BLIND AT GIGAPIXEL SCALE: VISUAL INFORMATION UTILIZATION IN MLLMS FOR WHOLE-SLIDE IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology multimodal large language models (PMLLMs) have attracted increasing attention for their ability to process whole-slide images (WSIs) and support tasks such as diagnosis assistance, question answering, and report generation. Despite strong performance on existing benchmarks, whether these models effectively utilize slide-specific visual evidence remains unclear. Unlike normal images, WSIs require highly compressed visual representations due to their gigapixel scale, raising questions about whether diagnostically relevant information can be accessed and incorporated into model decisions. We systematically analyze visual evidence utilization in PMLLMs by examining their inputs, representations, and predictions. We evaluate representative models across different input paradigms under controlled visual interventions and introduce VAPath, a WSI–QA–ROI benchmark with pathologist-annotated visual anchors for assessing whether model-derived visual signals align with pathology-relevant regions. Our analysis reveals a gap between visual information availability and decision-level utilization: visual interventions often have limited effects on model predictions, while an independent pathology retrieval procedure identifies stronger ROI-associated patterns in WSI patch features than model-derived signals. We further show that multimodal instruction tuning can improve answer fitting without increasing the measured marginal contribution of WSI visual information under our evaluation protocol. These findings demonstrate that benchmark accuracy alone provides an incomplete measure of WSI understanding and motivate evaluations that directly characterize how visual evidence contributes to PMLLM decisions.

Code & Dataset: Blind at Gigapixel

1 INTRODUCTION

Generalist multimodal large models (MLLMs) have demonstrated strong capabilities across a broad range of vision-language tasks, including visual question answering and image understanding. (Achiam et al., 2023; Bai et al., 2025). These advances have motivated the adaptation of MLLMs to the medical domain, where recent studies have developed pathology MLLMs (PMLLMs) for whole-slide image (WSI) understanding. By integrating visual representations with medical knowledge, PMLLMs have achieved impressive performance across various pathology benchmarks for diagnosis assistance, question answering, and report generation (Lu et al., 2024b; Sun et al., 2024; Chen et al., 2025; Liang et al., 2025; Huang et al., 2026), with performance approaching expert-level results on some pathology benchmarks. However, despite their strong benchmark performance, how PMLLMs acquire and utilize WSI visual information remains insufficiently understood.

Unlike natural images, WSIs are gigapixel-scale images containing extremely long visual sequences, complex spatial structures, and sparse diagnostic patterns distributed across localized regions (Chen et al., 2022; Lu et al., 2021; Shao et al., 2021). To enable WSI understanding under the computational constraints of MLLMs, existing PMLLMs typically introduce a pathology-specific visual encoding stage, where pathology foundation models transform numerous patch-level regions into compact visual representations before integrating them with an LLM (Lu et al., 2024a; Chen et al., 2024b; Vorontsov et al., 2024; Xu et al., 2024). Consequently, each visual representation typically corresponds to a larger tissue region and contains highly compressed morphological information compared with the fine-grained visual tokens commonly used in general-purpose MLLMs. These

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.

representations are then aligned with pathology knowledge and task-specific responses through multimodal instruction tuning, allowing the LLM to generate predictions conditioned on compressed pathology visual representations. This paradigm raises two fundamental questions:

What pathology-relevant information can be identified from WSI features, and to what extent do evaluated PMLLMs rely on slide-specific visual inputs for decision making?

To investigate the visual evidence utilization capability of PMLLMs, we perform a systematic analysis through four fundamental questions:

Q1: *Do PMLLMs genuinely rely on WSI visual evidence for pathology reasoning?*—counterfactual visual interventions quantify whether removing or replacing WSI evidence changes model decisions and performance (§4);

Q2: *How does visual evidence utilization vary across PMLLMs?*—a cross-model analysis disentangles visual sensitivity from evidence utilization across different input paradigms (§5);

Q3: *How is visual information utilized in PMLLM decisions?*—attention, causal intervention, and retrieval analyses compare model-derived signals with independent pathologist annotations (§6);

Q4: *What limits the effective utilization of visual evidence in PMLLMs?*—controlled analyses examine how answer fitting relates to the measured contribution of visual inputs to final predictions (§7).

To answer these questions, we establish a systematic analysis framework for analyzing visual information utilization in pathology MLLMs. We curate 3,605 WSI-based question answering samples from diverse data sources and construct VAPath, a WSI–QA–ROI benchmark with pathologist-annotated visual anchors that enable evaluation of whether model-derived visual signals align with ROI. We further evaluate several representative PMLLMs spanning WSI-input and general-image-input paradigms through controlled visual interventions, cross-model behavioral analysis, spatial evidence auditing, and representation-to-decision analysis. Together, these evaluations provide a comprehensive characterization of how visual information is represented, utilized, and transformed throughout the WSI-to-LLM pipeline.

Our analysis reveals a progressive decoupling between visual information and model decisions in current PMLLM pipelines. Although current PMLLMs can achieve strong benchmark performance, visual information is not necessarily required, selected, or sufficiently weighted during decision making. We identify three critical gaps: correctness does not imply visual reliance, visual sensitivity does not imply effective utilization, and visual representation does not imply decision influence. Further analysis shows that multimodal instruction tuning can improve answer fitting without increasing the marginal contribution of slide-specific visual information.

These findings indicate that evaluating pathology MLLMs requires moving beyond result and towards directly characterizing how visual information contributes to model decisions. By identifying where visual information can be retrieved but insufficiently translated into decisions, our analysis provides insights for developing pathology MLLMs with stronger connections between WSI representations and decision making.

2 Related work

Pathology MLLMs. Recent pathology MLLMs extend vision–language modeling to gigapixel whole-slide images (WSIs). Representative approaches, including PathChat (Lu et al., 2024b), PathAsst (Sun et al., 2024), SlideChat (Chen et al., 2025), WSI-LLaVA (Liang et al., 2025), HistoSelect (Huang et al., 2026), and Patho-R1 (Zhang et al., 2026b), explore different strategies for integrating pathology visual representations with large language models. However, despite their promising performance, it remains unclear whether these models can reliably utilize slide-specific visual evidence.

WSI-Level Pathology Benchmarks. Recent benchmarks have enabled systematic evaluation of pathology MLLMs on whole-slide understanding. WSI-VQA (Chen et al., 2024a), SlideBench (Chen et al., 2025), and WSI-Bench (Liang et al., 2025) introduce large-scale WSI question answering and report-generation tasks. These benchmarks' primary metrics measure final answers or generated reports rather than the necessity, localization, or causal contribution of slide-

specific evidence. High benchmark performance does not necessarily indicate that models solve questions through the provided WSI.

Visual Evidence Reliance and Grounding Failures. Recent studies in multimodal learning have revealed that strong visual-language performance may not always reflect genuine visual understanding (Goyal et al., 2017; Agrawal et al., 2018; Cadene et al., 2019; Dancette et al., 2021). MIRAGE (Asadi et al., 2026), Liu et al. (Liu et al., 2026), and Zhang et al. (Zhang et al., 2026a) show that MLLMs can suffer from weak visual grounding, image neglect, or limited visual-semantic binding (Rahmanzadehgervi et al., 2024; Tong et al., 2024). We further investigate how visual evidence is represented, utilized, and transformed throughout the complete WSI-to-LLM pipeline.

3 Experimental Setup

We establish a systematic evaluation framework to investigate how visual information is utilized in pathology MLLMs. We summarize the shared experimental settings below and describe the specific protocols in the corresponding sections.

Models. We evaluate seven representative models spanning two input paradigms. **WSI-input pathology MLLMs** directly process WSI-derived patch representations, including SlideChat (Chen et al., 2025), WSI-LLaVA (Liang et al., 2025), and HistoSelect (Huang et al., 2026). **General-image-input MLLMs** process slide thumbnails or selected regions as conventional images, including the pathology-tuned Patho-R1 (Zhang et al., 2026b), the general-purpose Intern-VL3.5 (Wang et al., 2025) and Qwen3-VL (Bai et al., 2025), and the medically tuned MedGemma-1.5 (Sellergren et al., 2026). Additional implementation details are provided in Appendix C.1.

WSI QA Benchmarks We evaluate on four widely used WSI question-answering datasets. SlideBench (Chen et al., 2025) provide slide-level VQA over the TCGA and BCNB cohorts, respectively. WSI-Bench (Liang et al., 2025) contains both closed-set multiple-choice questions and open-ended reporting tasks, while WSI-VQA (Chen et al., 2024a) provides generative WSI-level VQA with a multiple-choice subset.

VAPath: Visual Anchor Pathology Benchmark To enable direct analysis of visual evidence utilization in PMLLMs, we construct **VAPath**, a WSI–QA–ROI benchmark containing 3,605 samples with 19,956 explicit visual anchors. VAPath integrates four public pathology sources, BCNB (Chen et al., 2025), BRACS (Brancati et al., 2022), CAMELYON+ (Ehteshami Bejnordi et al., 2017), and TCGA (Tomczak et al., 2014), together with a private soft-tissue sarcoma cohort, SARC, covering 3,556 unique whole-slide images from 2,940 patients. Because these datasets differ substantially in label systems, annotation formats, and task definitions, we develop a unified processing pipeline to harmonize diagnostic labels and convert heterogeneous spatial annotations into a consistent ROI representation. The pipeline further consolidates fragmented regions, filters annotations with insufficient or excessively broad tissue coverage, and manually reviews candidate samples under the supervision of pathology experts to remove ambiguous or unreliable visual anchors. For each curated WSI–ROI pair, we construct morphology-oriented questions based only on visually observable histopathological patterns, ensuring that answering the question requires reference to the annotated pathology regions rather than non-visual clinical information. The resulting WSI–QA–ROI triplets provide independent visual anchors for quantitatively comparing model attention, causal influence, and retrieval signals with pathology-relevant regions. Details of VAPath construction are provided in Appendix C.4.

4 Do MLLMs genuinely rely on WSI visual evidence for pathology reasoning?

We first examine when the visual input is altered while keeping the question unchanged, do predictions and accuracy change accordingly? Specifically, we perform counterfactual evaluation under three visual intervention conditions. **Text-only** removes visual inputs and tests whether models can answer from language information alone. **Blank** preserves the visual pathway while removing pathology content by replacing WSI features with blank-image features for WSI-input models or a blank image for general-image-input models. **Replaced** substitutes the original slide with a real

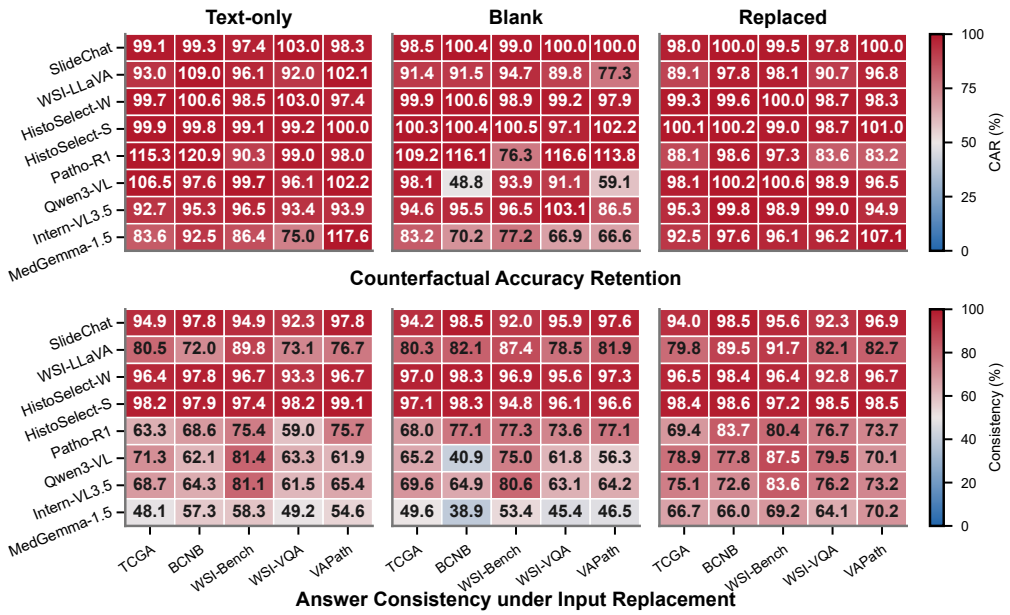

Figure 1: **Counterfactual evaluation across MLLMs and datasets.** Top: Counterfactual Accuracy Retention (CAR); bottom: answer consistency after visual interventions.

WSI from another patient, preserving a valid pathology input while removing slide-specific correspondence. Then we quantify each intervention using two paired metrics.

The *Counterfactual Accuracy Retention* (CAR) is defined as

$$\mathrm{CAR}(c) = \frac{\mathrm{Acc}(c)}{\mathrm{Acc}(\mathrm{Full})} \times 100\%, \tag{1}$$

where values close to 100% indicate limited performance gain from visual information. We also measure the *Answer Consistency under Input Replacement* (ACIR):

$$\mathrm{ACIR}(c) = \frac{1}{N} \sum_i \mathbb{1}[\hat{y}_i(\mathrm{Full}) = \hat{y}_i(c)], \tag{2}$$

which captures whether individual predictions remain unchanged after intervention. CAR and ACIR capture complementary aspects of visual reliance: CAR measures whether visual perturbations affect final performance, whereas ACIR measures whether they alter model decisions.

Figure 1 summarizes the counterfactual results. The CAR results show that visual interventions often have limited effects on benchmark accuracy, although their effects vary across model–dataset pairs. Under the **Replaced** condition, CAR ranges from 83.2% to 107.1%, with a average of 97.1%. The **Blank** condition shows a similarly heterogeneous pattern: several configurations remain close to their full-input performance, whereas others show larger gains or drops after pathology content is removed. Even under **Text-only**, where all visual information is removed, accuracy matches or exceeds the full-input setting in 27.5% of model–dataset pairs.

However, ACIR reveals behaviors that are not reflected by accuracy alone: some interventions alter predictions despite negligible accuracy changes. These results show that visual sensitivity and effective utilization of slide-specific evidence are not equivalent. Models may change their decisions after visual perturbations without consistently benefiting from the corresponding pathology evidence.See Appendix D for full results.

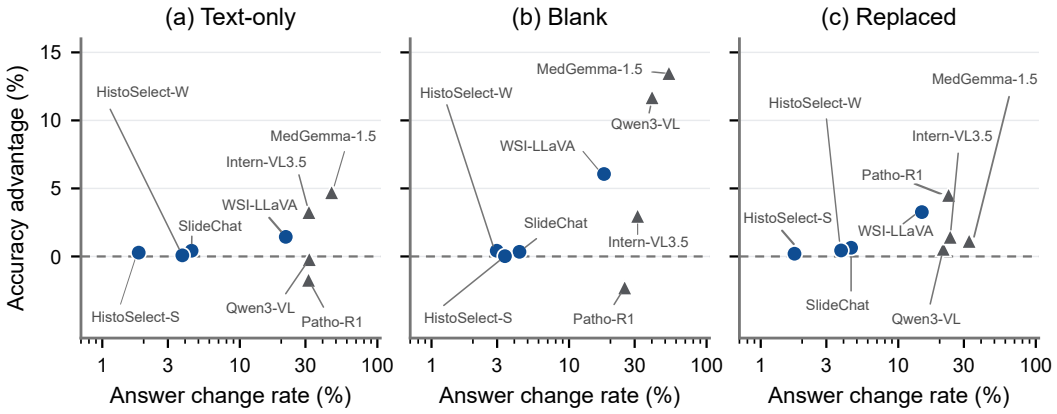

Figure 2: **Visual evidence utilization profiles across MLLMs.** Each point represents one model configuration with averages over the five accuracy-bearing datasets. **(a)** Text-only, **(b)** Blank, and **(c)** Replaced conditions. The x-axis shows the answer-change rate, and the y-axis shows the accuracy advantage of the original visual input over the corresponding counterfactual condition. The dashed line indicates zero accuracy advantage.

Takeaway 1: Benchmark performance is an insufficient measure of visual information utilization. Counterfactual evaluation reveals a gap between correct answering and reliance on the presented WSI: models may maintain predictions after visual information removal, or change predictions after visual perturbations without corresponding accuracy improvements.

5 How does visual evidence utilization vary across MLLMs?

We further characterize the behavioral patterns of different MLLMs across input paradigms. Based on the counterfactual evaluation, we analyze visual evidence utilization using two complementary measures: prediction sensitivity to visual intervention and the accuracy advantage of retaining the original visual input. For each condition family $c \in \{\text{Text-only}, \text{Blank}, \text{Replaced}\}$, we compute the mean answer-change rate $1 - \text{ACIR}_m(c)$ and the corresponding accuracy advantage across the five accuracy-bearing datasets (Figure 2).

Figure 2 reveals diverse visual response patterns. WSI-input models such as SlideChat and HistoSelect show consistently low sensitivity to visual intervention, with answer-change rates of only 1.8–4.5% and accuracy advantages of 0.02–0.64 percentage points, indicating that their predictions remain largely invariant to slide changes. WSI-LLaVA exhibits higher sensitivity, but the additional prediction changes do not consistently translate into larger accuracy gains.

General-image-input models are generally more responsive to visual changes, with answer-change rates of 21.2–53.2% across the three intervention conditions and 21.2–32.8% under the slide-specific **Replaced** condition. However, this sensitivity does not consistently correspond to stronger reliance on the queried slide evidence. In particular, Qwen3-VL, InternVL3.5, and MedGemma-1.5 show substantially larger accuracy advantages under **Blank** (11.7, 3.0, and 13.5 percentage points, respectively) than under **Replaced** (0.6, 1.4, and 1.1 percentage points). This indicates that their visual reliance is driven more by the presence of visual inputs than by slide-specific pathology evidence.

Takeaway 2: Visual responsiveness is not equivalent to visual evidence utilization. Effective use of WSI evidence requires both sensitivity to visual changes and consistent performance gains from the corresponding visual information, which are not simultaneously observed across current MLLMs.

6 How is visual information utilized in PMLLM decisions?

The previous sections measure how predictions respond to slide-level visual interventions but do not establish which regions contribute to those responses. We therefore investigate their alignment with pathologist-identified regions using VAPath, where each WSI–question pair is associated with independent pathologist-annotated ROIs. We analyze three complementary signals: visual attention, causal influence, and information retrieval from visual representations.

To compare heterogeneous spatial signals under a unified criterion, we use Average Precision (AP) to measure whether high-scoring visual units are preferentially located within pathologist-annotated ROIs. Since the three signals have different score distributions and cases may contain different ROI proportions, we further normalize AP by its expected value under random allocation:

$$\mathrm{AP}_{\mathrm{norm}} = \frac{\mathrm{AP}(\mathbf{s}, \mathbf{m}) - \mathrm{AP}_{\mathrm{random}}}{1 - \mathrm{AP}_{\mathrm{random}}}, \tag{3}$$

where $\mathbf{s}$ denotes spatial scores, $\mathbf{m}$ denotes the binary ROI mask, and $\mathrm{AP}_{\mathrm{random}}$ denotes the expected AP under random allocation. $\mathrm{AP}_{\mathrm{norm}} = 0$ indicates no improvement over random allocation, while $\mathrm{AP}_{\mathrm{norm}} = 1$ indicates that all high-scoring units are concentrated within ROIs.

Attention is weakly associated with pathologist-identified regions. We first examine whether attention highlights pathology-relevant regions. Attention from language tokens to visual tokens is aggregated and mapped back to WSI coordinates (see Appendix E.2). Figure 3 shows that attention-based signals achieve limited ROI localization performance across models. To distinguish slide-content sensitivity from structural effects, we perform a patch-order shuffle control and compare whether attention patterns are represented by token positions or by tissue identities (see Appendix E.2). Figure 4(a) shows that position-based similarity is consistently higher than identity-based similarity after shuffling, indicating that attention patterns are more strongly associated with token positions than with the underlying tissue regions.

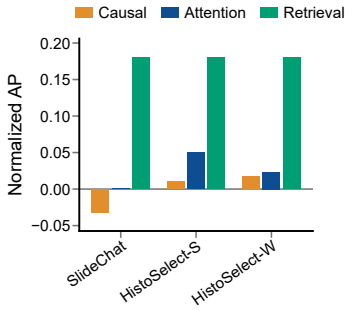

Figure 3: **ROI alignment measured by normalized AP.**

Causal visual influence is weak and poorly localized. Attention characterizes how visual information is internally allocated. We further perform controlled spatial interventions on visual representations to quantify whether individual regions causally affect model predictions. Specifically, we divide each WSI into cells and replace the corresponding patch features with blank features(see Appendix E.3). For each intervention, we measure the change in the model's answer margin:

$$\Delta_k = M(q, I) - M(q, I^{G_k \leftarrow \mathrm{blank}}), \tag{4}$$

where $M(q, I) = z_{y^*}(q, I) - \log \sum_{y \neq y^*} \exp(z_y(q, I))$ denotes the margin of the correct answer over competing candidates, I is the original visual input, $I^{G_k \leftarrow \mathrm{blank}}$ denotes the input after replacing

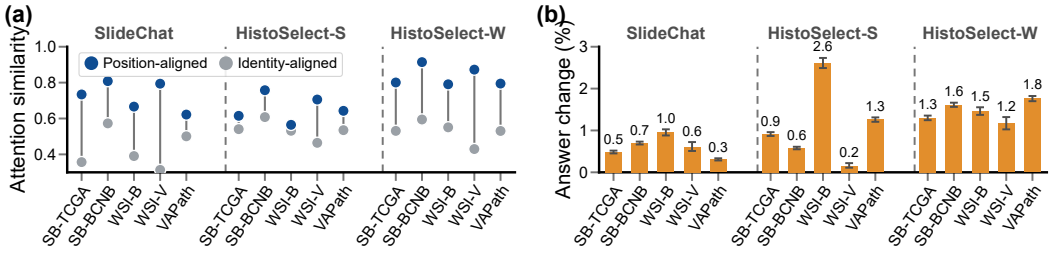

Figure 4: **Spatial controls.** **(a)** Attention similarity after patch-order shuffle. **(b)** Answer change after spatial interventions; error bars indicate 95% confidence intervals.

cell G_k with blank features, and $z_y(q, I)$ denotes the logit of candidate answer y. We use $|\Delta_k|$ to quantify the magnitude of each region's influence on the prediction.

Figure 3 shows that causal influence achieves limited ROI localization performance, with normalized AP values close to the random baseline. Moreover, spatial interventions produce negligible changes in the prediction margin (mean $\Delta_k = 0.0014 \pm 0.0972$) and rarely change the final prediction (Figure 4(b)), indicating that individual regions exert limited influence on model decisions. These findings show that visual influence is limited in both localization and strength: high-influence regions are not preferentially concentrated within pathologist-annotated ROIs and rarely become sufficient to alter final predictions.

visual information is identifiable but weakly reflected in decisions. The weak attention localization and causal influence raise the question of whether pathology-relevant visual evidence is missing from the representation or simply not effectively used by the model. We therefore evaluate an independent pathology retrieval model using answer-associated visual entities (see Appendix E.4). Retrieval achieves substantially higher normalized AP than attention or causal influence (Figure 3), indicating that task-relevant visual semantics can be recovered by an independent pathology vision-language retrieval model from the visual representation, while the evaluated PMLLM signals show limited correspondence with such evidence.

Additional experimental results and qualitative visualization analyses are provided in Appendix E.5.

Takeaway 3: Visual accessibility and decision-level utilization are not equivalent: task-relevant visual patterns can be identified from upstream WSI patch features, while evaluated PMLLM-derived signals show limited localization and influence.

7 What limits the effective utilization of visual evidence in PMLLMs?

The previous sections show limited measured effects of visual interventions on predictions in several evaluated settings. In this section, we further investigate the extent to which visual information can influence model predictions and analyze why models fail to effectively incorporate visual cues into final decisions.

7.1 Controlled slide pairing reveals weak visual reliance

Previous analyses show that visual interventions often have limited effects on model predictions. However, they do not directly assess whether models utilize slide-specific visual information. Pathology QA provides a natural setting for a stricter evaluation because many tasks follow standardized diagnostic guidelines, use template-based formulations, and share a fixed answer space across patients, while the correct answer depends on the corresponding slide. This structure enables us to isolate the influence of slide-specific information on model decisions under otherwise identical task conditions.

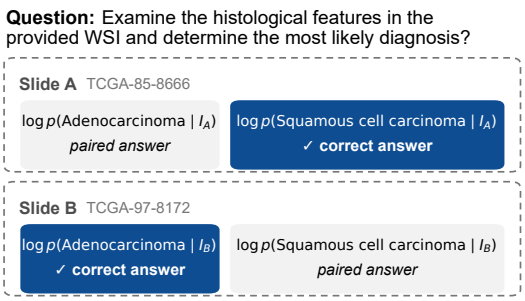

Figure 5: **Slide-swap evaluation protocol.** Paired slides share the same question and option set but correspond to different correct answers.

Specifically, we construct patient-disjoint slide pairs in which the same question and candidate answer set are paired with two different slides that require different answers based on their slide-specific pathology findings (Figure 5).Let I_a and I_b denote a slide pair, and a_a and a_b denote their corresponding correct answers. Details are provided in Appendix F.2.

We define the paired visual preference score as

$$G_{\text{pair}} = \frac{1}{2}\left[\log p(a_a|q, I_a) - \log p(a_b|q, I_a) + \log p(a_b|q, I_b) - \log p(a_a|q, I_b)\right]. \tag{5}$$

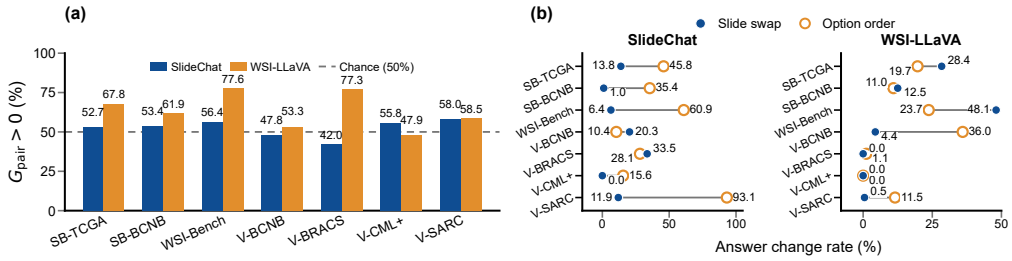

Figure 6: **Visual influence is inconsistent and rarely decisive across models.** **(a)** Fraction of slide pairs with positive paired visual preference ($G_{\text{pair}} > 0$). **(b)** Prediction change rate after slide swapping and option-order reversal.

A positive G_{pair} indicates that changing the slide identity shifts the relative answer preference toward the corresponding answer of the matched slide, suggesting that the model incorporates slide-specific information when making predictions.

Figure 6(a) shows that slide-specific visual preference remains limited overall, although its magnitude varies across models. The 50% reference corresponds to a directionally balanced outcome, where positive and negative paired preferences are equally likely after slide–answer correspondence is randomized. SlideChat stays close to this reference (52.3% average), indicating that slide changes do not consistently shift its preference toward the corresponding slide. WSI-LLaVA achieves a higher average positive preference rate (63.5%), suggesting stronger slide-specific sensitivity, although the effect remains limited on several sources.

Figure 6(b) further examines whether these slide-specific preference shifts propagate to final predictions. Slide swapping changes the selected answer in only 12.44% of samples for SlideChat and 13.41% for WSI-LLaVA on average. We further use option-order reversal as a non-visual control to distinguish visual insensitivity from general prediction stability. For SlideChat, option reversal changes predictions substantially more often than slide swapping (41.34% versus 12.44%), indicating that its weak response to slide changes is specific to slide identity rather than general prediction stability. WSI-LLaVA shows comparable sensitivity to the two perturbations, with an option-reversal rate of 14.72%. Together, these results indicate that even when slide-specific information affects likelihood-level preferences, such influence rarely propagates to final decisions.

Takeaway 4: Matched slide-pair controls provide a evaluation of slide-specific visual reliance, revealing a gap between likelihood-level visual influence and decision changes: slide-specific information may affect model preferences, but rarely determines the final prediction.

7.2 Weak visual necessity during training.

Training data contain textual cues that may provide alternative prediction paths. The limited influence of slide-specific information on final decisions motivates further analysis of whether current training settings encourage sufficient reliance on visual inputs. One possible explanation is that training samples may contain textual cues that provide alternative signals for answer prediction. To examine this possibility, we evaluate the training data used by SlideChat and WSI-LLaVA with two text-only LLMs that serve as language backbones in representative PMLLMs, Qwen2.5-7B-Instruct (Qwen et al., 2025) and Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024), without providing any visual input. Figure 7(a) shows that a substantial fraction of questions in both training corpora can be answered from text alone, with the effect consistently stronger in SlideChat.

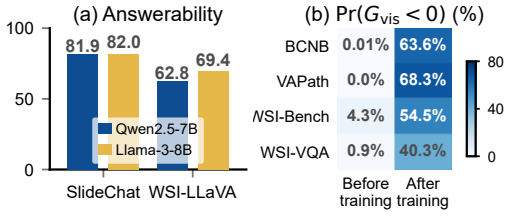

Figure 7: **Visual contribution diminishes after training.** **(a)** Text-only answerability of training samples with two backbone LLMs. **(b)** Fraction of patient-disjoint test samples with negative visual gain ($G_{\text{vis}} < 0$) before and after training.

Table 1: **Correct-answer NLL and visual gain before and after training.** $\Delta\mathrm{NLL}$ denotes the NLL reduction after training: $\mathrm{NLL}_{\mathrm{Before}} - \mathrm{NLL}_{\mathrm{After}}$.

		Before training			After training			Training improvement	
Training data	Split	w/o image	w/ image	G_{vis}	w/o image	w/ image	G_{vis}	$\Delta\mathrm{NLL}_{\mathrm{w/o}}$	$\Delta\mathrm{NLL}_{\mathrm{w/}}$
SlideChat	Train	4.8728	4.5961	$+0.2766$	0.2488	0.2480	$+0.0008$	**4.6240**	**4.3481**
	Test	8.8743	4.8005	$+4.0738$	4.4694	4.5316	-0.0622	**4.4049**	**0.2689**
WSI-LLaVA	Train	5.9682	5.9967	-0.0285	0.8220	0.7849	$+0.0371$	**5.1462**	**5.2118**
	Test	8.8743	4.8005	$+4.0738$	4.6954	4.9374	-0.2419	**4.1789**	**-0.1369**

This observation is consistent with the possibility that textual shortcuts exist in the training data: when textual cues provide strong predictive signals, optimization may not necessarily encourage stronger dependence on slide-specific visual evidence (Cadene et al., 2019; Dancette et al., 2021). Appendix F provides the full evaluation setup, implementation details, and scoring protocol.

Training improves answer fitting without strengthening visual dependence. We next investigate whether multimodal instruction tuning increases the contribution of WSI information to model predictions. Specifically, we perform controlled before-and-after training analyses by fine-tuning the SlideChat's initialization with different pathology instruction datasets, including SlideChat and WSI-LLaVA training samples. We quantify the marginal contribution of visual inputs using

$$G_{\mathrm{vis}} = \log p(y^* \mid q, I) - \log p(y^* \mid q). \tag{6}$$

A larger G_{vis} indicates that visual inputs provide greater improvement in predicting the correct answer, whereas negative values indicate that visual inputs do not improve the prediction.

As shown in Table 1, instruction tuning substantially improves answer fitting on train split across both training configurations, with large reductions in the NLL of the correct answer under both image-conditioned and text-only settings. However, across both training configurations, the reduction in NLL is substantially larger without visual inputs than with visual inputs on the test set, suggesting that the improvement is not primarily associated with additional visual contribution. Correspondingly, G_{vis} consistently decreases after training, from positive values before instruction tuning to approximately zero or negative values after training across test splits. These results indicate that instruction tuning improves answer fitting without increasing the marginal contribution of visual inputs. Figure 7(b) further shows that the proportion of samples with negative visual gain increases across evaluation sources after training.

Takeaway 5: Multimodal instruction tuning improves answer fitting but does not necessarily strengthen visual dependence: when the optimization objective can be improved without increasing the contribution of visual inputs, better prediction performance does not imply stronger reliance on WSI information.

8 Conclusion

In this work, we systematically investigate visual information utilization in pathology multimodal large models under evaluated WSI question-answering settings. Through controlled interventions, spatial auditing, and training analysis, we reveal a gap between available visual information and decision-level utilization: although current PMLLMs achieve strong benchmark performance, visual interventions often have limited effects on predictions, and independently recoverable visual semantics are not consistently reflected in model decisions. We further show that multimodal instruction tuning improves answer fitting without increasing the measured marginal contribution of WSI visual information, highlighting that benchmark accuracy alone does not fully characterize WSI understanding and motivating future evaluation and training strategies that better connect visual representations, pathology evidence, and decision making.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were used to assist with language refinement, LaTeX formatting, and limited code-related assistance. Generative AI tools were not used to generate experimental results, perform data analysis, or make scientific conclusions. We independently reviewed and verified all AI-assisted content, including text, equations, claims, and code-related materials, and take full responsibility for the final content of this work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980. IEEE, 2018.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

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*, 2025.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022.

Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 401–417. Springer, 2024a.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16123–16134. IEEE, 2022.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5134–5143. IEEE, 2025.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1554–1563. IEEE, 2021.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Camelyon16 Consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wentao Huang, Weimin Lyu, Peiliang Lou, Qingqiao Hu, Xiaoling Hu, Shahira Abousamra, Wenchao Han, Ruifeng Guo, Jiawei Zhou, Chao Chen, et al. Act like a pathologist: Tissue-aware whole slide image reasoning. *arXiv preprint arXiv:2603.00667*, 2026.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: A multimodal large language model for whole slide image. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22718–22727. IEEE, 2025.

Guimeng Liu, Tianze Yu, Somayeh Ebrahimkhani, Shawn Lin, Kok Pin Ng, and Man Cheung. How do medical mllms fail? a study on visual grounding in medical images. In *International Conference on Learning Representations*, volume 2026, pp. 95215–95268, 2026.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024a.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024b.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. 2025. URL `https://arxiv.org/abs/2412.15115`.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Asian Conference on Computer Vision*, pp. 293–309. Springer, 2024.

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.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5034–5042, 2024.

K Tomczak, P Czerwińska, and M Wiznerowicz. The cancer genome atlas (tcga): An immeasurable source of knowledge. wspolczesna onkol. 2015; 1a: A68–a77, 2014.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10): 2924–2935, 2024.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Chengyang Zhang, Wenchuan Zhang, Bo Li, Xinyu Liu, Jiaming Yang, Mengran Li, Chenxun Deng, Jie Chen, Yang Zhang, Wei Ju, et al. Do pathology vision-language models truly see pathology? *arXiv preprint arXiv:2607.21065*, 2026a.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28418–28426, 2026b.

A LIMITATIONS

Our findings characterize PMLLM configurations under the evaluated WSI question-answering tasks, prompts, and input pipelines. VAPath covers selected diagnostic tasks with fixed answer sets and curated ROIs; its purpose is to provide a controlled evaluation of visual evidence utilization rather than a comprehensive assessment of pathology reasoning trajectories or clinical decision workflows. Within this scope, the benchmark enables analysis of whether models identify and use pathology-relevant visual information, while broader evaluations are needed to study complete diagnostic processes.

Each analysis is designed to address a specific aspect of visual evidence utilization. Image removal and replacement evaluate prediction sensitivity to controlled changes in visual input, while matched slide-pair evaluations further assess whether decisions follow slide-specific evidence when task conditions are held constant. Annotated ROIs provide independent pathology references for evaluating spatial correspondence, although they represent explicit reference regions rather than a complete characterization of all diagnostic evidence. Similarly, external CONCH retrieval evaluates whether an independent procedure can identify ROI-associated visual information from upstream slide patches. This analysis characterizes the accessibility of task-relevant visual patterns in the visual pipeline rather than tracing information flow through every aggregation or projection stage. Accordingly, our conclusions focus on the relationship between identifiable visual evidence and measured decision-level influence under the evaluated protocols, without attributing observed behaviors to a specific internal module.

The spatial intervention analysis is designed to assess whether regional visual evidence contributes to model decisions beyond isolated local features. By perturbing aggregated tissue regions while preserving the overall visual context, this analysis evaluates whether distributed spatial information has measurable effects on final predictions. The limited changes in answer margins and predictions indicate that substantial regions rarely exert decisive influence under the evaluated protocols. Together with whole-slide interventions and matched slide-pair evaluations, these results provide evidence for limited decision-level reliance on slide-specific visual evidence in current PMLLMs. The training analysis addresses a complementary question: before-and-after comparisons using one model initialization and two instruction corpora show that improved answer fitting does not necessarily increase the measured contribution of visual information.

B FUTURE WORK

These findings suggest that future PMLLM evaluation should measure the contribution of slide-specific evidence alongside answer accuracy. In particular, answer-changing slide pairs, controlled visual interventions, and spatial reference annotations can reveal behaviors that conventional QA scores leave unresolved. For model development, training objectives and data construction should make relevant visual evidence consequential to the answer, while evaluation should verify that improvements persist when textual cues are insufficient and the queried slide changes. In pathology applications, strong benchmark performance should be accompanied by evidence-sensitive validation on the intended task and input pipeline before predictions are interpreted as grounded in the patient's slide.

C EXPERIMENTAL SETUP

C.1 BASELINE IMPLEMENTATIONS, CHECKPOINTS, AND INPUTS

We evaluate each model using its official implementation whenever available. For WSI-input models, we preserve the released slide-token order and modify only the content-bearing patch features under counterfactual interventions. Patch extraction uses 224×224 tissue patches at $20\times$ magnification, and features are cached before intervention so that the slide aggregator, projection module, and language model are re-evaluated under each condition. For conventional image-input MLLMs, we retain the model-specific image processor and replace only the raster image input. Table 2 summarizes the implementation details and official code sources used for reproducible evaluation.

Table 2: **Implementation details of evaluated models.**

Model	Input modality	Intervention-preserving preprocessing	Official implementation
SlideChat	WSI patch features	Cached CONCH 224×224, $20\times$ patch features; preserve token order and slide aggregation	`https://github.com/uni-medical/SlideChat`
WSI-LLaVA	WSI slide features	Released WSI tokenizer/projector; replace patch features after tokenization	`https://github.com/XinhengLyu/WSI-LLaVA`
HistoSelect	WSI patch features	Released selection order and backbone interface; no reimplementation of selection modules	`https://github.com/winston52/HistoSelect`
Patho-R1	Raster image input	Original image processor unchanged; interventions applied before preprocessing	`https://github.com/Wenchuan-Zhang/Patho-R1`
InternVL3.5	Raster image input	Official dynamic tiling and conversation template	`https://github.com/OpenGVLab/InternVL`
Qwen3-VL	Raster image input	Official processor and visual token configuration	`https://github.com/QwenLM/Qwen3-VL`
MedGemma-1.5	Raster image input	Official processor and prompting interface	`https://github.com/Google-Health/medgemma`

Checkpoint loading and decoding. All checkpoints are evaluated in inference mode using the official released tokenizers and processors. We use deterministic decoding with temperature 0 for answer generation and compute candidate likelihoods or answer margins under teacher forcing when required. Across Full, Blank, and Replaced conditions, we keep the prompt format, candidate order, maximum generation length, and image-token budget identical. The Text-only condition removes visual inputs and therefore does not use an image-token budget. For slide replacement experiments, donor slides are sampled from the same evaluation source with disjoint patient identities and processed using the same preprocessing pipeline as the original slides.

C.2 Scope of evaluated models

Our evaluation focuses on pathology multimodal large language models (PMLLMs) that directly process whole-slide images or WSI-derived visual representations for question answering. We distinguish these models from conventional image-input pathology MLLMs, which operate on cropped or rasterized images rather than native WSI inputs. The latter models are included only as reference baselines to characterize the difference between WSI-oriented and image-input multimodal systems.

For WSI-based models, we evaluate each model according to its available implementation and representation design. SlideChat is used as the primary baseline for spatial and causal analyses because it is a representative open-source WSI PMLLM with publicly available training data and inference pipeline. HistoSelect is further included for comparison because it adopts the same publicly available instruction-tuning data, enabling a controlled analysis of model behavior under related training conditions.

Not all WSI PMLLMs support spatial intervention analysis. In particular, WSI-LLaVA uses precomputed Gigapath slide representations in its released inference pipeline. These slide-level representations aggregate patch information before model input and do not preserve the original patch-level spatial correspondence required for controlled region-level intervention. Therefore, WSI-LLaVA is evaluated in prediction-based analyses but is not included in spatial causal intervention experiments.

For conventional pathology foundation model pipelines and MIL-based methods, we do not include them as primary PMLLM comparisons. MIL approaches generally provide slide-level prediction through learned aggregation rather than generating answers conditioned on open-ended visual question prompts, and therefore cannot be directly evaluated under the same MCQ-based VQA framework.

C.3 Treatment of existing WSI benchmarks

Existing WSI benchmarks differ in answer formats, slide identifiers, and task formulations. We convert each sample into a unified tuple $(I, q, \mathcal{A}, y^*, g)$, where I denotes the slide, q the question, $\mathcal{A}$ the ordered semantic answer set, y^* the correct semantic label, and g the patient identifier. Answer strings are normalized by lower-casing, whitespace normalization, option-marker removal, and dataset-specific alias mapping. These transformations are used only for output evaluation and do not introduce additional clinical information into the model input.

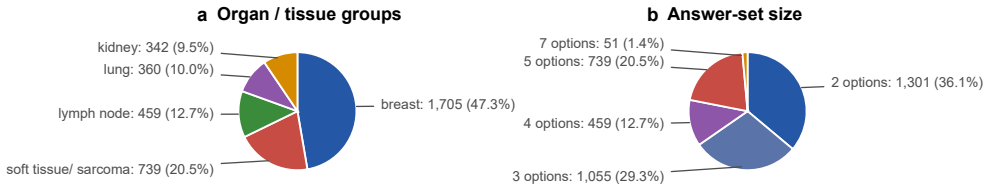

Figure 8: **VAPath composition by organ group and answer-set size.** (a) Organ/tissue-group composition. (b) Number of answer options per sample. Direct labels report the sample count and percentage of the 3,605-sample release; labels for small wedges use guide lines.

C.4 VAPath construction

Dataset composition. VAPath is a curated whole-slide image question-answering benchmark with explicit visual anchors. The dataset contains 3,605 WSI–QA–ROI samples from 3,556 unique WSIs and 2,940 case entities. It covers seven source groups: BCNB, BRACS, CAMELYON+, SARC, and three TCGA-derived cancer groups (breast, kidney, and lung cohorts). Table 3 summarizes the source-level composition and task assignment.

Table 3: **VAPath source and task composition.**

Source	Task	#Class	Samples	WSIs	Patients	ROIs
BCNB	Tumor type	3	974	974	974	8,640
BRACS	Lesion class	3;7	132	83	56	2,278
CAMELYON+	Metastasis	4	459	459	283	5,160
SARC	Sarcoma subtype	5	739	739	398	1,570
TCGA-Breast	Histological subtype	2	599	599	586	649
TCGA-Kidney	Histological subtype	3	342	342	330	862
TCGA-Lung	Histological subtype	2	360	360	313	797
Total	–	–	3,605	3,556	2,940	19,956

Annotation representation and harmonization. VAPath consolidates heterogeneous spatial annotations from multiple pathology sources into a unified polygon-based representation. For public datasets, including BCNB, BRACS, CAMELYON+, and TCGA, we retain ROI annotations released with the original datasets. These annotations were generated under source-specific annotation protocols with documented human supervision and quality-control procedures, providing a reliable and traceable foundation for benchmark construction.

For the private SARC cohort, ROI annotations were newly generated through a blinded pathology annotation process. Regions were independently annotated by attending pathologists and subsequently reviewed by a senior pathologist with more than 10 years of diagnostic experience to ensure annotation consistency and clinical relevance.

To enable unified spatial analysis, all retained ROI contours are converted into normalized coordinates relative to the original WSI dimensions and stored using a consistent polygon representation across sources. After harmonization and quality filtering, VAPath contains 19,956 polygon contours, which serve as explicit visual anchors for ROI-based evaluation.

Question construction and quality control. VAPath questions are constructed from visually observable morphological patterns and do not use pathology reports, demographics, treatment history, or other non-visual metadata. We design 12 question templates covering major diagnostic tasks, including tumor subtyping, lesion classification, metastasis assessment, and cancer subtype identification (Table 4). Each question is paired with a fixed semantic option set, and the correct answer is determined by the corresponding slide-level annotation.

After automated preprocessing, candidate samples undergo quality control to ensure reliable visual evidence and task consistency. All candidate samples are independently reviewed by two pathologists, with disagreements resolved through consensus discussion with a senior consultant pathol-

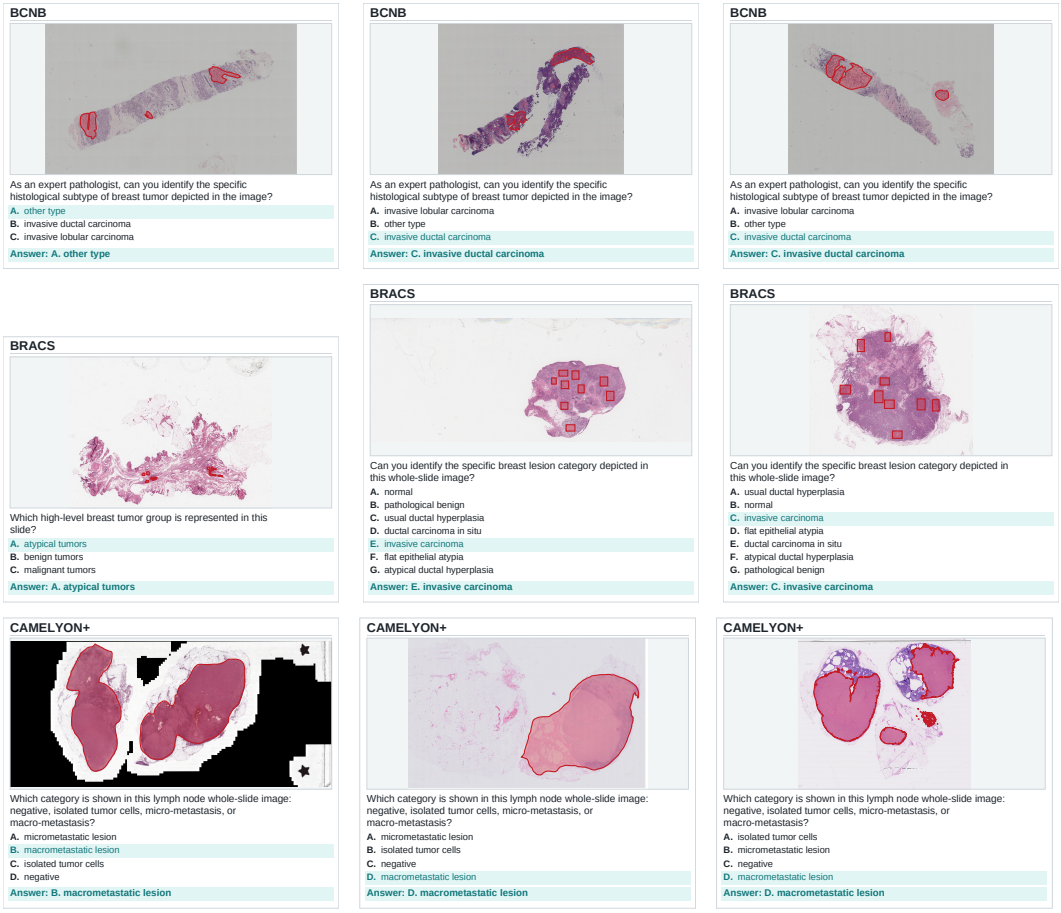

Figure 9: **Representative VAPath samples: BCNB, BRACS, and CAMELYON+.** Rows correspond to one source each and show three examples with their WSI thumbnails, spatial annotations, questions, candidate options, and correct answers.

ogist. The quality-control process removes samples with incompatible labels or spatial annotations, including BRACS samples whose filtered regions do not match the target classification space, CAMELYON+ contours corresponding to normal or exclusion regions, and non-diagnostic marker annotations in TCGA. Spatial annotations are further validated by removing invalid polygons and clipping partially out-of-bound contours to WSI boundaries. Samples with ambiguous diagnosis, insufficient visual support, duplicate content, or potential non-visual shortcuts are excluded.

For the final dataset, answer options and labels are normalized to lowercase semantic strings. Within identical option sets, option orders are deterministically shuffled with fixed seed to balance answer positions while preserving the original task definition. The final 3,605 WSI–QA–ROI samples pass the quality-control procedure and are used for evaluation.

Representative samples. Figures 9–11 show three representative samples for each VAPath source. Each sample card preserves the WSI thumbnail, the spatial annotation, the visually grounded question, the candidate options, and the correct answer.

D Complete Counterfactual Results

Table 5 provides the complete Q1/Q2 counterfactual evaluation results corresponding to Figures 1 and 2. Full denotes the accuracy under the original-slide condition. Counterfactual conditions include visual removal and replacement interventions. For Blank conditions, the intervened regions are replaced with blank images, which are processed through the same CONCH feature extraction

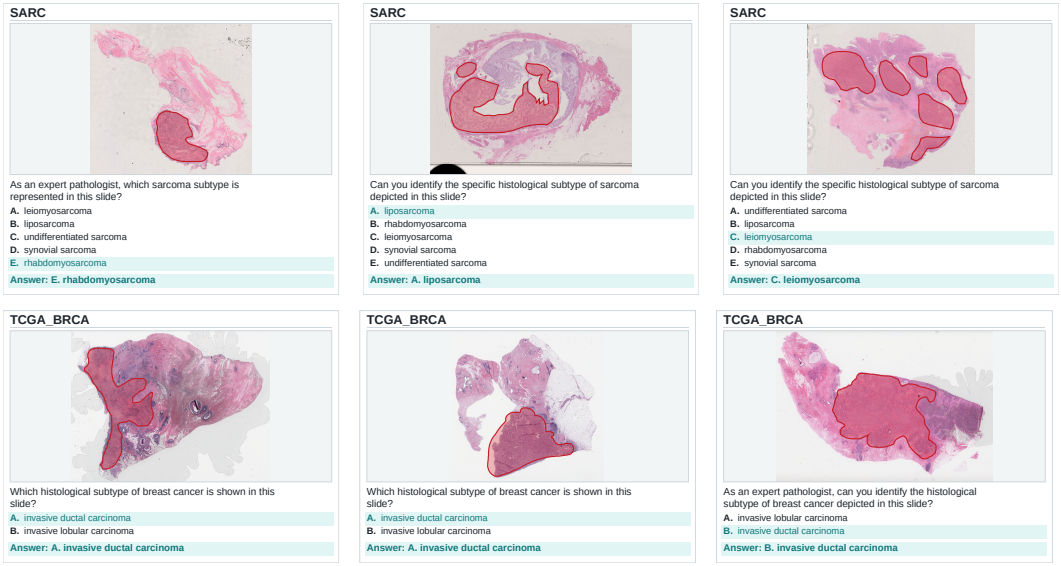

Figure 10: **Representative VAPath samples: SARC and TCGA-Breast.** The top row contains sarcoma-subtype examples; the bottom row contains breast cancer histological-subtype examples.

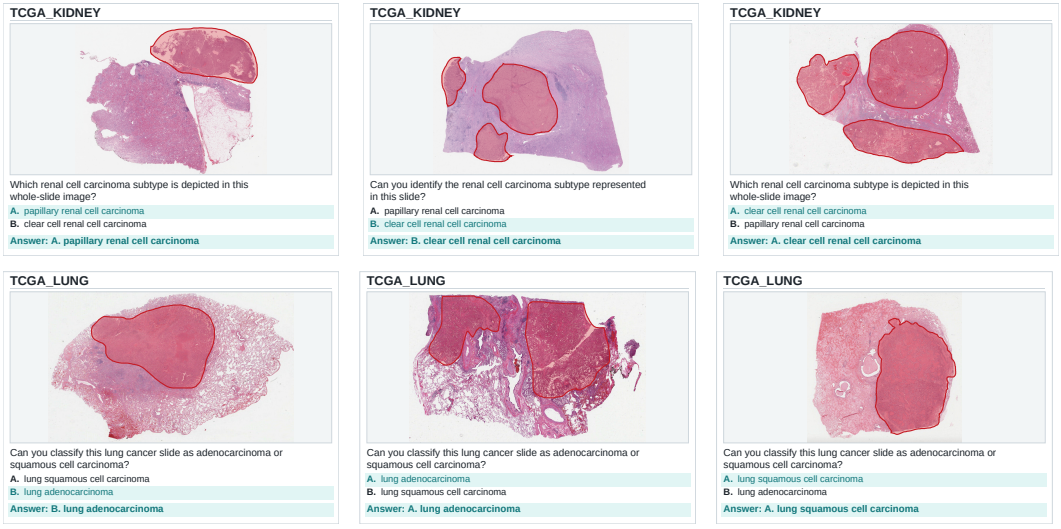

Figure 11: **Representative VAPath samples: TCGA-KIRC and TCGA-LUAD.** The top row contains renal cell carcinoma examples; the bottom row contains lung cancer histological-subtype examples.

pipeline to preserve the input representation format and feature distribution. ΔAcc is computed as Full accuracy minus the accuracy under each counterfactual condition (in percentage points), where positive values indicate performance degradation after removing or modifying the original visual input. Consistency denotes ACIR (%) computed from paired predictions of the same samples across the original and counterfactual conditions. Replaced-slide interventions are designed as slide-level perturbations rather than complete removal of all slide-specific visual cues, because replacement slides may retain generic or partially overlapping visual characteristics. Therefore, these interventions evaluate whether model predictions are sensitive to changing the provided slide context, rather than isolating the effect of eliminating every original-slide feature.

Table 4: **Question templates used in VAPath.** Templates are instantiated according to the corresponding pathology task while preserving the same visual evidence requirement.

Task category	Question template
Breast tumor subtype	As an expert pathologist, can you identify the specific histological subtype of breast tumor depicted in the image?
Breast tumor subtype	Based on the histopathological appearance, which histological subtype of breast tumor is shown in this slide?
Breast lesion classification	Which high-level breast tumor group is represented in this slide?
Breast lesion classification	Can you identify the specific breast lesion category depicted in this whole-slide image?
Lymph node metastasis classification	As a pathologist, which category best describes this lymph node slide?
Lymph node metastasis classification	Which category is shown in this lymph node whole-slide image: negative, ITC, micro-metastasis, or macro-metastasis?
Sarcoma subtype	As an expert pathologist, which sarcoma subtype is represented in this slide?
Sarcoma subtype	Can you identify the specific histological subtype of sarcoma depicted in this slide?
Renal cell carcinoma subtype	Can you identify the renal cell carcinoma subtype represented in this slide?
Renal cell carcinoma subtype	Which renal cell carcinoma subtype is depicted in this whole-slide image?
Lung cancer subtype	Can you classify this lung cancer slide as adenocarcinoma or squamous cell carcinoma?
Lung cancer subtype	Which histological subtype of lung cancer is shown in this slide?

E Spatial-Signal Computation and Result Interfaces

E.1 Normalized Average Precision for spatial alignment

To quantitatively compare heterogeneous spatial signals, we measure whether visual units receiving higher scores are preferentially located within pathologist-annotated ROIs. Let $\mathbf{s} = \{s_i\}_{i=1}^{P}$ denote the spatial scores over P visual units, where larger values indicate stronger responses from a specific signal. Let $\mathbf{m} = \{m_i\}_{i=1}^{P}$ denote the corresponding binary ROI mask, where $m_i = 1$ indicates that visual unit i overlaps with a pathologist-annotated ROI and $m_i = 0$ otherwise.

We treat the spatial score map as a retrieval problem, where visual units are ranked according to their scores s_i. For a given ranking, let $R(k)$ denote the number of ROI-positive units among the top-k ranked visual units. The precision at rank k is defined as

$$\mathrm{Prec}(k) = \frac{R(k)}{k}. \tag{7}$$

The Average Precision is computed as the average precision over all ROI-positive units:

$$\mathrm{AP}(\mathbf{s}, \mathbf{m}) = \frac{1}{\sum_i m_i} \sum_{k=1}^{P} \mathrm{Prec}(k) m_{\pi(k)}, \tag{8}$$

where $\pi(k)$ denotes the visual unit ranked at position k according to descending spatial scores. Higher AP indicates that high-scoring visual units are more concentrated within pathologist-annotated ROIs.

Because different samples may contain different ROI proportions and the three spatial signals have different score distributions, we normalize AP with respect to its expected value under random allocation:

$$\mathrm{AP}_{\mathrm{norm}} = \frac{\mathrm{AP}(\mathbf{s}, \mathbf{m}) - \mathrm{AP}_{\mathrm{random}}}{1 - \mathrm{AP}_{\mathrm{random}}}, \tag{9}$$

where $\mathrm{AP}_{\mathrm{random}}$ denotes the expected AP when visual units are randomly ranked. Therefore, $\mathrm{AP}_{\mathrm{norm}} = 0$ corresponds to random spatial allocation, while $\mathrm{AP}_{\mathrm{norm}} = 1$ indicates perfect localization of ROI units at the highest-ranked positions. Negative values indicate worse-than-random alignment.

Table 5: **Complete counterfactual result grid for Q1 and Q2.**

Configuration	Dataset	Full	ΔAcc (pp)			Consistency (%)		
			Text	Blank	Repl.	Text	Blank	Repl.
SlideChat	SlideBench-TCGA	74.15	+0.7	+1.1	+1.5	94.9	94.2	94.0
	SlideBench-BCNB	53.73	+0.4	-0.2	-0.0	97.8	98.5	98.5
	WSI-Bench	81.60	+2.1	+0.8	+0.4	94.9	92.0	95.6
	WSI-VQA	59.49	-1.8	+0.0	+1.3	92.3	95.9	92.3
	VAPath	41.58	+0.7	+0.0	+0.0	97.8	97.6	96.9
WSI-LLaVA	SlideBench-TCGA	64.20	+4.5	+5.5	+7.0	80.5	80.3	79.8
	SlideBench-BCNB	45.74	-4.1	+3.9	+1.0	72.0	82.1	89.5
	WSI-Bench	87.89	+3.4	+4.7	+1.7	89.8	87.4	91.7
	WSI-VQA	54.87	+4.4	+5.6	+5.1	73.1	78.5	82.1
	VAPath	46.66	-1.0	+10.6	+1.5	76.7	81.9	82.7
HistoSelect-W	SlideBench-TCGA	73.58	+0.2	+0.1	+0.5	96.4	97.0	96.5
	SlideBench-BCNB	51.33	-0.3	-0.3	+0.2	97.8	98.3	98.4
	WSI-Bench	81.06	+1.2	+0.9	+0.0	96.7	96.9	96.4
	WSI-VQA	60.00	-1.8	+0.5	+0.8	93.3	95.6	92.8
	VAPath	42.05	+1.1	+0.9	+0.7	96.7	97.3	96.7
HistoSelect-S	SlideBench-TCGA	73.33	+0.1	-0.2	-0.1	98.2	97.1	98.4
	SlideBench-BCNB	51.70	+0.1	-0.2	-0.1	97.9	98.3	98.6
	WSI-Bench	79.97	+0.7	-0.4	+0.8	97.4	94.8	97.2
	WSI-VQA	61.03	+0.5	+1.8	+0.8	98.2	96.1	98.5
	VAPath	41.30	+0.0	-0.9	-0.4	99.1	96.6	98.5
Patho-R1	SlideBench-TCGA	43.65	-6.7	-4.0	+5.2	63.3	68.0	69.4
	SlideBench-BCNB	41.69	-8.7	-6.7	+0.6	68.6	77.1	83.7
	WSI-Bench	55.75	+5.4	+13.2	+1.5	75.4	77.3	80.4
	WSI-VQA	42.05	+0.4	-7.0	+6.9	59.0	73.6	76.7
	VAPath	50.06	+1.0	-6.9	+8.4	75.7	77.1	73.7
Qwen3-VL	SlideBench-TCGA	47.64	-3.1	+0.9	+0.9	71.3	65.2	78.9
	SlideBench-BCNB	53.67	+1.3	+27.5	-0.1	62.1	40.9	77.8
	WSI-Bench	62.03	+0.2	+3.8	-0.4	81.4	75.0	87.5
	WSI-VQA	45.90	+1.8	+4.1	+0.5	63.3	61.8	79.5
	VAPath	54.23	-1.2	+22.2	+1.9	61.9	56.3	70.1
Intern-VL3.5	SlideBench-TCGA	57.42	+4.2	+3.1	+2.7	68.7	69.6	75.1
	SlideBench-BCNB	57.68	+2.7	+2.6	+0.1	64.3	64.9	72.6
	WSI-Bench	71.87	+2.5	+2.5	+0.8	81.1	80.6	83.6
	WSI-VQA	48.46	+3.2	-1.5	+0.5	61.5	63.1	76.2
	VAPath	60.67	+3.7	+8.2	+3.1	65.4	64.2	73.2
MedGemma-1.5	SlideBench-TCGA	47.69	+7.8	+8.0	+3.6	48.1	49.6	66.7
	SlideBench-BCNB	50.96	+3.8	+15.2	+1.2	57.3	38.9	66.0
	WSI-Bench	59.62	+8.1	+13.6	+2.3	58.3	53.4	69.2
	WSI-VQA	47.16	+11.8	+15.6	+1.8	49.2	45.4	64.1
	VAPath	44.94	-7.9	+15.0	-3.2	54.6	46.5	70.2

The same $\mathrm{AP}_{\mathrm{norm}}$ computation is applied to attention scores, causal intervention scores, and external retrieval scores, enabling direct comparison of their spatial correspondence with pathology-relevant regions.

E.2 Attention computation

Let $V = \{1, \ldots, P\}$ denote the visual tokens, and let p denote the last token position of the multimodal prompt before answer generation or candidate scoring. For transformer layer ℓ and head h, $A_{\ell h}[p, i]$ represents the attention weight from the prompt token to visual token i. We compute the visual attention score as

$$a_i = \frac{1}{|\mathcal{L}||\mathcal{H}|} \sum_{\ell \in \mathcal{L}} \sum_{h \in \mathcal{H}} A_{\ell h}[p, i], \qquad \bar{a}_i = \frac{a_i}{\sum_{j \in V} a_j}. \tag{10}$$

The same layer and head sets are used for all samples of each model. The obtained token-level scores are mapped back to spatial locations according to the corresponding visual-token support. Attention scores are non-negative and are evaluated as spatial signals using the normalized Average Precision ($\mathrm{AP}_{\mathrm{norm}}$) defined in the main text.

To analyze whether attention reflects tissue content or token organization, we perform a patch-order shuffle control. Given a permutation π over visual tokens, we compute the position-aligned similarity

$$S_{\mathrm{pos}} = \mathrm{corr}(\bar{\mathbf{a}}, \bar{\mathbf{a}}^{\pi}), \tag{11}$$

and the identity-aligned similarity

$$S_{\mathrm{id}} = \mathrm{corr}(\bar{a}_i, \bar{a}^{\pi}_{\pi^{-1}(i)}). \tag{12}$$

A larger S_{pos} than S_{id} indicates that attention patterns are better preserved by token positions than by the underlying tissue regions.

E.3 Causal influence computation

We perform controlled spatial interventions by partitioning each slide into an 8×8 grid and retaining non-empty cells $\{G_k\}_{k=1}^{K}$. For each cell, the corresponding raw CONCH patch features are replaced with matched blank features, while patch features outside the cell and textual inputs are left unchanged. Model-specific aggregation, selection, and projection are then re-evaluated, so downstream visual tokens need not remain identical to the original run.

For each intervention, the model output margin is measured as

$$\Delta_k = M(q, I) - M(q, I^{G_k \leftarrow blank}), \tag{13}$$

where $M(q, I)$ denotes the margin of the correct answer over alternative candidates. Positive Δ_k indicates that removing cell G_k decreases the correct-answer margin.

The intervention pipeline follows each model's original visual processing procedure. In SlideChat, the visual aggregation module and multimodal projector are re-evaluated after feature replacement while preserving the visual-token structure. In HistoSelect, the visual selection module is also re-evaluated, allowing the selected tokens to adapt after intervention.

For ROI analysis, each cell is assigned a score $s_k = |\Delta_k|$. Scores are normalized over valid non-empty cells and compared with pathology-annotated regions using $\mathrm{AP}_{\mathrm{norm}}$. Samples without valid cells or evaluable interventions are excluded before aggregation. We report source-level statistics with patient- or slide-clustered bootstrap confidence intervals.

Algorithm 1: Cell-wise causal influence computation

Input: Raw CONCH feature sequence F, question q, candidate set $\mathcal{A}$, correct answer y^*, cells $\{G_k\}_{k=1}^K$, blank feature b
Output: Signed cell influences $\{\Delta_k\}_{k=1}^K$ and ROI-alignment scores
Compute the answer margin: $M(X) = z_{y^*}(X) - \log \sum_{y \in \mathcal{A} \setminus \{y^*\}} \exp(z_y(X))$
for $k = 1$ **to** K **do**
  Initialize $F^{(k)} \leftarrow F$
  Replace patch features assigned to G_k with blank feature b
  Obtain intervened input $X^{(k)}$ using the model-specific visual pipeline
  Compute cell influence: $\Delta_k = M(X) - M(X^{(k)})$
  Store prediction change indicator: $\mathbb{1}[\arg\max_y z_y(X) \neq \arg\max_y z_y(X^{(k)})]$
Set ROI score $s_k \leftarrow |\Delta_k|$ and compute ROI alignment
return $\{\Delta_k\}_{k=1}^K$ and ROI-alignment scores

E.4 External retrieval computation

External retrieval provides an independent evaluation of whether answer-relevant visual information remains accessible from the visual representation. We use a frozen CONCH model to compute image-text similarity between WSI patches and candidate answer descriptions. For each visual patch x_i, the image encoder produces a feature $v_i = f_{\text{img}}(x_i)$, while each candidate answer $a \in \mathcal{A}$ is encoded by the text encoder as $u_a = f_{\text{text}}(a)$.

We define the retrieval score of patch i as the maximum similarity with the candidate answers:

$$r_i = \max_{a \in \mathcal{A}} \frac{v_i^\top u_a}{\|v_i\|_2 \|u_a\|_2}, \qquad s_i^{\text{ret}} = \max(r_i, 0). \tag{14}$$

The patch score takes the maximum similarity over all candidate answer descriptions, regardless of which candidate is correct for the slide. Consequently, ROI alignment evaluates whether candidate-associated patterns receive higher spatial scores inside annotated regions. It does not test whether retrieval selects the correct answer, distinguishes the correct candidate from distractors, or solves the QA task.

The resulting patch-level scores are mapped back to the WSI spatial lattice and evaluated against pathology-annotated ROIs using the same AP_{norm} metric as attention and cell-wise intervention signals. This retrieval analysis does not involve PMLLM predictions or probe the representations after model-specific aggregation, selection, or projection. For a pipeline that consumes CONCH patch features, it provides a reference related to an upstream feature family used by that pipeline; for a pipeline with different visual inputs, it remains an external reference computed from WSI patches.

E.5 Qualitative visualization of spatial signals

E.6 Spatial alignment results

Table 6: **Complete spatial alignment results.** Source-level normalized Average Precision (AP_{norm}) for attention, causal influence, and external retrieval with respect to pathologist-annotated ROIs. Higher values indicate stronger alignment beyond the random-allocation expectation; bold denotes the highest value in each row.

	SlideChat		HistoSelect-S		HistoSelect-W		Retrieval
Source	Attn.	Causal	Attn.	Causal	Attn.	Causal	CONCH
BCNB	0.00	-0.03	**0.07**	0.02	0.05	0.01	0.03
BRACS	0.00	-0.01	0.00	0.07	0.01	0.02	**0.08**
CAMELYON+	0.01	-0.02	0.09	0.02	0.06	0.05	**0.12**
SARC	-0.08	-0.03	0.12	-0.01	-0.01	0.02	**0.28**
TCGA	0.02	-0.04	0.02	-0.01	0.02	0.01	**0.25**

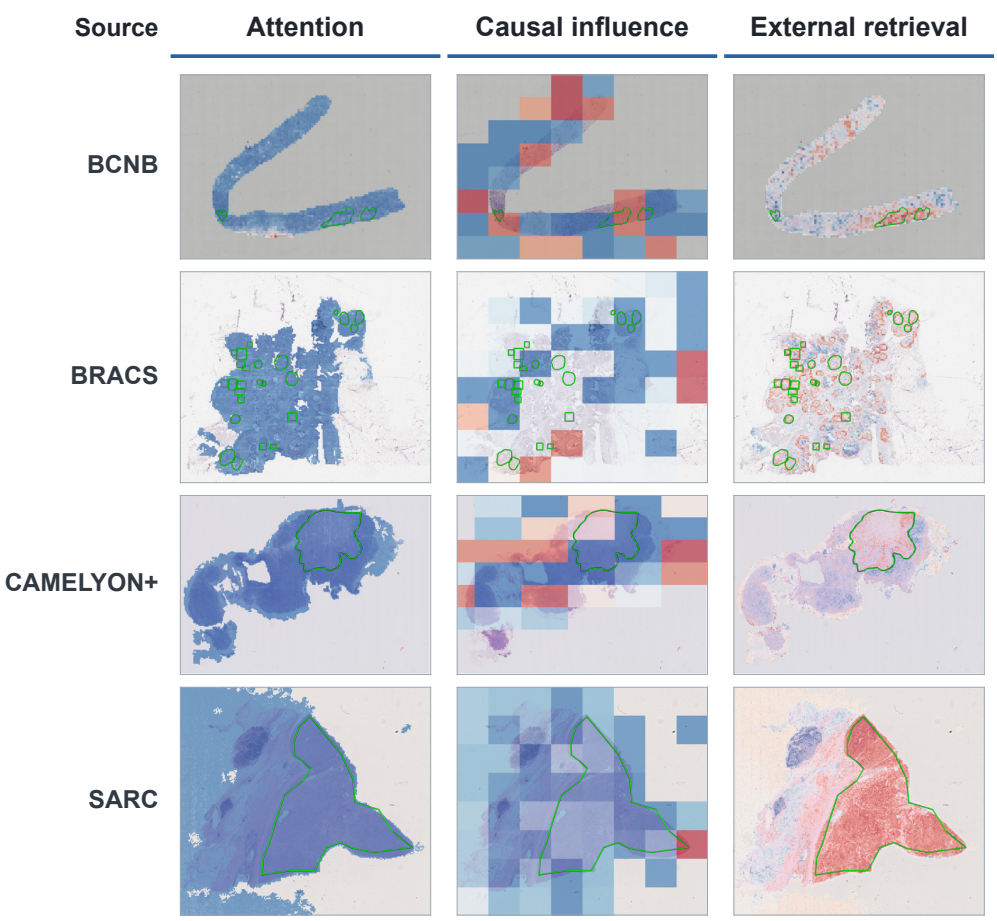

Figure 12: **Qualitative visualization of spatial signals.** Each row contains one representative sample from a VAPath source. Columns show spatial maps derived from attention, causal intervention, and external retrieval, respectively. Green contours mark pathologist-annotated ROIs. The continuation presents the remaining three sources.

F Training and Paired-Evaluation Details

F.1 Text-only backbone implementation

For the text-only answerability audit, we use the official Qwen2.5-7B-Instruct release (`https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`) and Meta-Llama-3-8B-Instruct release (`https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`), each loaded with its official chat template. Images, slide identifiers, ROI descriptions, and all image-derived metadata are omitted. For a question q and candidate a, we score

$$\text{score}(a \mid q) = \frac{1}{|a|} \sum_{t=1}^{|a|} \log p(a_t \mid q, a_{<t}), \qquad \hat{a} = \arg\max_{a \in \mathcal{A}} \text{score}(a \mid q), \tag{15}$$

which avoids confounding candidate length with likelihood. The exact prompt contains the question, the complete ordered candidate set, and a request to return the selected option only. The scoring implementation uses the released tokenizer, disables sampling, and records the model revision and prompt template hash.

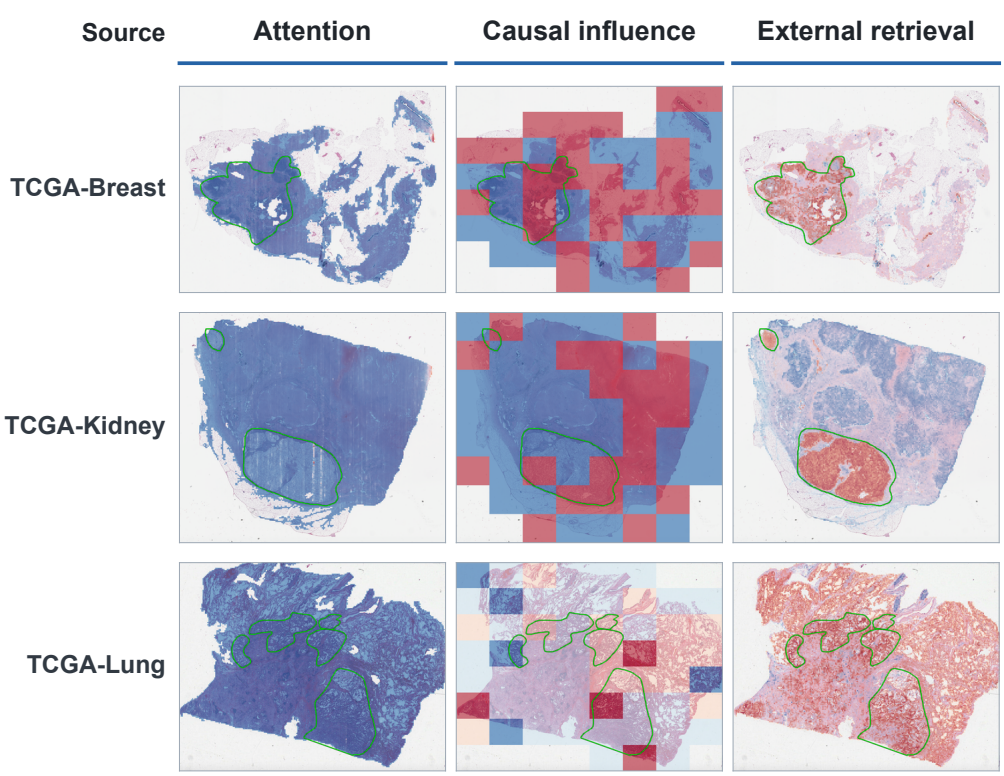

Figure 12 (continued).

F.2 Paired evaluation-set construction and statistics

The slide-pair evaluation is designed to isolate slide-specific visual influence under otherwise identical task conditions. Rather than comparing arbitrary QA samples, we construct matched pairs where the question formulation, task identity, and semantic option space are fixed, while only the underlying WSI changes. This design allows us to evaluate whether model preference follows slide-specific pathology evidence.

Canonicalization of QA items. For each source, we first normalize questions and answer options following Appendix C.3. Question identity is defined at the task level rather than exact string matching. Specifically, we normalize prompt templates, option meanings, and answer labels into canonical semantic representations. Each item is represented as

$$(I, q, \mathcal{A}, y^*, g),$$

where I denotes the slide, q the canonical question, $\mathcal{A}$ the semantic option set, y^* the semantic gold answer, and g the patient identifier.

Two samples form a valid slide pair only if they satisfy all of the following criteria:

(1) they share the same canonical task identity and semantic option set;

(2) they originate from different patients;

(3) their semantic gold answers are different.

The original answer letters are not used for pair construction because option ordering varies across datasets. Instead, all answers are mapped to semantic labels before matching. For a valid pair (I_a, I_b), the two slides therefore correspond to matched tasks with different correct answers (a_a, a_b).

Pair construction and dependency control. After canonicalization, items are grouped by $(source, task)$ and used to construct bounded pair graphs. We remove repeated slide appearances within a pair, duplicate or uncertain targets, and samples overlapping with instruction-tuning data. To avoid dependency inflation, the graph construction applies the following constraints:

(1) each pair contains different endpoint slides and patients;

(2) each endpoint slide has bounded degree;

(3) connected components contain limited numbers of endpoint slides;

(4) under-represented answer combinations are preferentially retained.

The resulting connected components are used as bootstrap units rather than individual pairs, preventing multiple pairs involving the same slide from being treated as independent observations.

For each valid pair, both slide assignments are evaluated under identical prompts, candidate ordering, and decoding settings. The paired visual preference score G_{pair} is then computed from the two directional evaluations:

$$G_{\text{pair}} = \frac{1}{2}\left[\log p(a_a|q, I_a) - \log p(a_b|q, I_a) + \log p(a_b|q, I_b) - \log p(a_a|q, I_b)\right].$$

Dataset statistics. Table 7 summarizes the resulting paired evaluation sets. We report the number of canonical QA items, endpoint slides participating in the pair graph, and final valid pairs. Pair counts are descriptive statistics; uncertainty estimation is performed at the component level rather than treating individual pairs as independent samples.

Table 7: **Statistics of the paired evaluation sets.** Canonical items denote unique slide-task samples after semantic normalization. Pair-endpoint slides are slides participating in the final bounded pair graph. Pairs are descriptive counts and are not treated as independent statistical units.

Source	Canonical items	Pair-endpoint slides	Task groups	Pairs
SlideBench-TCGA	3,176	574	75	1,011
WSI-Bench	879	107	20	156
SlideBench-BCNB	6,581	1,058	7	8,703
VAPath-BCNB	974	446	1	364
VAPath-BRACS	132	83	2	176
VAPath-CAMELYON+	459	459	1	754
VAPath-SARC	739	739	1	1,432
Total	12,940	3,466	–	12,596

F.3 Model and Controlled Training Setup

Model initialization. For the controlled experiments using the SlideChat and WSI-LLaVA training corpora, we fine-tuned the same SlideChat architecture under an identical training protocol, with only the source training corpus changed. The model consists of a LongNet-based slide encoder, a visual projector, and Qwen2.5-7B-Instruct as the language model. Each experiment was initialized from the official stage-1 SlideChat checkpoint. Whole-slide images were represented using pre-extracted CONCH patch features, which were processed by the slide encoder and projector before being integrated into the multimodal chat prompt.

Training configuration. Both training conditions followed the official SlideChat instruction tuning pipeline with the default optimization configuration. The LongNet slide encoder and pretrained language model were kept frozen, while the visual projector and the trainable adaptation components specified by the SlideChat training procedure were optimized. Training used the corresponding SlideChat and WSI-LLaVA training corpora without additional data modification. Each condition was trained with three independent random seeds, and we report the average performance across the three runs.

Optimization. Training was performed for one epoch using the SlideChat default optimization settings. We used AdamW optimization with gradient clipping and mixed-precision training. The effective batch size and learning-rate configuration followed the official SlideChat implementation.

Before–after NLL evaluation. We evaluated the official stage-1 checkpoint and the fine-tuned models using identical prompts, reference-answer tokenization, and evaluation procedures. For each example, the reference answer was scored with and without the corresponding slide visual features. The negative log-likelihood (NLL) was averaged over answer tokens and then over all evaluation examples. We defined the marginal visual contribution as

$$G_{\mathrm{vis}} = \mathrm{NLL}_{\mathrm{w/o\ image}} - \mathrm{NLL}_{\mathrm{w/\ image}}, \tag{16}$$

where positive values indicate that visual input improves the likelihood of the reference answer. Before-training results were obtained from the shared stage-1 checkpoint, while after-training results were averaged across the three fine-tuned runs.

Table 8: Controlled fine-tuning configuration used for both training-corpus conditions.

Setting	Value
Initialization	Official SlideChat stage-1 checkpoint
Trainable components	SlideChat default trainable modules
Training data	SlideChat or WSI-LLaVA training corpus
Training epochs	1
Optimizer	AdamW
Precision	Mixed precision training
Random seeds	3 independent runs