	21.35	29.74	16.04	38.22	23.47	44.82
GPT-4o (Thumbnail)	0.00	0.00	0.00	0.00	0.00	0.00	11.05	-
GPT-4o (Patch)	34.69	59.32	23.95	28.63	23.15	38.94	27.57	46.59
Medical domain VLMs								
LLaVA-Med (Thumbnail)	0.01	0.01	0.00	0.00	0.00	0.01	13.20	-
LLaVA-Med (Patch)	23.95	42.52	23.72	18.99	15.05	30.10	21.55	27.78
MedDr (Thumbnail)	28.92	48.08	20.65	29.96	23.88	35.48	43.69	-
MedDr (Patch)	45.46	40.44	22.73	30.28	15.49	33.67	45.42	29.35
Quilt-LLaVA (Thumbnail)	67.41	55.33	15.97	22.89	16.27	41.55	27.53	-
Quilt-LLaVA (Patch)	77.14	56.46	23.18	18.23	19.82	44.43	30.17	20.76
Patch-aggregated Methods								
WSI-VQA (Slide)	3.90	40.82	10.53	30.00	0.00	23.35	46.90	33.59
SlideChat (Slide)	90.17	73.09	25.05	23.11	17.49	54.14	-	55.03
WSI-LLaVA (Slide)	85.32	60.89	24.75	46.28	29.20	52.68	54.63	54.97
Agentic Systems								
GIANT (Slide)	57.37	42.55	14.79	48.52	26.63	44.97	47.05	-
PathAgent-offline	87.52	61.33	25.34	55.95	30.21	55.72	56.32	58.36
PathAgent-online	**91.31**(+1.14)	**75.87**(+2.78)	**27.26**(+2.21)	**56.21**(+9.93)	**35.88**(+6.68)	**59.31**(+5.17)	**60.74**(+6.11)	**62.23**(+7.20)

the final prediction, we utilize the Navigator for patch selection among X_{mag}. Next, the Perceptor is applied to the selected patch x_{mag} at the location $Loc(x_{mag})$ to generate a supplementary description $Des(x_{mag})$. The above processes enable PathAgent to collect question-relevant information. The analytic state S_t can be updated as follows:

$$S_t = S_t \cup \{(M_t, Loc(x_{mag}), Des(x_{mag}))\}, \tag{5}$$

which is then fed into the Executor, where Action 3 produces the final answer.

3.5 Concluding the Analysis

When Action 3 is selected in iteration t, the Executor integrates the analytic states from all iterations $(S_1, \ldots, S_{t-1})$ together with the accumulated reasoning traces. Before generating the final answer, the Executor summarizes the accumulated analytic states when the textual context becomes excessively long, retaining only question-relevant diagnostic evidence, including morphological findings, cellular patterns, tissue organization and intermediate reasoning conclusions. By jointly analyzing aggregated evidence with the question q, PathAgent performs a comprehensive synthesis that unifies textual descriptions and prior inferences across different magnification levels. This procedure enables PathAgent to generate a coherent analytic conclusion accompanied by a detailed reasoning explanation, providing an evidence-grounded rationale that supports the final answer. The overall procedure of PathAgent is summarized in Algorithm 1.

Table 2: Comparison of zero-shot visual question answering between our PathAgent and current state-of-the-art methods on the PathMMU dataset. The best performance is in **bold**, while the second-best performance is underlined.

Method	PathMMU					
	PubMed	SocialPath	EduContent	Atlas	PathCLS	Accuracy
Random Choice	22.13	25.48	25.49	19.66	15.31	22.14
Expert performance	72.85	71.46	68.95	68.32	78.85	71.82
General domain VLMs						
Qwen3-VL [1]	44.63	49.12	44.31	45.88	27.67	45.82
GPT-4o [46]	50.54	53.64	50.12	55.29	23.47	49.31
Medical domain VLMs						
LLaVA-Med [17]	27.69	27.36	27.18	30.73	20.31	26.25
MedDr [12]	21.37	25.34	32.95	31.45	22.17	28.33
Quilt-LLaVA [28]	42.63	46.68	45.31	42.79	29.28	41.54
Patch-aggregated Methods						
WSI-VQA [6]	32.69	35.93	34.63	50.37	21.38	33.64
SlideChat [8]	50.49	55.64	50.33	56.97	25.31	50.82
WSI-LLaVA [18]	47.35	53.26	47.22	53.81	24.87	49.57
Agentic Systems						
PathAgent-offline	51.26	57.71	51.46	59.39	48.53	53.19
PathAgent-online	**73.88**(+23.34)	**71.53**(+15.89)	**75.29**(+24.96)	**76.22**(+19.25)	**68.27**(+38.99)	**72.54**(+21.72)

4 Experiments

4.1 Implementation Details

In the experiments, we use the Trident [24, 35, 47] tool to crop the input WSI. Each WSI is decomposed into non-overlapping patches of size 512×512 at the $5\times$ magnification level, which also serves as the initial analytic magnification scale M_1. The maximum number of iterations T is set to 5 as a safe computational budget, and the numbers of patches selected by the Navigator in the first iteration and subsequent iterations are $k_1 = \lceil 0.1N \rceil$ and $k_t = \lceil 0.05N \rceil$, respectively. In PathAgent-offline, PLIP [15], Patho-R1 [49], and Qwen3-4B [45] are employed as the Navigator, Perceptor, and Executor, respectively, if not specifically noted. For PathAgent-online, the Executor is replaced with Qwen3-235B-A22B [45] using the Qwen API.

4.2 Datasets and Metrics

We evaluate PathAgent on five visual question answering (VQA) benchmarks: SlideBench-VQA [8] is a slide-level dataset with closed-ended questions; WSI-VQA [6] and WSI-Bench [18] (we evaluate the morphological analysis and diagnosis subsets) are two slide-level datasets with both close- and open-ended questions; PathMMU [32] and PathVQA [13] are two RoI-level datasets with closed-ended and open-ended questions, respectively. For the closed-ended visual question answering task, accuracy is used as the evaluation metric, while for the open-ended visual question answering task, BLEU [26], METEOR [2], and ROUGE [19] are adopted. Following [18], we additionally report WSI-Precision

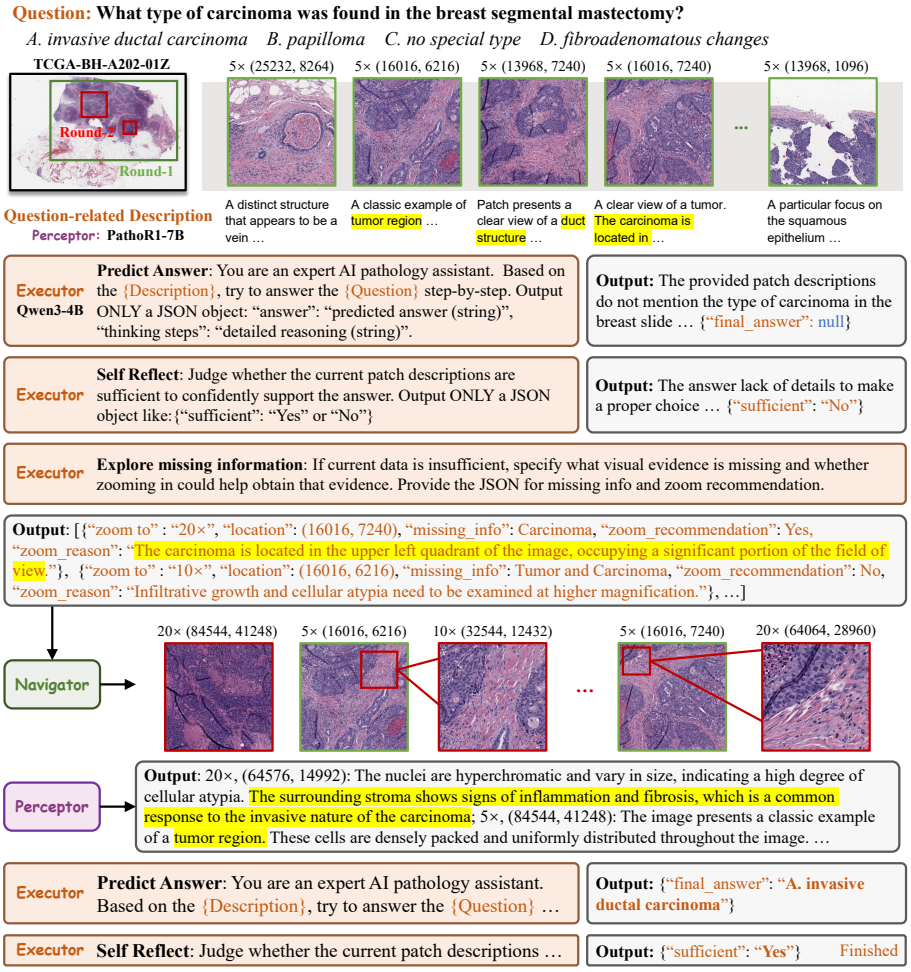

Fig. 4: Case study of a closed-ended question from WSI-VQA. PathAgent progressively retrieves task relevant patches and identifies key evidence for the diagnosis. Yellow highlights indicate the critical cues discovered during the reasoning process.

and WSI-Relevance to account for synonymous diagnostic expressions that may not exactly match the ground-truth. All metrics are presented in percentages in the experiments. Dataset statistics and evaluation details are provided in the Supplementary Materials.

4.3 Quantitative Evaluation

We compare PathAgent with nine advanced models in zero-shot VQA scenarios, including VLMs (Qwen3-VL [1], MedDr [12], GPT-4o [46], LLaVA-Med [17],

Table 3: Ablation study on the Executor and Perceptor in PathAgent-offline. Here, the WSI-VQA dataset includes both open- and closed-ended questions. The best performance is in **bold**, while the second-best performance is underlined.

Executor	Perceptor	WSI-VQA				SlideBench-VQA (BCNB)
		BLEU-1	BLEU-4	METEOR	Accuracy	Accuracy
WSI-VQA method		34.21	19.65	21.12	46.90	23.35
Qwen3-4B	Quilt-LLaVA	44.53	43.46	25.03	52.95	49.33
	Patho-R1-3B	48.17	46.79	29.61	53.67	51.43
	Patho-R1-7B	52.89	49.69	31.29	56.32	54.72
Qwen3-32B	Quilt-LLaVA	45.35	44.41	25.70	51.58	50.47
	Patho-R1-3B	52.46	50.47	31.50	54.77	53.26
	Patho-R1-7B	**55.58**	**52.76**	**34.10**	**57.29**	**55.02**

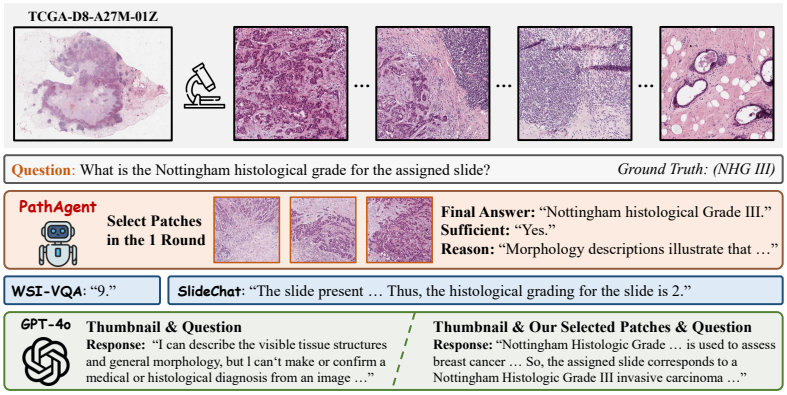

Fig. 5: Case study of an open-ended question in WSI-VQA. PathAgent retrieves task-relevant RoIs and answers correctly, while WSI-VQA and SlideChat fail. Providing the selected patches also enables GPT-4o to produce the correct response.

Quilt-LLaVA [28]), patch-aggregated methods (WSI-VQA [6], WSI-LLaVA [18], SlideChat [8]), and GIANT [3]. Since VLMs have constraints on image scale, we evaluate them using the following strategies [5, 8]: (1) "Thumbnail" adjusts the size of the WSI to 1024×1024 as input. (2) "Patch" randomly selects 30 patches from each WSI at $20\times$ and aggregates predictions via majority voting. Tab. 1 and Tab. 2 list the quantitative results, showing that PathAgent achieves the best average performance across the four zero-shot VQA datasets, indicating that PathAgent has advantages over the slide-based models and the general-purpose VLMs. Results on WSI-Bench are provided in the Supplementary Materials.

On the SlideBench-VQA (BCNB), general-purpose VLMs like Qwen3-VL and GPT-4o demonstrate impressive zero-shot performance but remain inferior to WSI-specific methods (SlideChat and WSI-LLaVA). This gap reflects the challenge of capturing fine-grained spatial and morphological information from WSIs using standard VLM pipelines. In contrast, PathAgent retrieves task-relevant RoIs and achieves 55.72% overall accuracy in the PathAgent-offline setting, out-

Table 4: Ablation study on Multi-Step Reasoning and Navigator use on the SlideBench-VQA (BCNB) dataset. The best performance is in **bold**, while the second-best performance is underlined.

Multi-Step Reasoning	Navigator	SlideBench-VQA (BCNB)						
		Tumor Type	Receptor Status	HER2 Expression	Histological Grading	Molecular Subtype	Accuracy	Average Iterations
		38.46	52.38	20.35	39.42	20.19	41.62	1.83
✓		52.43	53.89	23.13	44.16	26.28	44.97	1.59
	✓	78.88	58.27	23.40	45.46	27.42	50.68	1.32
✓	✓	**87.52**	**61.33**	**25.34**	**55.95**	**30.21**	**55.72**	**1.21**

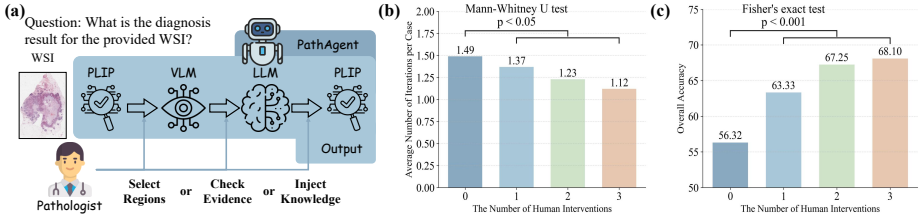

Fig. 6: Human collaborative experiment. (a) Pathologists can intervene during the reasoning process. (b, c) Effect of the number of human interventions on the average iterations per case and overall accuracy on WSI-VQA.

performing all baseline methods. With a stronger Executor, PathAgent-online further improves the overall accuracy to 59.31%. However, improvements are limited on HER2 Expression and Histological Grading. This observation is expected, as these tasks in clinical practice often require immunohistochemical testing rather than relying solely on morphological features.

On the PathMMU dataset, the performance gap among general-purpose VLMs, medical domain VLMs, and patch-aggregated methods becomes smaller because the dataset provides patch-level inputs rather than gigapixel WSIs. Under this setting, models can directly access local diagnostic information without severe resolution compression. Despite this, PathAgent still achieves the best performance. PathAgent-offline reaches 53.19% overall accuracy, outperforming SlideChat (50.82%) and GPT-4o (49.31%). PathAgent-online further improves the accuracy to 72.54%, surpassing the expert-level performance of 71.82%. At the task level, PathAgent-online achieves the highest scores on most subsets. The only exception occurs in PathCLS, where the limitation mainly arises from the Navigator module rather than the Executor.

4.4 Qualitative Evaluation

Fig. 4 illustrates the reasoning process of PathAgent on a WSI-VQA case. In iteration 1, the patch description at 5× magnification captures only the duct structure and carcinoma location, without providing sufficient evidence about

the cancer type. PathAgent therefore determines that the current information is insufficient. Based on the missing-information analysis, the Navigator is guided to potential cancer regions, and the model requests higher magnification to examine invasive growth and cellular atypia. In iteration 2, PathAgent retrieves a stromal region showing inflammation and fibrosis, which provides key evidence supporting the diagnosis of invasive ductal carcinoma and leads to the answer.

As shown in Fig. 5, for an open-ended question, PathAgent correctly answers in the first iteration using the selected RoIs. In contrast, WSI-VQA and SlideChat produce incorrect predictions. GPT-4o also generates the correct answer when provided with the patches selected by PathAgent, suggesting that PathAgent effectively selects question-relevant RoIs and can enhance the performance of general-purpose VLMs.

4.5 Ablation Study

As shown in Tab. 3, the performance of PathAgent-offline improves consistently when stronger foundation models are adopted. The combination of Qwen3-32B and Patho-R1-7B achieves the best results, reaching an accuracy of 57.29% on WSI-VQA. These results demonstrate that PathAgent effectively leverages stronger foundation models to improve reasoning and prediction accuracy. Additional ablation on the Navigator is provided in the Supplementary Materials.

We evaluate the effects of the Multi-Step Reasoning Executor and the Navigator. The baseline replaces Multi-Step Reasoning with single-step reasoning and substitutes the Navigator with uniform random patch sampling. As shown in Tab. 4, removing both modules yields 41.62% overall accuracy with 1.83 average iterations. Adding Multi-Step Reasoning alone increases accuracy to 44.97% and reduces the average iterations to 1.59. Using the Navigator alone further improves accuracy to 50.68% and reduces the average iterations to 1.32, indicating that the Navigator mainly contributes to efficient RoI retrieval. Task-level results further reveal the roles of each component. The Navigator yields notable improvements on Receptor Status, raising the accuracy from 52.38% to 58.27%, suggesting that effective RoI selection helps capture discriminative regions. In comparison, Multi-Step Reasoning shows greater benefits for Molecular Subtype, where the accuracy increases from 20.19% to 26.28%. This task often depends on implicit morphological patterns and their associated clinical knowledge, making Multi-Step Reasoning more effective than single-step inference. The best performance is achieved when both modules are combined, reaching 55.72% overall accuracy with 1.21 average iterations.

We further analyze the effect of the number of patches provided to the Perceptor. Fig. 8 shows that increasing the number of patches provided to the Perceptor does not consistently improve performance. The accuracy peaks at a 10% patch sampling ratio and then gradually declines, as redundant information interferes with the decision process. This observation highlights the importance of the Navigator in selecting informative regions efficiently. Therefore, we use 10% of patches in the initial iteration.

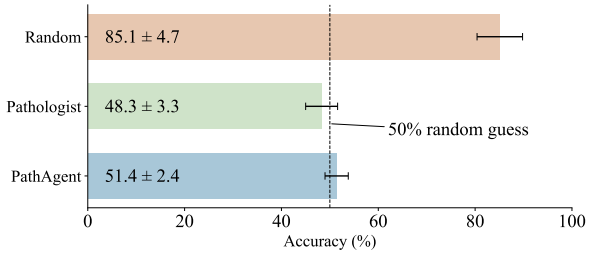

Fig. 7: Visual Turing test on SlideChat-VQA (BCNB). The dashed line marks the ideal indistinguishable level.

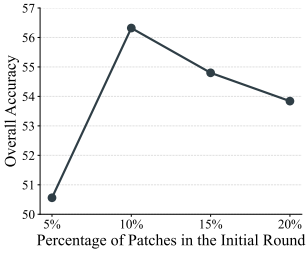

Fig. 8: Initial patch proportion test on WSI-VQA.

4.6 Diagnosis in Collaboration with Pathologists

Acknowledging the heterogeneity of diseases, the responses of the agent system are inevitably limited. Timely human intervention helps promptly correct the diagnostic reasoning and improve the prediction accuracy. We therefore conduct a human collaborative experiment with two pathologists, each with 5 years of experience, on 80 questions. As shown in Fig. 6 (a), the entire analytic process of PathAgent is fully transparent to pathologists, allowing them to intervene at any inference step, including selecting RoIs manually, reviewing VLM descriptions, supplementing missing information, or adjusting image magnification. Experimental results (Fig. 6 (b, c)) show that increasing interventions improve accuracy while reducing the average number of examined iterations. To further quantify this effect, we group the cases into no intervention (0) and intervention ($\geq$1). The Mann–Whitney U test compares the distributions of iteration counts, and Fisher's exact test evaluates differences in accuracy. The results statistically confirm that human intervention significantly enhances both diagnostic accuracy and decision efficiency, showcasing the strong potential of PathAgent in interactive clinical scenarios.

We conduct a visual Turing test on the RoIs selected by PathAgent involving 22 human judges. In each trial, judges identify which of two RoI sets is selected by the machine. As shown in Fig. 7, the identification accuracy for the random vs pathologist condition reaches 85.1%, indicating that randomly selected RoIs can be easily recognized by human judges. In contrast, the accuracies for PathAgent vs pathologist and pathologist vs pathologist are 48.3% and 51.4%, respectively, both close to the ideal level of 50.0%. These results suggest that the RoIs selected by PathAgent are largely indistinguishable from those selected by pathologists.

4.7 Task Complexity and Reasoning Mechanism Analysis

To further analyze the reasoning mechanism of PathAgent, we examine the average number of iterations required for different question types. As shown in Fig. 9, the average number of iterations generally increases with the clinical task difficulty, indicating that more complex problems require additional rounds of evidence gathering and reasoning, consistent with clinical practice. An exception

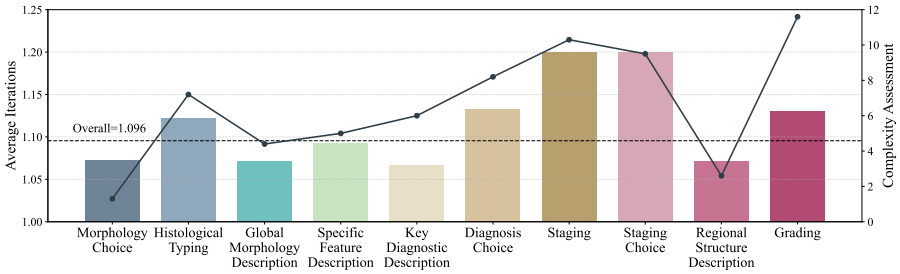

Fig. 9: Average attempts and task complexity across task types on WSI-Bench.

is observed in grading tasks: although tumor grading is clinically challenging, the model typically requires fewer iterations. This is because grading decisions mainly rely on localized morphological features, such as nuclear atypia and mitotic figures, which typically require lower recognition cost.

5 Conclusion

In this work, we introduce PathAgent, a training-free large language model-based agent that mirrors the reflective, stepwise thinking of human experts in computational pathology. Through iterative analysis, PathAgent retrieves RoIs and aggregates visual evidence for accurate and interpretable WSI analysis. Extensive experiments across multiple datasets demonstrate strong reasoning capability and remarkable generalization. Ablation studies and the human collaborative experiment highlight that PathAgent is an evolving intelligent agent with broad clinical applications. Moreover, reasoning mechanism analysis and the visual Turing test show that the RoIs selected by PathAgent are visually indistinguishable from those of pathologists, and that the required reasoning iterations across tasks follow the difficulty of clinical diagnosis, indicating strong alignment with expert diagnostic behavior.

Acknowledgements

We would like to acknowledge the support from the National Natural Science Foundation of China (No. 62331011).

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)

2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. 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)
4. Cai, L., Huang, S., Zhang, Y., Lu, J., Zhang, Y.: Attrimil: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. Medical Image Analysis p. 103631 (2025)
5. Chen, C., Weishaupt, L.L., Williamson, D.F., Chen, R.J., Ding, T., Chen, B., Vaidya, A., Le, L.P., Jaume, G., Lu, M.Y., et al.: Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv preprint arXiv:2506.20964 (2025)
6. 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 (2024)
7. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine (2024)
8. 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 Computer Vision and Pattern Recognition Conference. pp. 5134–5143 (2025)
9. 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)
10. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model (2023), https://arxiv.org/abs/2303.03378
11. Ghezloo, F., Seyfioglu, 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. arXiv preprint arXiv:2502.08916 (2025)
12. He, S., Nie, Y., Wang, H., Yang, S., Wang, Y., Cai, Z., Chen, Z., Xu, Y., Luo, L., Xiang, H., et al.: Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. arXiv preprint arXiv:2404.15127 (2024)
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. 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. arXiv preprint arXiv:2511.16635 (2025)
15. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
16. Jiang, S., Liu, F., Wang, Z., Cai, L., Zhang, Y.: Pathreasoner-r1: Instilling structured reasoning into pathology vision-language model via knowledge-guided policy optimization. arXiv preprint arXiv:2601.21617 (2026)

17. 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)
18. 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)
19. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
20. Liu, F., Jiang, S., Cai, L., Wang, Z., Zhang, Y.: Pathflip: Fine-grained language-image pretraining for versatile computational pathology. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 7132–7140 (2026)
21. Liu, T., Huang, T., Ding, T., Wu, H., Humphrey, P., Perincheri, S., Schalper, K., Ying, R., Xu, H., Zou, J., et al.: Leveraging multi-modal foundation models for analysing spatial multi-omic and histopathology data. Nature Biomedical Engineering pp. 1–18 (2026)
22. 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**, 863–874 (2024)
23. 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)
24. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)
25. 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. arXiv preprint arXiv:2507.14680 (2025)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
28. Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: 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. pp. 13183–13192 (2024)
29. Shaikovski, G., Vorontsov, E., Casson, A., Viret, J., Zimmermann, E., Tenenholtz, N., Wang, Y.K., Bernhard, J.H., Godrich, R.A., Retamero, J.A., et al.: Prism2: Unlocking multi-modal general pathology ai with clinical dialogue. arXiv preprint arXiv:2506.13063 (2025)
30. Sun, Y., Si, Y., Zhu, C., Gong, X., Zhang, K., Chen, P., Zhang, Y., Shui, Z., Lin, T., Yang, L.: Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10360–10371 (2025)
31. 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. arXiv preprint arXiv:2505.20510 (2025)
32. 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)
33. Sun, Y., Zhang, Y., Si, Y., Zhu, C., Shui, Z., Zhang, K., Li, J., Lyu, X., Lin, T., Yang, L.: Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. arXiv preprint arXiv:2407.00203 (2024)
34. Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: Medagents: Large language models as collaborators for zero-shot medical reasoning. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 599–621 (2024)
35. Vaidya, A., Zhang, A., Jaume, G., Song, A.H., Ding, T., Wagner, S.J., Lu, M.Y., Doucet, P., Robertson, H., Almagro-Perez, C., et al.: Molecular-driven foundation model for oncologic pathology. arXiv preprint arXiv:2501.16652 (2025)
36. Wang, S., Wu, R., Herndon, C., Liu, Y., Koga, S., Shen, J., Huang, Z.: Pathology-cot: Learning visual chain-of-thought agent from expert whole slide image diagnosis behavior. arXiv preprint arXiv:2510.04587 (2025)
37. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)
38. 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)
39. Wang, Z., Cai, L., Low, C.H., Liu, H., Wu, J., Wang, J., Wang, R., Song, L., Bian, J., Fu, J., et al.: 3dmedagent: Unified perception-to-understanding for 3d medical analysis. arXiv preprint arXiv:2602.18064 (2026)
40. Wang, Z., Wu, J., Cai, L., Low, C.H., Yang, X., Li, Q., Jin, Y.: Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. arXiv preprint arXiv:2503.18968 (2025)
41. Xia, P., Wang, J., Peng, Y., Zeng, K., Wu, X., Tang, X., Zhu, H., Li, Y., Liu, S., Lu, Y., et al.: Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. arXiv preprint arXiv:2506.00555 (2025)
42. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al.: A vision–language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)
43. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)
44. 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)
45. 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)
46. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 (2023)

47. Zhang, A., Jaume, G., Vaidya, A., Ding, T., Mahmood, F.: Accelerating data processing and benchmarking of ai models for pathology. arXiv preprint arXiv:2502.06750 (2025)
48. Zhang, K., Xu, H., Wang, S.: Pathreasoning: A multimodal reasoning agent for query-based roi navigation on whole-slide images. arXiv preprint arXiv:2511.21902 (2025)
49. 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)
50. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning. pp. 2165–2183. PMLR (2023)