PATHOARGUS: ADVANCING COMPREHENSIVE, EVIDENCE-GROUNDED VISUAL REASONING FROM GIGAPIXEL SLIDES TO MULTI-SLIDE CASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide pathology question answering requires finding relevant tissue and answering from visual evidence, yet textual cues can also yield correct answers. We introduce the PATHOARGUS dataset to evaluate six pathology capabilities and test whether answers follow controlled changes in target evidence. It contains 22,078 four-choice questions covering 4,809 target cases across 15 TCGA projects, including 4,281 test questions in PATHOARGUS-BENCH. Evidence-Set Grounding (ESG) forms counterfactual quartets by placing a target in each of three candidate sets or removing it; quartet exact-match accuracy (QExact) requires all four answers to be correct. Across 24 baselines, GPT-5.6 attains 57.09% overall accuracy but only 3.93% QExact on 483 quartets. We further propose PATHOARGUS, a framework that selects question-relevant evidence patches from whole-slide images to answer pathology questions. It reaches 65.15% overall accuracy, 68.27% ESG accuracy, and 43.27% QExact using only 512 selected patches per question, retaining 1.51% of available bench features. These results show more consistent answers across evidence states, while fine-grained local findings remain challenging.

1 INTRODUCTION

Whole-slide pathology question answering (QA) requires locating visual evidence relevant to a question within gigapixel whole-slide images (WSIs). This evidence may occupy a small region or span multiple slides. For example, identifying which candidate WSI set contains a specified lesion requires finding the corresponding tissue; if the target is absent from all sets, the correct answer is None.

Existing work studies slide-level QA, evidence search, and visual grounding using different tasks and input interfaces (Chen et al., 2024c; 2025; Liang et al., 2025; Lu et al., 2026; Buckley et al., 2025; Liao et al., 2026; Zhang et al., 2026; Chen et al., 2026b). Answer accuracy alone, however, does not establish whether a model is able to effectively use visual evidence. MIRAGE shows that models can answer many medical and general VQA questions without images (Asadi et al., 2026). We also observe this critical limitation in WSI VQA. As shown in Fig. 1a, without any visual input, SlideChat is able to reach 71.06% on SlideBench (Chen et al., 2025), 80.20% on the four-choice subset of WSI-Bench (Liang et al., 2025), and 60.26% on WSI-VQA (Chen et al., 2024c). Furthermore, the baseline fails to follow the target evidence when it moves or disappears in Fig. 1b.

We therefore ask: **Can current multimodal large language models (MLLMs) ground pathology predictions in visual evidence from WSIs and respond correctly when that evidence changes?** To evaluate this capability, we propose six benchmark tasks related to crucial pathological questions.

In this paper, we present the PATHOARGUS dataset, covering six pathology capabilities across 15 TCGA projects (Fig. 2a). It contains 22,078 four-choice questions, including 4,281 test questions in PATHOARGUS-BENCH. Labels are derived from case-linked reports. Alongside per-task accuracy and text-only controls, we further propose Evidence-Set Grounding (ESG), which constructs counterfactual quartets. The target case appears in Set 1, 2, or 3, or is absent, while the question and answer choices remain fixed. The bench split includes 483 quartets. ESG accuracy measures whether a model identifies which candidate set contains the target evidence, or recognizes its ab-

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

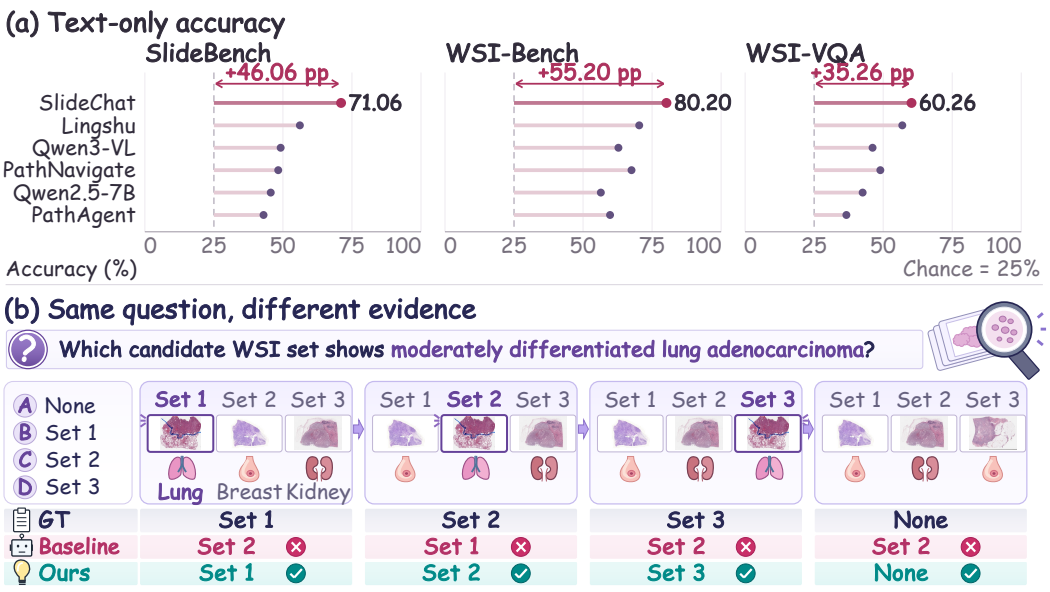


Figure 1: (a) Text-only accuracy on three benchmarks; arrows mark gaps above 25% chance (WSI-Bench: four-choice subset). (b) A separate Qwen2.5 qualitative quartet compares the F2000 baseline with Ours as target evidence moves or disappears.

sence. QExact tests whether all four answers correctly follow the target evidence as it moves or disappears.

Evaluation of 24 baseline systems reveals a gap between individual-answer accuracy and quartet consistency. GPT-5.6 reaches 57.04% ESG accuracy but only 3.93% QExact. These results reveal failures to reliably identify the candidate set containing the target finding as the evidence changes.

We further propose PATHOARGUS, a novel framework for evidence-grounded whole-slide pathology QA. It uses the question to select relevant evidence patches from the full collection of supplied WSIs and answers based on the selected evidence, retaining each patch's set and slide identity.

PATHOARGUS reaches 65.15% Overall accuracy, 68.27% ESG accuracy, and 43.27% QExact using only 512 selected patches per question, leading the evaluated systems on all three metrics. Fine-grained pathology assessment remains challenging even with more consistent answers across evidence states.

Our contributions are threefold:

- We introduce the PATHOARGUS dataset for training and evaluating six pathology capabilities, with ESG and QExact testing whether answers follow changes in target evidence.
- We evaluate 24 baseline systems on the 4,281-question PATHOARGUS-BENCH, identifying differences across pathology tasks and a gap between individual-answer accuracy and quartet consistency.
- We propose PATHOARGUS, a framework that selects question-relevant evidence patches from whole-slide images to answer pathology questions.

2 RELATED WORK

Whole-slide QA and evidence acquisition. WSI-VQA, SlideBench, and WSI-Bench evaluate pathology question answering from whole-slide inputs (Chen et al., 2024c; 2025; Liang et al., 2025). GIANT, PathNavigate, and PathAgent search for evidence by navigating WSIs (Buckley et al., 2025; Yang et al., 2026; Chen et al., 2026a), while EviPathBench and PathView-Bench evaluate evidence search and reasoning (Liao et al., 2026; Chen et al., 2026b). HistoSelect (Huang et al., 2026a) selects question-relevant patches through tissue-group and patch-level selection for slide-level QA. PATHOARGUS allocates patches across candidate sets, slides, and spatial regions before local selec-

Table 1: Medical and pathology QA benchmark coverage; "P" denotes partial coverage. Column definitions appear in Appendix H.

Benchmark	Whole-slide context	Evidence acquisition	Visual dependence	Multi-slide context	Counterfactual intervention
PathVQA (He et al., 2021)	✗	✗	✗	✗	✗
OmniMedVQA (Hu et al., 2024)	✗	✗	✗	✗	✗
GMAI-MMBench (Chen et al., 2024b)	✗	✗	✗	✗	✗
PathMMU (Sun et al., 2024)	✗	✗	✓	✗	✗
WSI-VQA (Chen et al., 2024c)	✓	✗	✗	✗	✗
MicroVQA (Burgess et al., 2025)	✗	✗	✓	✗	✗
SlideBench (Chen et al., 2025)	✓	✗	P	✗	✗
WSI-Bench (Liang et al., 2025)	✓	✗	✗	✗	✗
PathoArgus-Bench	✓	✓	✓	✓	✓

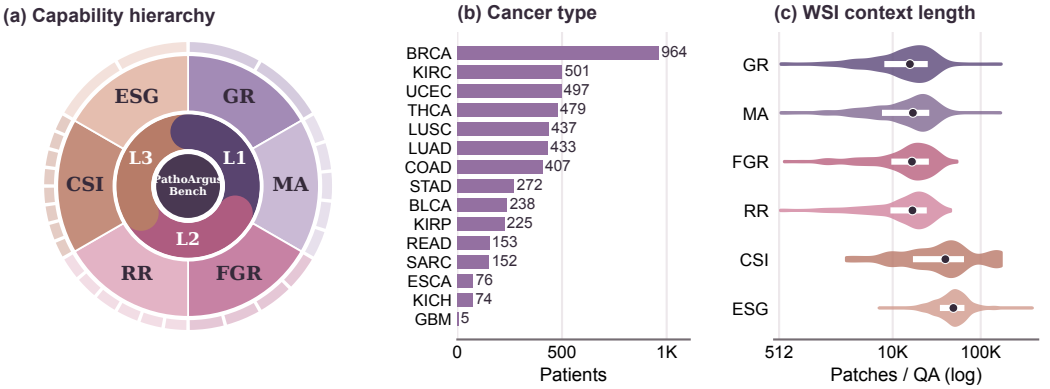


Figure 2: PATHOARGUS-BENCH overview: (a) capability taxonomy (equal sectors denote categories), (b) cancer type by TCGA project, and (c) WSI context length (available patches per question; questions with at least 512 patches). Dots and white bars indicate medians and interquartile ranges on a log scale.

tion, preserving each patch's source identity. Coverage constraints limit competition across sources, reducing the risk that a few high-scoring regions crowd out evidence needed for comparison.

Visual dependence. MIRAGE and PathBind test whether answers depend on visual content (Asadi et al., 2026; Zhang et al., 2026). PATHOARGUS-BENCH tests this dependence in WSI QA using text-only controls and ESG. The former measure accuracy without images; the latter changes the target case's location or presence while keeping the question fixed. QExact requires correct predictions across all four evidence states. Table 1 compares benchmark coverage.

3 DATASET CONSTRUCTION

3.1 EVALUATION TARGET

Each input contains a question q, four answer choices $\mathcal{A} = \{a_1, \ldots, a_4\}$, and patch features $\mathcal{X} = \{x_1, \ldots, x_N\}$ from one or more case-linked WSIs. A system reads at most K features and predicts one choice. PATHOARGUS-BENCH records both the available patch count N and input limit K for answer generation. ESG provides three candidate WSI sets and a None option for questions whose target is absent from all sets.

Patch counts describe available and selected input. Text-only and ESG controls test visual dependence and responses to evidence changes, respectively.

3.2 PATHOLOGY TASKS AND EVIDENCE REQUIREMENTS

We abbreviate the six capabilities as GR (Global Recognition), MA (Morphologic Analysis), FGR (Fine-Grained Recognition), RR (Regional Reasoning), CSI (Cross-Slide Integration), and ESG (Evidence-Set Grounding). The tasks cover different types of visual evidence (Fig. 2(a)). GR and MA assess broad tissue identity and morphology; FGR and RR assess local findings. CSI uses

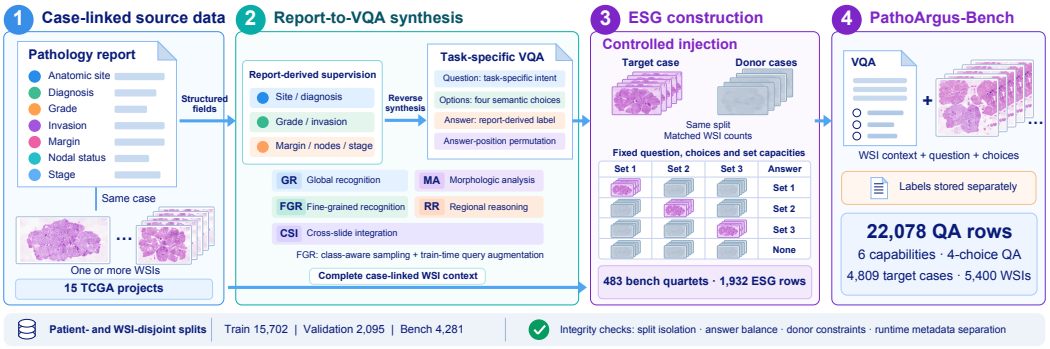


Figure 3: Dataset construction from case-linked reports and WSIs. ESG moves or removes target evidence while preserving questions, choices, and WSI counts per set.

multiple slides to answer case-level questions with report-derived labels. ESG tests whether the model identifies the candidate set containing the target or correctly answers None. Its 483 underlying questions each have four evidence states, forming 1,932 evaluation rows. Appendix A.3 provides task counts and subtask abbreviations.

3.3 Construction Pipeline

Starting from 4,962 patients and 5,516 locally registered TCGA WSIs (Cancer Genome Atlas Research Network et al., 2013), we follow WSI-Bench (Liang et al., 2025) in constructing QA pairs from case-linked reports (Fig. 3). We link each report to the available WSIs from the same case and use its structured fields for anatomic site, diagnosis, grade, invasion, margin, nodal status, and stage as answer supervision.

For GR, MA, FGR, RR, and CSI, each report field is mapped to a task-specific question and four clinical answer choices. For example, a grade field supplies the reference answer to a grading question, while a margin field supports a margin-status question. We permute answer positions and pair each question with the complete available case-linked WSI context; CSI questions include multiple slides from the same case. For FGR, class-aware sampling and question augmentation during training preserve rare combinations of perineural and lymphovascular invasion labels.

ESG then constructs four evidence configurations for the same question and choices. The target case occupies Set 1, Set 2, or Set 3, with donor cases filling the other sets. In the fourth configuration, donor slides replace the target, producing the None reference answer. Donors come from the same split with matched WSI counts, so every configuration preserves the number of slides in each set. Approximately half of the bench ESG quartets use same-organ donors, and the remainder use cross-organ donors.

We stratify the pool of 4,913 patients by TCGA project, FGR label, and WSI count. Automated checks verify patient/WSI disjointness, balanced answer positions, and ESG donor constraints. Each final QA item contains the WSI context, question, and four choices, with the reference label stored separately. Model inputs exclude answers, case identifiers, counterfactual conditions, decision metadata, and construction fields.

Size and patch counts. The PathoArgus dataset contains 22,078 questions over 4,809 target cases and 5,400 WSIs: PathoArgus-Instruct has 15,702 training questions, PathoArgus-Val has 2,095 validation questions, and PathoArgus-Bench has 4,281 test questions. Splits share no patients or WSIs. Bench questions provide 33,743 patch features on average and up to 388,637; selecting up to 512 retains 1.51% of available features (Fig. 2c). Appendix A reports detailed split and workload statistics.

Non-visual controls. Answer-position counts differ by at most one question in each split and by at most two questions within each bench task. A zero-shot Qwen2.5-7B text-only control obtains 24.39% Overall accuracy and 0/483 QExact; 96.07% of its ESG quartets receive a constant prediction across all evidence states. FGR text-only accuracy remains above the nominal 25% chance level

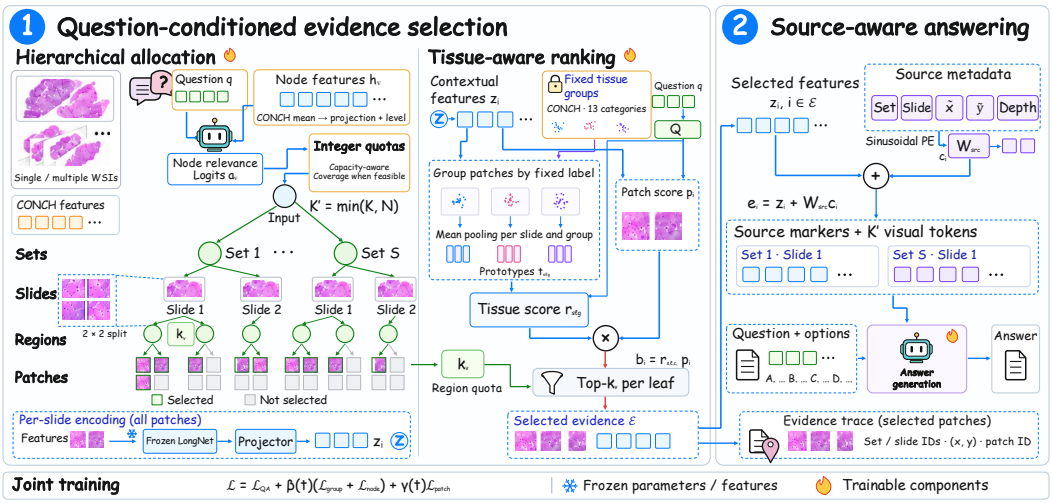


Figure 4: PATHOARGUS (1) selects question-relevant patches through hierarchical allocation and tissue-aware ranking, then (2) answers using the selected evidence and its source information.

at 30.87%, despite balanced answer positions. Appendix B audits the clinical choices represented by these positions.

4 PATHOARGUS

4.1 OVERVIEW

PATHOARGUS selects question-relevant evidence patches from all supplied WSIs and preserves their set and slide identities for the reader, the model that answers the question (Fig. 4).

Let $\mathcal{C} = \{\mathcal{S}_s\}_{s=1}^{S}$ denote the input collection, where $\mathcal{S}_s = \{W_{s\ell}\}_{\ell=1}^{L_s}$ contains the slides of set s, with $S = 3$ for ESG and $S = 1$ otherwise. Each patch i has a visual feature $\mathbf{x}_i \in \mathbb{R}^{d_v}$, set and slide identifiers (s_i, ℓ_i), and slide-local coordinates $\mathbf{p}_i$. Given question q and all N available patches, we select up to K patches to form the evidence set $\mathcal{E}$:

$$|\mathcal{E}| = K' = \min(K, N), \qquad \hat{a} = \text{Reader}_\theta\left(q, \text{Serialize}(\mathcal{E})\right). \tag{1}$$

All tasks share the selector and reader, without task-specific answering heads.

4.2 QUESTION-CONDITIONED EVIDENCE SELECTION

To preserve evidence from multiple sources, we allocate patch quotas across sets, slides, and regions according to question relevance. These quotas determine how many patches each region contributes; local ranking determines which patches it contributes.

Spatial structure and slide context. Per-slide coordinate quadtrees form the hierarchy input → set → slide → spatial region → patch. Splitting stops at B patches, depth D, or indivisible coordinates; spatial levels group patches at a single representation scale.

To represent patches in their slide context, we encode each slide independently: $H_{s\ell} = F(X_{s\ell})$ and $Z_{s\ell} = P(H_{s\ell})$, where F is a frozen LongNet encoder and P is a trainable visual projector. Here $X_{s\ell}$ contains the slide's patch features $\mathbf{x}_i$, and $Z_{s\ell}$ contains their projected contextual features $\mathbf{z}_i$. The encoder processes all candidate patches; K limits only the features passed to the reader.

Node relevance. For a nonvirtual node v, let I_v be its descendant patch indices and $n_v = |I_v|$ its capacity. We summarize its visual content and score it against the precomputed question embedding

$\mathbf{q} \in \mathbb{R}^{d_v}$:

$$\boldsymbol{\mu}_v = \frac{1}{n_v} \sum_{i \in I_v} \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}, \qquad \mathbf{h}_v = \phi_v(\boldsymbol{\mu}_v) + \mathbf{e}_{\lambda(v)},$$
$$\mathbf{u} = \phi_q(\mathbf{q}), \qquad a_v = \phi_s([\mathbf{h}_v; \mathbf{u}; \mathbf{h}_v \odot \mathbf{u}]), \quad g_v = \sigma(a_v). \tag{2}$$

Here $\mathbf{e}_{\lambda(v)}$ encodes hierarchy level: sets have level zero and slide roots level one. The feature mappings and scoring MLP are shared across nodes.

Integer quotas. Starting with K' at the input root, we use node scores to recursively allocate each parent's quota k_u among its children:

$$0 \le k_v \le n_v, \qquad \sum_{v \in \mathrm{Ch}(u)} k_v = \min\left(k_u, \sum_{v \in \mathrm{Ch}(u)} n_v\right). \tag{3}$$

When the parent quota permits, each nonempty child receives at least one patch. Remaining slots follow a softmax over child logits a_v, subject to each child's patch count. A child with zero quota contributes no patches. Appendix D.1 details rounding, redistribution, and tie handling.

Tissue-aware ranking within regions. Within each regional quota, selection considers both the tissue relevant to the question and the content of individual patches. We assign each patch a fixed label c_i from 13 tissue categories using zero-shot CONCH image–text matching with category prompts (Appendix D.2). For each nonempty tissue category g on slide (s, ℓ), let $\mathbf{t}_{s\ell g}$ be the mean of its projected contextual features $\mathbf{z}_i$. With a trainable question projection $\bar{\mathbf{q}} = Q(\mathbf{q})$, we multiply tissue-group and patch relevance scores (Huang et al., 2026a):

$$r_{s\ell g} = \sigma(f_g([\mathbf{t}_{s\ell g}; \bar{\mathbf{q}}])),$$
$$p_i = \sigma(f_p([\mathbf{z}_i; Q_p(\bar{\mathbf{q}})])), \qquad b_i = r_{s_i \ell_i c_i}\, p_i. \tag{4}$$

Selectors share parameters across slides; tissue prototypes are slide-specific. Each leaf contributes its k_v highest-scoring patches to $\mathcal{E}$. Tissue categories affect ranking, without separate quotas.

4.3 Source-Aware Answering

Visual findings must remain associated with their source slides and, for ESG, their candidate sets. We therefore augment each selected patch with source information, concatenating sinusoidal encodings of its set, slide, normalized local coordinates, and spatial leaf depth:

$$\mathbf{c}_i = [\mathrm{PE}(s_i); \mathrm{PE}(\ell_i); \mathrm{PE}(\tilde{x}_i); \mathrm{PE}(\tilde{y}_i); \mathrm{PE}(d_i)],$$
$$\mathbf{e}_i = \mathbf{z}_i + W_{\mathrm{src}} \mathbf{c}_i. \tag{5}$$

We serialize embeddings by set and slide with `Candidate WSI Set s, Slide l:` markers, then score option letters conditioned on this sequence and the question. Markers occupy text positions beyond the K' visual embeddings. Selected patch indices are retained for inspection.

4.4 Joint Training

QA examples from all six tasks supervise the shared selector and answering model through their reference answers.

Counterfactual evidence-set supervision. ESG quartets keep the question and choices fixed while placing the target in Set 1, 2, or 3, or removing it. The reference answer follows the target's location or absence. Each variant receives independent cross-entropy supervision on the correct set or None answer.

Joint optimization. A straight-through surrogate sends answer-loss gradients to allocation and local-selection scores along the tree paths of selected patches. Integer quotas and top-k membership remain discrete. Auxiliary similarity-prior losses supervise node, tissue-group, and patch scores across the candidate pool using visual–question similarities as targets:

$$\mathcal{L} = \mathcal{L}_{\mathrm{QA}} + \beta(t)(\mathcal{L}_{\mathrm{group}} + \mathcal{L}_{\mathrm{node}}) + \gamma(t)\mathcal{L}_{\mathrm{patch}}, \tag{6}$$

Table 2: Bench accuracy (%); the human baseline averages two experts' scores on a separate 250-question subset (Appendix C.3). Bold/underline mark best/second-best open-source results; † marks original agent backbones.

Model	Size	Overall↑	GR↑	MA↑	FGR↑	RR↑	CSI↑	ESG↑	QExact↑
Human experts (subset)	N/A	99.20	100.00	100.00	100.00	96.67	96.67	100.00	100.00
text-only	7B	24.39	25.87	11.19	30.87	26.95	26.53	24.95	0.00
▼ *General-purpose MLLMs*									
GPT-5.6 (OpenAI, 2026)	N/R	57.09	85.20	46.85	24.16	42.93	46.26	57.04	3.93
InternVL3 (Zhu et al., 2025)	8B	27.12	37.44	16.32	<u>37.58</u>	22.80	26.53	26.29	0.00
Qwen3-VL (Bai et al., 2025a)	8B	26.47	36.19	23.54	18.12	22.44	31.29	25.05	0.00
Qwen3-VL-A3B (Bai et al., 2025a)	30B-A3B	26.42	33.96	25.41	24.83	24.63	25.85	24.43	0.00
GLM-4.6V-Flash (GLM-V Team et al., 2025)	10B	24.69	25.00	23.08	24.83	25.00	23.81	24.84	0.00
Vision-DeepResearch-8B (Huang et al., 2026b)	8B	24.67	26.87	17.72	23.49	24.39	29.93	25.10	0.00
▼ *Medical MLLMs*									
LLaVA-Med v1.5 (Li et al., 2023)	7B	24.83	25.00	24.94	24.83	24.27	24.49	25.00	0.00
MedVLM-R1-2B (Pan et al., 2025)	2B	29.60	35.07	16.55	28.86	42.68	27.89	24.84	0.00
HuatuoGPT-V (Chen et al., 2024a)	7B	25.67	38.68	16.55	20.13	19.51	27.21	25.21	0.00
HuatuoGPT-V (Chen et al., 2024a)	34B	27.54	39.30	<u>42.42</u>	26.85	15.24	28.57	24.53	1.24
Lingshu (LASA Team et al., 2025)	7B	35.76	64.93	41.72	33.56	27.93	29.93	26.24	0.21
Lingshu (LASA Team et al., 2025)	32B	36.16	66.04	41.49	36.24	25.24	32.65	27.43	0.83
MedGemma 1.5 (Sellergren et al., 2026)	4B	26.14	36.07	15.38	26.85	23.66	29.25	25.16	0.00
MedGemma (Sellergren et al., 2025)	27B	29.01	33.21	28.21	<u>37.58</u>	33.05	31.97	24.84	0.00
▼ *Pathology-specific methods*									
Quilt-LLaVA (Seyfioglu et al., 2024)	7B	25.20	25.12	24.01	24.83	26.71	23.81	25.00	0.00
Patho-R1-7B (Zhang et al., 2025)	7B	28.85	56.59	13.29	29.53	19.39	25.17	25.00	0.00
PathGen-LLaVA (Sun et al., 2025)	13B	33.52	69.28	28.44	16.11	23.41	29.93	25.67	0.00
W2T (Chen et al., 2024c)	∼24M	24.43	26.00	40.79	22.82	12.68	27.89	25.00	0.00
SlideChat (Chen et al., 2025)	∼7B	25.91	25.25	23.08	28.86	26.83	27.21	26.09	0.00
WSI-LLaVA (Liang et al., 2025)	∼7B	30.46	51.74	25.17	24.16	26.22	29.93	25.10	0.00
HistoSelect (Huang et al., 2026a)	7B	28.66	50.00	21.45	30.20	22.20	23.81	24.38	0.00
MLLM-HWSI (Alawode et al., 2026)	7B	27.31	45.27	21.21	36.24	17.68	21.09	25.05	0.00
PathNavigate (Yang et al., 2026)	7B+4B†	25.91	32.96	21.21	20.13	23.05	31.29	25.26	0.21
PathAgent (Chen et al., 2026a)	7B+4B†	25.46	32.84	20.28	19.46	23.05	34.69	24.33	0.00
▼ *PathoArgus*									
SFT (Qwen2.5)	7B	32.38	54.48	31.93	24.83	29.51	33.33	25.00	0.00
PathoArgus (Qwen2.5) (Bai et al., 2025b)	7B	<u>62.04</u>	**87.56**	**42.66**	**38.26**	**64.39**	**52.38**	57.30	26.09
PathoArgus (Qwen3-VL) (Bai et al., 2025a)	8B	**65.15**	86.19	33.57	32.89	<u>64.27</u>	38.78	**68.27**	**43.27**
PathoArgus (Lingshu) (LASA Team et al., 2025)	7B	61.39	<u>86.82</u>	37.76	32.21	62.44	<u>44.22</u>	<u>59.16</u>	<u>30.02</u>

where $\mathcal{L}_{\mathrm{QA}}$ is mean answer-token negative log-likelihood. We train the scorers, projections, selectors, and reader LoRA while keeping patch features, question embeddings, and LongNet fixed. Training and inference share a single allocation and selection pass before answering. Optimization details and model configurations appear in Appendix D.

5 Experiments

5.1 Evaluation Protocol

We evaluate 24 pretrained systems on the 4,281-question bench split using WSI patches as visual inputs and closed-set A–D answers. PathoArgus uses Qwen2.5, Qwen3-VL, and Lingshu readers. The limit of 512 selected patches per question was chosen on the validation split. Diagnostic analyses use Qwen3-VL. In the visual-input ablations, each visual/text-only pair uses the same trained checkpoint, differing only in visual input. Input interfaces and implementation details are provided in Appendix C.1. We report Overall accuracy (total correct answers divided by total questions), per-capability accuracy, and quartet exact-match accuracy (QExact), which tests correct responses across all four counterfactual ESG states:

$$\mathrm{QExact} = \frac{1}{G}\sum_{g=1}^{G}\mathbf{1}\left[\bigwedge_{c=1}^{4}\hat{y}_{g,c} = y_{g,c}\right], \tag{7}$$

where $G = 483$ is the number of ESG quartets, and $\hat{y}_{g,c}$ and $y_{g,c}$ denote the predicted and reference answers for condition c.

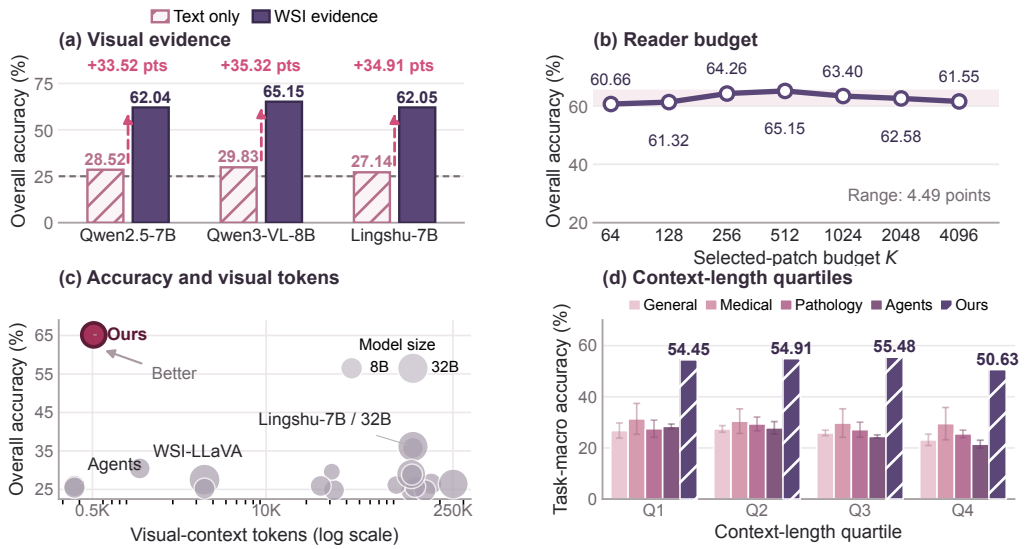


Figure 5: PathoArgus accuracy by (a) visual input, (b) selected-patch count, (c) visual-token input cost, and (d) available context. Panels (b–d) use Qwen3-VL.

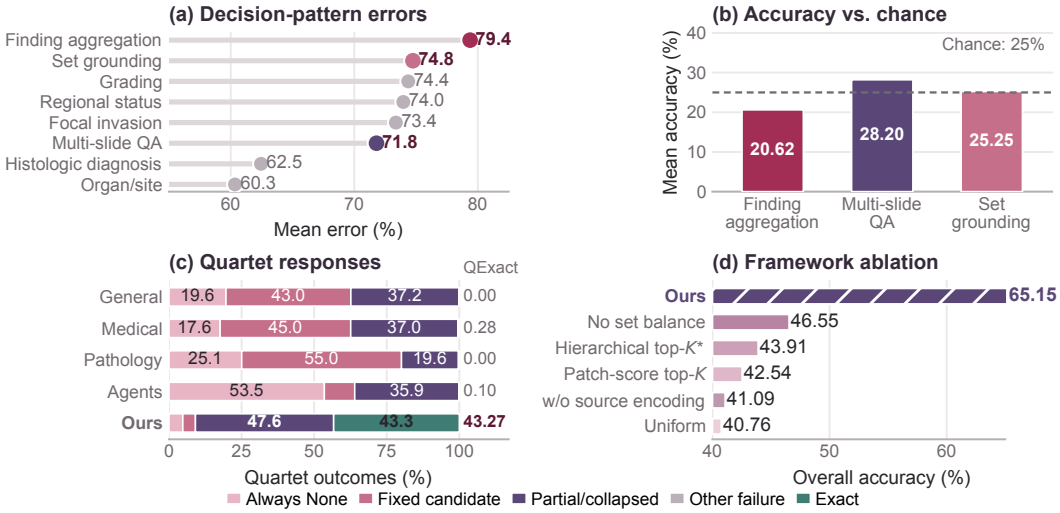


Figure 6: Across 22 external baselines, (a) decision errors and (b) accuracy on three evidence requirements; the dashed line marks 25% chance. (c) Quartet responses alongside Ours and (d) Qwen3-VL framework ablations at $K = 512$; * removes balancing and coverage together.

5.2 Performance across Pathology Tasks

Finding 1. PathoArgus improves pathology QA and evidence-set grounding. PathoArgus reaches 65.15% Overall accuracy using only 512 selected patches per question, exceeding GPT-5.6 by 8.06 points on the same bench split. It attains the highest ESG accuracy and QExact in Tab. 2, at 68.27% and 43.27%, respectively. With the same training split and experimental configuration, Qwen2.5 PathoArgus improves all six tasks over the SFT baseline, gaining 29.66 points Overall. ESG accuracy rises from 25.00% to 57.30%, and QExact from 0.00% to 26.09%. Matched visual-input ablations show gains of 33.52–35.32 points Overall across all three PathoArgus configurations (Fig. 5a). For reader-side cost-effectiveness, Fig. 5c compares Overall accuracy against reader-side visual-token count as a proxy for input cost; bubble area denotes parameter count. PathoArgus reaches 65.15% using only 512 selected patch embeddings, versus 36.16% for Lingshu-32B at an estimated 125.6K visual tokens. Counts include both measurements and estimates from each model's input interface; they exclude feature extraction and candidate encoding costs.

5.3 PATHOLOGY DECISIONS AND EVIDENCE RESPONSIVENESS

We analyze eight pathology decision types across 22 external baselines with complete predictions (Appendix C.5). **Finding 2. Models struggle with decisions across foci, slides, and candidate sets.** Extending morphology and diagnostic QA in WSI-Bench (Liang et al., 2025) and SlideBench (Chen et al., 2025), PATHOARGUS-BENCH tests decisions across foci, slides, and candidate sets. Across 22 external baselines, mean errors are 79.38% for finding aggregation (identifying the most adverse finding across multiple foci), 71.80% for multi-slide case QA, and 74.75% for candidate-set grounding (Fig. 6a). Accuracy is 20.62–28.20% against 25% chance (Fig. 6b). These errors motivate question-guided patch selection that preserves slide and set identities. **Finding 3. PathoArgus more reliably follows target evidence across counterfactual states.** PATHOARGUS reaches 68.27% ESG accuracy and 43.27% QExact, versus 57.04% and 3.93% for GPT-5.6 (209 versus 19 correct quartets). Each correct quartet identifies the set containing the target evidence in all three placements and selects None when it is absent. Fig. 6c distinguishes four failure patterns: *Always None* selects None in all four states; *Fixed candidate* selects the same set throughout; *Partial/collapsed* repeats an option for states requiring different answers, using only two or three distinct options; and *Other failure* includes selecting all four options in an incorrect order or producing invalid outputs. The 22 external baselines average 0.11% QExact, often repeating one candidate or always None; PATHOARGUS makes these constant responses in 9.11% of quartets. Appendices C.6 and E detail response patterns and examples.

5.4 ABLATION STUDY

Finding 4. Structured allocation and source encoding improve QA accuracy. With Qwen3-VL and 512 selected patches, full PATHOARGUS reaches 65.15% Overall accuracy (Fig. 6d). **Uniform sampling** selects without question-relevance scores and reaches 40.76%, favoring question-guided evidence selection. **Patch-score Top-K** ranks patches globally instead of assigning hierarchical quotas, reaching 42.54%. **Hierarchical Top-K** retains hierarchy-informed scores but selects globally without balancing or coverage, reaching 43.91%. Their lower accuracy shows that relevance scores alone do not replace structured allocation. **No set balance** removes candidate-set balancing from evidence allocation and reaches 46.55%, supporting allocation across sets. **Without source encoding** removes patch-level source vectors but retains set/slide text markers, reaching 41.09%. This 24.06-point drop shows that source encoding adds value beyond text markers.

5.5 HYPERPARAMETER SENSITIVITY AND CONTEXT-LENGTH ROBUSTNESS

We first evaluate sensitivity to K, the number of patches selected for answering (Fig. 5b). Across $K = 64$–$4{,}096$, Overall accuracy ranges from 60.66% to 65.15%. Selecting only 256 patches achieves 64.26%, just 0.89 points below the best result at 512 patches, showing that strong performance does not require a narrowly tuned K. Increasing K beyond 512 provides no further gain, with accuracy falling to 61.55% at 4,096 patches. We chose the default $K = 512$ on the validation split. We then assess robustness to context length, measured by the total number of available patches. We sort questions into four equally sized groups, from Q1 (shortest contexts) to Q4 (longest contexts), while keeping $K = 512$ (Fig. 5d). PATHOARGUS leads in all four groups. In Q4, accuracy averaged equally across tasks reaches 50.63%, exceeding the strongest baseline family mean of 29.51% by 21.12 points. Thus, the advantage persists when the model must select evidence from the largest WSI collections. The groups contain different questions; they do not test context expansion for the same question. Appendix C.7 provides aggregation details and full results.

6 CONCLUSION

PATHOARGUS-BENCH evaluates six pathology tasks using complete WSI context, text-only controls, and tests that move or remove the target evidence. The PATHOARGUS framework reaches 65.15% Overall accuracy and 43.27% QExact using only 512 selected patches per question, leading the evaluated systems on both metrics. Broader verification of report-derived labels against the supplied WSIs and evaluation beyond TCGA remain needed.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and polishing, figure and table preparation, and consistency-checking code. The authors reviewed the AI-assisted materials and take responsibility for the final content.

ETHICS STATEMENT

This work evaluates computational pathology systems using TCGA whole-slide images and pathology reports. Model inputs use question identifiers that do not expose case identity and exclude case identifiers, answers, and construction metadata. The benchmark is intended for research rather than clinical diagnosis, and its distribution follows the applicable TCGA data-use and provenance requirements. Differences between TCGA projects may reflect patient composition, specimen processing, and archival practices, which may also affect aggregate scores.

REPRODUCIBILITY STATEMENT

Sections 3 and 5 describe data construction, split assignment, tasks, metrics, model inputs, and patch counts. Section 4 explains how PathoArgus scores and selects patches across candidate sets, slides, and regions. Appendix Table 14 lists the configuration. Supplementary files provide task and project statistics, patch counts per question, coverage of clinical answer choices, and data and scripts for the quantitative figures.

REFERENCES

Basit Alawode, Arif Mahmood, Muaz Khalifa Al-Radi, Shahad Albastaki, Asim Khan, Muhammad Bilal, Moshira Ali Abdalla, Mohammed Bennamoun, and Sajid Javed. MLLM-HWSI: A multimodal large language model for hierarchical whole slide image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Alawode_MLLM-HWSI_A_Multimodal_Large_Language_Model_for_Hierarchical_Whole_Slide_CVPR_2026_paper.html.

Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. MIRAGE: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026. URL https://arxiv.org/abs/2603.21687.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Thomas A. Buckley, Kian R. Weihrauch, Katherine Latham, Andrew Z. Zhou, Padmini A. Manrai, and Arjun K. Manrai. Navigating gigapixel pathology images with large multimodal models. *arXiv preprint arXiv:2511.19652*, 2025. URL https://arxiv.org/abs/2511.19652.

James Burgess, Jeffrey J. Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G. Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D. Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. MicroVQA: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19552–19564, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Burgess_MicroVQA_A_Multimodal_Reasoning_Benchmark_for_Microscopy-Based_Scientific_Research_CVPR_2025_paper.html.

Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10): 1113–1120, 2013. doi: 10.1038/ng.2764.

Jingyun Chen, Linghan Cai, Zhikang Wang, Yi Huang, Songhan Jiang, Shenjin Huang, Hongpeng Wang, and Yongbing Zhang. Pathagent: Toward interpretable analysis of whole-slide pathology images via large language model-based agentic reasoning. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2511.17052.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision: Towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024a. URL https://arxiv.org/abs/2406.19280.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Yu Qiao, and Junjun He. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-2992. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ab7e02fd60e47e2a379d567f6b54f04e-Abstract-Datasets_and_Benchmarks_Track.html.

Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. WSI-VQA: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision*, 2024c. URL https://arxiv.org/abs/2407.05603.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. 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*, pp. 5134–5143, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_SlideChat_A_Large_Vision-Language_Assistant_for_Whole-Slide_Pathology_Image_Understanding_CVPR_2025_paper.html.

Zongyi Chen, Yu Liang, Jie Lin, and Liansheng Wang. Pathview-bench: Can multimodal large language models achieve fine-grained multiscale understanding of pathology images? *arXiv preprint arXiv:2607.28318*, 2026b. URL https://arxiv.org/abs/2607.28318.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. URL https://arxiv.org/abs/2507.01006.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718, 2021. doi: 10.18653/v1/2021.acl-short.90. URL https://aclanthology.org/2021.acl-short.90/.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22170–22183, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Hu_OmniMedVQA_A_New_Large-Scale_Comprehensive_Evaluation_Benchmark_for_Medical_LVLM_CVPR_2024_paper.html.

Wentao Huang, Weimin Lyu, Peiliang Lou, Qingqiao Hu, Xiaoling Hu, Shahira Abousamra, Wenchao Han, Ruifeng Guo, Jiawei Zhou, Chao Chen, and Chen Wang. Act like a pathologist: Tissue-aware whole slide image reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6972–6981, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_Act_Like_a_Pathologist_Tissue-Aware_Whole_Slide_Image_Reasoning_CVPR_2026_paper.html.

Wenxuan Huang, Yu Zeng, Qiuchen Wang, Zhen Fang, Shaosheng Cao, Zheng Chu, Qingyu Yin, Shuang Chen, Zhenfei Yin, Lin Chen, et al. Vision-deepresearch: Incentivizing deepresearch capability in multimodal large language models. *arXiv preprint arXiv:2601.22060*, 2026b. URL https://arxiv.org/abs/2601.22060.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. URL https://arxiv.org/abs/2306.00890.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. 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. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liang_WSI-LLaVA_A_Multimodal_Large_Language_Model_for_Whole_Slide_Image_ICCV_2025_paper.html.

Dankai Liao, Tianyi Zhang, Yufeng Wu, Xinyue Zhang, Qiaochu Xue, Zeyu Liu, Dachun Zhao, Linghan Cai, and Yueming Jin. Evipathbench: Benchmarking evidence acquisition and reasoning in vision-language models for whole-slide pathology. *arXiv preprint arXiv:2607.19261*, 2026. URL https://arxiv.org/abs/2607.19261.

Hao Lu, Ziniu Qian, Yifu Li, Yang Zhou, Bingzheng Wei, and Yan Xu. CTIS-QA: Clinical template-informed slide-level question answering for pathology. *arXiv preprint arXiv:2601.01769*, 2026. URL https://arxiv.org/abs/2601.01769.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention*, pp. 337–347, 2025. doi: 10.1007/978-3-032-04981-0_32. URL https://arxiv.org/abs/2502.19634.

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

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. URL https://arxiv.org/abs/2604.05081.

Mehmet Saygin Seyfioglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 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*, 2024. URL https://arxiv.org/abs/2312.04746.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. PathMMU: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pp. 56–73, 2024. doi: 10.1007/978-3-031-73033-7_4. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/7899_ECCV_2024_paper.php.

Yuxuan Sun, Zhongyi Shui, Yunlong Zhang, Jingxiong Li, Yixuan Si, Xuan Gong, Chenglu Zhu, Xinheng Lyu, Kai Zhang, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rFpZnn11gj.

Chunze Yang, Qidong Liu, Wenjie Zhao, Yue Tang, Jiusong Ge, Di Zhang, Jiashuai Liu, Lei Wu, Junbo Lu, Ni Zhang, Xian Wu, Zeyu Gao, and Chen Li. Pathnavigate: A training-free pathology agent with surprise-guided scan and shared slide memory for whole-slide image vqa. *arXiv preprint arXiv:2605.23559*, 2026. URL https://arxiv.org/abs/2605.23559.

Chengyang Zhang, Wenchuan Zhang, Bo Li, Xinyu Liu, Jiaming Yang, Mengran Li, Chenxun Deng, Jie Chen, Yang Zhang, Wei Ju, Yuhao Yi, Hong Bu, and Jiancheng Lv. Do pathology vision-language models truly see pathology? *arXiv preprint arXiv:2607.21065*, 2026. URL https://arxiv.org/abs/2607.21065.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-R1: A multimodal reinforcement learning-based pathology expert reasoner. *arXiv preprint arXiv:2505.11404*, 2025. URL https://arxiv.org/abs/2505.11404.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.