XPATHAGENT: CONCEPT-TO-EVIDENCE REASONING VIA RELIABILITY-GUIDED ROUTING ON HETEROGENEOUS PATHOLOGY TOOLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Computational pathology provides diverse specialist models spanning organs, spatial resolutions, and tissue properties. Integrating these heterogeneous pathology tools is essential for whole-slide image analysis, yet existing pathology agent methods rely on fixed foundation models and validation strategies, limiting generalizability while providing limited transparency into how tool outputs are selected, validated, and integrated. We present XPATHAGENT, a plug-and-play framework for concept-to-evidence reasoning over heterogeneous pathology tools via reliability-guided routing. The framework operates through three key modules: (1) a Diagnostic Atlas that maps clinical queries to pathology concepts and corresponding feasible tool candidates; (2) a Concept–Evidence Tree that dynamically refines tool-to-region assignments as new evidence accrue; and (3) Invariance-Calibrated Reliability (ICR), which guides tool routing by measuring tool invariance under controlled input degradation while penalizing correlated errors. Validated measurements propagate into concept- and task-level estimates, yielding a fully traceable reasoning path. Compared with existing foundation models and agentic approaches, XPATHAGENT attains top performance on all 13 benchmarks across subtyping, grading, metastasis detection, stromal TIL scoring, and survival prediction, establishing concept-driven orchestration as a robust paradigm for trustworthy pathology agents.

1 INTRODUCTION

Computational pathology has developed increasingly capable specialist models for representation learning, segmentation, cell analysis, slide-level prediction, and vision-language understanding Lu et al. (2024); Shaikovski et al. (2024); Xiang et al. (2025); Vorontsov et al. (2023); Hörst et al. (2026); Ding et al. (2025); Qu et al. (2026); Lyu et al. (2025). However, pathology tasks require different types of evidence and ways of gathering them. For instance, cancer grading relies on detailed tissue morphology, including cellular appearance and signs of proliferation, whereas tumor-infiltrating lymphocyte scoring requires spatial reasoning over multiple tissue components. Such evidence can differ in visual scale, spatial organization, and the type of analysis needed to extract it, making it difficult for a single approach to support the full range of pathology tasks. More recently, agentic systems have emerged as a flexible alternative for reasoning over whole-slide images using specialized models and iterative evidence gathering Lyu et al. (2025); Sun et al. (2026); Chen et al. (2025a); Huang et al. (2026).

Recent pathology agents explore several strategies for structuring WSI reasoning. PathFinder and CPathAgent perform iterative region navigation and multi-scale evidence collection, while PathAgent uses self-reflection to determine whether additional evidence is needed (Ghezloo et al., 2025; Sun et al., 2026; Chen et al., 2025a). WSI-Agents coordinates specialized expert models with task allocation, external pathology knowledge, and model-based verification, whereas HistoSelect performs question-conditioned selection of relevant slide regions (Lyu et al., 2025; Huang et al., 2026). Despite these advances, two challenges remain central to our setting: making the required morphological evidence explicit and determining how much individual tool-derived measurements should be trusted.

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.

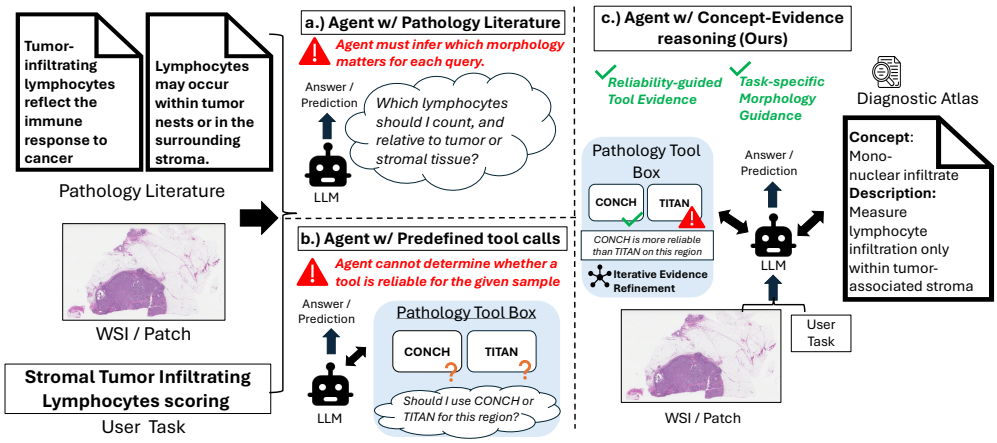

Figure 1: Illustration of current pathology agent paradigms and XPathAgent. (a) Literature-augmented agents provide pathology knowledge but still require the LLM to infer which morphological evidence is relevant to the query. (b) Tool-augmented agents can invoke specialist pathology models but lack an explicit mechanism to assess whether their outputs are reliable for the current sample. (c) XPathAgent combines task-specific morphology guidance from the Diagnostic Atlas with reliability-guided tool evidence for concept-to-evidence reasoning

First, pathology questions often require multifaceted evidence grounded in established pathology knowledge. For example, survival prognosis may require identifying and quantifying multiple tissue components and reasoning over their spatial and functional relationships. Existing agents largely delegate this decomposition to an orchestrating LLM or VLM, guided by self-reflection Chen et al. (2025a), external knowledge retrieval Lyu et al. (2025), or predefined tools and analysis steps Ghezloo et al. (2025); Sun et al. (2026). The required morphological evidence and its relationships therefore remain largely implicit, making it difficult to trace how individual measurements support the final decision.

Second, access to multiple pathology models does not ensure reliable evidence gathering. Existing agents either combine specialized models through predefined workflows or delegate tool selection to an orchestrating LLM (Lyu et al., 2025; Chen et al., 2025a; Ghezloo et al., 2025), but do not explicitly estimate whether a model output is reliable for the specific tissue, image quality, resolution, or region being analyzed. They therefore lack an input-conditional reliability estimate for down-weighting or pruning unreliable tool-region measurements. Moreover, when multiple models measure the same signal, agreement does not necessarily provide independent corroboration if the models exhibit correlated failures (Kuncheva & Whitaker, 2003).

We therefore propose XPATHAGENT, an agent framework that grounds pathology tasks in established morphological concepts and dynamically routes evidence through specialist tools based on measurement reliability. The Diagnostic Atlas maps each task to supporting concepts and candidate models, from which XPATHAGENT constructs a Concept-Evidence Plan Tree to guide evidence collection across tissue regions. As measurements are collected, Invariance-Calibrated Reliability (ICR) estimates concept-specific tool reliability and revises routes by pruning unreliable measurements. The remaining evidence is integrated across tools and concepts according to tool-concept suitability and complementary strengths. Across five pathology task categories and 13 benchmark–task combinations, XPATHAGENT achieves the strongest performance on all benchmarks. In the zero-shot setting, it improves over prior state-of-the-art methods by an average of 1.33 and 4.5 QWK points on cancer grading and sTIL scoring, respectively, and by relative gains of 3.1 and 1.6 on tumor subtyping and metastasis detection. For survival prediction, XPATHAGENT achieves a mean C-index of 0.726, compared with 0.704 for the second-best method. Our main contributions are summarized as follows:

- We propose a concept-grounded evidence reasoning framework that decomposes pathology tasks into morphological concepts and candidate specialist tools, enabling systematic and transparent evidence collection for interpretable pathology decision-making.

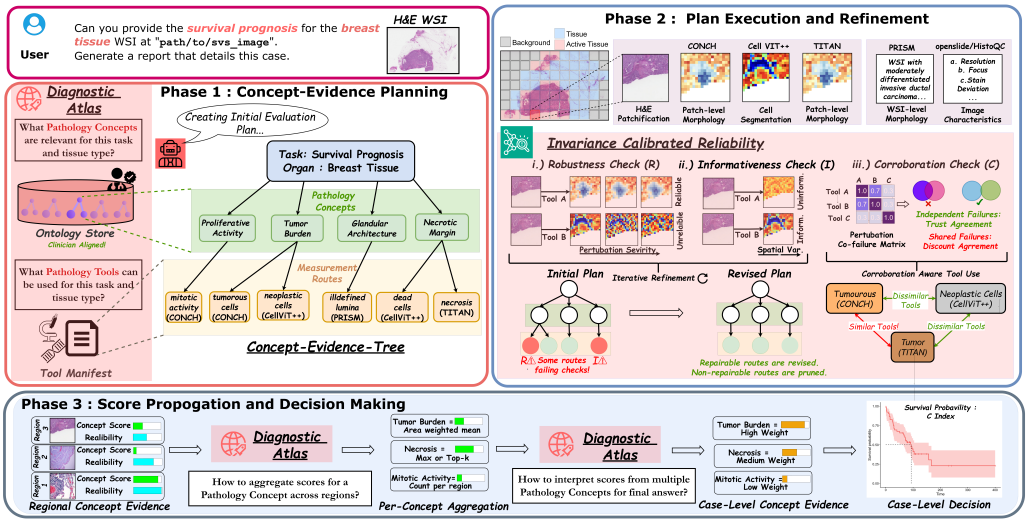

Figure 2: Overview of `XPathAgent`. Phase 1 grounds the query in task-relevant pathology concepts and candidate measurement routes using the Diagnostic Atlas. Phase 2 executes and refines these routes with invariance-calibrated reliability and cross-tool corroboration. Phase 3 aggregates regional concept evidence into slide-level concept scores and the final case-level decision.

- To leverage the varying strengths of diverse pathology foundation models across subtasks and inference settings, we design a reliability-guided adaptive routing mechanism that uses invariance-calibrated reliability to dynamically assess tool–concept suitability, prune unreliable routes, and integrate complementary evidence for robust task-level inference.
- We demonstrate broad generalization across 13 benchmark-task combinations spanning five pathology tasks, achieving the strongest performance across all benchmarks under both zero-shot and in-context learning settings.

2 METHOD

On a H&E image, `XPathAgent` responds to a task by decomposing it into evidentiary pathology concepts, measuring each with pathology model tools, estimating the reliability of every measurement, and aggregating the resulting evidence. The agent proceeds in three phases (Figure 2). In the *Concept-Evidence planning* phase, it converts the query into a plan specifying the task-relevant concepts and the tool routes available to measure them (§2.1). In the *Plan Execution and Refinement* phase, it executes this plan across image regions, assesses the reliability of each measurement, and revises unreliable tool routes (§2.2). In the *Score Propogation and Decision Making* phase, it aggregates the reliable measurements into concept-level and task-level estimates to produce the final answer (§2.3). These phases are supported by three expert-guided mechanisms for transparent reasoning: a *Diagnostic Atlas* that encodes pathology knowledge and the tool ecosystem; a *Self-Evolving Concept-Evidence Tree* (CET) that tracks tool calls and measurement-guided reasoning and *Invariance-Calibrated Reliability* (ICR), a label-free scheme for scoring tool measurements and guiding tool routing.

Preliminaries. Given a free-text query q and an H&E image I, which is either a whole-slide image or a region patch, `XPathAgent` returns a report-style answer $\mathcal{A}$ along with citing a set of regions of interest (RoI) $\mathcal{O} = \{o_1, \ldots, o_m\}$ within the input image as annotated evidence. Specifically, the agent first parses q into a known task type $t \in \{\text{grading, subtyping, scoring, prognosis, metastasis}\}$, an organ $org \in \{\text{breast, prostate, kidney, ovary, lymph}\}$. After background removal, the image is partitioned into a grid of fixed-size non-overlapping regions $\mathcal{R} = \{r_1, \ldots, r_N\}$, and each region is further tiled at the resolution required by each tool (Appendix A). We refer to this spatial hierarchy as WSI $\rightarrow$ region $\rightarrow$ tile throughout.

2.1 CONCEPT-EVIDENCE PLANNING PHASE

The first phase constructs a plan specifying the evidence required to answer the query and the tools available to measure it. Rather than relying on an LLM to infer the decomposition and tool selection from scratch, `XPathAgent` grounds both in the Diagnostic Atlas, which encodes task-relevant pathology concepts together with the capabilities and operating constraints of available tools. The resulting task-to-concept and concept-to-tool mappings form the initial Concept–Evidence Plan Tree, which is refined as evidence is gathered in the next phase.

Diagnostic Atlas. To align task reasoning with established pathology literature and enable expert-level transparency, we distill recurring tissue-morphology patterns and task-specific interpretation rules from pathology literature and clinical guidelines. We pair this knowledge with a diverse ecosystem of state-of-the-art pathology models spanning organs, resolutions, and diagnostic capabilities. The Diagnostic Atlas represents these two forms of knowledge separately through an *Ontology Store* and a *Tool Manifest*.

The *Ontology Store* defines a shared vocabulary of clinician-aligned pathology concepts and the morphological cues associated with each concept. These concepts are distilled from established pathology literature and reporting conventions (Table 8) , and clinician-reviewed to ensure that their definitions and task-specific use reflect real-world pathological interpretation. For a given task and organ, the Ontology Store retrieves the task-relevant concepts and defines the interpretation scheme, including functional roles, cross-region evidence aggregation, and contributions to the final decision. For example, in sTIL scoring, mononuclear infiltrate provides the quantity of interest, stromal extent defines its denominator, and tumor epithelium identifies tissue that should be excluded. These task-specific attributes are used later during decision propagation. The complete concept inventory and clinician-validation procedure are provided in Appendix B.1.

The *Tool Manifest* specifies how each pathology model measures these concepts. For each tool, it records its concept coverage, organ specificity, utility, and evaluation settings. For instance, it specifies each tool's *mechanism family* (e.g., image–text models such as CONCH versus cell-analysis models such as CellViT++)Escobar Díaz Guerrero et al. (2022), operating constraints (e.g. CONCH requires at least 20x magnification), and known failure modes (e.g. CellViT++ is sensitive to blurring). It also defines how each tool's native output is mapped into concept measurement and which configurations can be adjusted during route refinement. More details and examples on Tool Manifest are provided in Appendix B.2. Together, the Ontology Store and Tool Manifest allow `XPathAgent` to map each task-relevant concept to one or more compatible measurement routes before observing the image.

Concept–Evidence Tree. The *Concept–Evidence Tree* (CET) represents the evidence collection plan for a given task. The root represents the task–organ pair (t, o), the second level contains the relevant pathology concepts $\mathcal{C}_{t,o}$ retrieved from the Ontology Store, and its leaves contain the candidate tools available to measure each concept. The task and concept levels remain fixed throughout a run, while the tool routes are instantiated and refined in the next phase (§2.2).

Let $\mathcal{K}$ denote the available tools. For each concept $C \in \mathcal{C}_{t,o}$, the CET defines a set of candidate measurement routes

$$\mathcal{L}_C = \{(k, C) \ : \ k \in \mathcal{K},\ k \text{ can measure } C\}\,, \tag{1}$$

where tool compatibility and operating requirements are obtained from the Tool Manifest. All compatible tools are included in the initial plan to provide alternative ways of measuring the same concept. The resulting CET is therefore defined by the task–organ pair, its associated concepts, and the candidate routes for each concept:

$$\mathcal{T}_{t,o} = \Big((t, o),\, \mathcal{C}_{t,o},\, \{\mathcal{L}_C\}_{C \in \mathcal{C}_{t,o}}\Big). \tag{2}$$

2.2 PLAN EXECUTION AND REFINEMENT PHASE

After initializing the Concept–Evidence Tree, XPATHAGENT executes its candidate measurement routes within each tissue region and refines them based on the observed image. Routes that produce unsuitable or uninformative outputs are revised or removed, while the remaining routes are retained

as evidence for the corresponding pathology concept. This phase therefore converts the initial plan into a region-specific set of validated measurements that is passed to the final decision stage.

2.2.1 INVARIANCE-CALIBRATED RELIABILITY

Recent reliability-aware pathology agents estimate tool reliability from accumulated experience on previously observed cases Zhang et al. (2026). While effective when sufficient comparable experience is available, such approaches depend on prior examples that can inform the reliability of a tool on a new input. We instead estimate tool reliability offline and in a label-free manner. Specifically, we use a small reference set of unlabeled images sampled from the same benchmark as the given case. Before inference, we extract tissue-containing regions from this reference set and apply controlled perturbations along three image-quality axes-resolution, focus, and stain deviation-at multiple severity levels. Because these perturbations alter image quality while preserving the underlying tissue morphology, they provide a label-free probe of how reliably each tool responds as its operating conditions deteriorate. The complete definitions of perturbation, severity levels and construction of the reference-set are provided in Appendix C.

Let $s_k(r)$ denote the grid of tile-level measurements returned by tool k on region r. For perturbation axis j and severity level q, let r_j^q denote the corresponding perturbed region, and let ρ denote Spearman rank correlation across tiles.

Before Inference: We first calibrate each tool offline using reference regions $x \in \mathcal{X}_{\text{ref}}$. For each axis j and severity q, we compare the tool's measurements on every reference region before and after perturbation. Averaging these correlations across the N reference regions gives

$$a_{kj}(q) = \frac{1}{N} \sum_{x \in \mathcal{X}_{\text{ref}}} \rho\big(s_k(x), s_k(x_j^q)\big)\,. \tag{3}$$

This is repeated across severity levels q that gives us an invariance curve $a_{kj}(q)$ for each tool k and perturbation axis j. The curve maps each severity level to the average rank correlation across the reference regions. We store these curves offline and use them to estimate tool reliability under different image conditions. Higher values indicate that the tool better preserves the relative ordering of tissue locations as the perturbation becomes more severe.

During Inference: Let q_{jr} denote the observed severity of axis j in region r. We look up the corresponding value on the stored invariance curve and define

$$\textbf{i.) Admissibility:} \qquad a_k(r) = \min_j a_{kj}(q_{jr}), \tag{4}$$

so that a route is only as admissible as its least stable observed perturbation axis.

Because a spatially constant output can be highly invariant yet provide little discriminative evidence, we additionally measure

$$\textbf{ii.) Informativeness:} \qquad d_k(r) = 1 - \exp\left(-2\,\frac{\sigma_k^2(r)}{\sigma_k^2(S)}\right), \tag{5}$$

where $\sigma_k^2(r)$ is the within-region variance and $\sigma_k^2(S)$ is the variance of the same tool across the slide.

The resulting route reliability is

$$w_k(r) = a_k(r)\, d_k(r), \tag{6}$$

requiring a route to be both stable under the observed perturbation severities and informative within the region.

Iterative route refinement. For each measurement route we check if $a_k(r) < a_k^{min}$ and $d_k(r) < d_k^{min}$. Routes that fail *admissibility* are pruned directly, since changing the tool configuration cannot correct a mismatch between the tool and the observed image conditions. In contrast, low *informativeness* may arise from a suboptimal configuration. We therefore re-parameterize and re-evaluate such routes for up to $R_{\max}$ rounds, pruning those that remain uninformative. For example, in CONCH, reparameterization replaces the text embedding with an alternative phrase describing the same pathology concept, to make the cosine-similarity grid sharper.

Individual route reliability alone does not determine the confidence gained from agreement among multiple tools. Two tools may produce similar output because they independently support the same morphology, or because they fail under the same image degradation axis. We therefore reuse the per-region rank-stability measurements underlying the invariance curves in Eq. 3 to estimate dependence between the failure patterns of different tools.

For each tool k, let

$$\delta_k(x, j, q) = 1 - \rho\big(s_k(x),\, s_k(x_j^q)\big) \tag{7}$$

denote its degradation under perturbation (j, q). For a pair of tools A and B, we define their co-failure dependence as

$$\lambda_{AB} = \max(0,\, \rho(\boldsymbol{\delta}_A, \boldsymbol{\delta}_B))\,, \tag{8}$$

where $\boldsymbol{\delta}_k$ collects these degradation values across the reference regions and perturbation sweep. A high λ_{AB} therefore indicates that the two tools tend to lose stability under the same image conditions.

At inference time, let $g^{AB}(r)$ denote the observed agreement between two surviving routes measuring the same concept on region r. We discount this agreement by their estimated dependence:

$$\textbf{iii.) Corroboration:} \qquad \widetilde{g}^{AB}(r) = g^{AB}(r)\,(1 - \lambda_{AB})\,. \tag{9}$$

Thus, agreement provides stronger corroboration when the tools have independent failure patterns, and is discounted when they are known to fail together. We compute $\widetilde{g}^{AB}(r)$ for every pair of surviving tools measuring the same concept in a region. These pairwise corroboration values are retained alongside the corresponding tool measurements and reliabilities, then verbalized and passed to the orchestrating agent for concept-level interpretation in the next phase.

2.3 Score Propagation and Decision Making Phase

The final phase converts the regional concept evidence produced in Phase 2 into a case-level decision using the aggregation rules specified in the Diagnostic Atlas. Each regional estimate carries both a concept score and its associated precision, which are propagated through two stages.

Concept-level Aggregation. Regional evidence is aggregated according to the interpretation rule associated with each pathology concept. The rule ϱ_C specifies how regional measurements are combined: `representative` computes a precision-weighted aggregate across regions, whereas `max` and `presence` select the most extreme region that exceeds a minimum precision threshold. This allows different concepts within the same task to be interpreted with distinct aggregation strategies. For example, Nottingham grading assesses glandular architecture across the tumor, while nuclear atypia is determined by the most abnormal region. Each concept therefore yields a slide-level estimate $\hat{z}_C^{\mathrm{WSI}}$ together with its propagated precision.

Task-level Interpretation. Concepts may contribute to the final decision with different directions and magnitudes. The Diagnostic Atlas specifies each concept's direction of association π_C with the target outcome and its importance weight $\beta_C \geq 0$. We use π_C to align concept scores so that larger values consistently indicate stronger evidence for the target outcome, yielding the aligned score $\tilde{z}_C^{\mathrm{WSI}}$. The case-level score is then

$$\eta = \frac{1}{B} \sum_C \beta_C\, \tilde{z}_C^{\mathrm{WSI}}, \qquad B = \sum_C \beta_C, \tag{10}$$

where β_C controls the relative contribution of each concept. The associated uncertainty is obtained by propagating the concept precisions and the reasoning LLM then converts the aggregated concept evidence into the requested output format, while preserving the concept-to-region evidence chain supporting the final decision.

3 Experiments

We evaluate XPathAgent across five computational pathology task categories: cancer grading, metastasis detection, tumor subtyping, and stromal tumor-infiltrating lymphocyte (sTIL) scoring,

and survival prognosis. Our experiments address three questions: (1) Generalizability: can the agentic framework support heterogeneous pathology tasks across organs and datasets? (2) Reasoning does invariance-calibrated reliability enable more effective evidence integration than direct aggregation of specialist tools? (3) Adaptability: can the framework adapt to new examples through in-context learning?

3.1 Experimental Setup

Datasets and metrics. We evaluate XPathAgent on 13 benchmark-task combinations. Cancer grading is evaluated on SICAP Silva-Rodríguez (2020), KMC-RCC Chanchal et al. (2023), and TCGA-UCEC Erickson et al. (2016), while metastasis detection is evaluated on CAMELYON16 Ehteshami Bejnordi et al. (2017) and CAMELYON17 Bándi et al. (2019). For tumor subtyping, we use TCGA-RCC Ricketts et al. (2018), and TCGA-NSCLC The Cancer Genome Atlas Research Network (2014; 2012), and evaluate stromal tumor-infiltrating lymphocyte (sTIL) scoring on TCGA-BRCA The Cancer Genome Atlas Network (2012); National Cancer Institute and TIGER-WSITILS TIGER Challenge. We report quadratic weighted kappa (QWK) for ordinal grading and sTIL scoring, balanced accuracy (BACC) for metastasis detection and subtyping, and Weighted-F1 as an additional measure. Survival prognosis is assessed on TCGA-BLCA Robertson et al. (2017), TCGA-BRCA The Cancer Genome Atlas Network (2012); National Cancer Institute, TCGA-LUAD The Cancer Genome Atlas Research Network (2014), and TCGA-UCEC Erickson et al. (2016) with concordance index (C-Index) Harrell Jr et al. (1996). Dataset details, evaluation protocols, and task-specific dataset and tool adaptations are provided in the Appendix D.

Table 1: **Survival prediction** across four cancer cohorts, evaluated using C-index. For pathology foundation models, the backbones are kept frozen, and only the ABMIL aggregator is trained.

Method	BLCA ↑	BRCA ↑	LUAD ↑	UCEC ↑	Overall ↑
MIL Baseline					
ResNet-50 + ABMIL Ilse et al. (2018)	0.528	0.596	0.587	0.632	0.586
Foundation Models					
PLIP Huang et al. (2023) + ABMIL	0.541	0.611	0.626	0.661	0.610
VIRCHOW Vorontsov et al. (2023) + ABMIL	0.557	0.645	0.627	0.637	0.617
CONCH Lu et al. (2024) + ABMIL	0.604	0.658	0.653	0.704	0.654
MUSK Xiang et al. (2025) + ABMIL	0.597	0.667	0.660	0.699	0.656
PRISM Shaikovski et al. (2024)	0.624	0.663	0.648	0.711	0.662
Survival-Specific					
SurvPath Jaume et al. (2024)	0.631	0.660	0.672	0.703	0.667
HFGPI Zhou et al. (2026)	0.665	<u>0.694</u>	0.707	<u>0.726</u>	0.698
MDCS-MoAME Qu et al. (2026)	<u>0.679</u>	0.692	<u>0.724</u>	0.721	<u>0.704</u>
XPathAgent-Base	0.662	0.684	0.700	0.717	0.691
XPathAgent	**0.713**	**0.721**	**0.729**	**0.741**	**0.726**

Baselines and implementation. We compare against pathology foundation models, task-specific approaches, and recent pathology reasoning agents, including PLIP Huang et al. (2023), CONCH Lu et al. (2024), MUSK Xiang et al. (2025), VIRCHOW Vorontsov et al. (2023), PRISM Shaikovski et al. (2024), PathAgent Chen et al. (2025a), WSI-Agent Lyu et al. (2025), and HistoSelect Huang et al. (2026). For survival analysis, we additionally compare with ABMIL Ilse et al. (2018) and specialized survival models including SurvPath Jaume et al. (2024), HFGPI Zhou et al. (2026), and MDCS-MoAME Qu et al. (2026). XPathAgent-Base retains the agent scaffolding but removes the Diagnostic Atlas, the Concept-Evidence Tree, and Invariance-Calibrated Reliability components. All modules in XPathAgent operate without requiring in-domain labeled data or task-specific tuning. For prognosis tasks, the resulting evidence is learnt using ABMIL under weak supervision from overall survival labels; all other tasks are evaluated in a training-free setting. ICR calibration constructs the reference set $\mathcal{X}_{\text{ref}}$ using 20 random, unlabelled WSIs/patches per dataset from samples excluded from the evaluation cohort. The invariance curves $a_{kj}(q)$ are fitted for each tool k and quality axis j at four severity levels q (plus a clean reference), using Gaussian defocus and HED stain jitter on a scanner-independent scale. For routes with just one measurement, we use corroboration baseline $c_0 = 0.5$. In all experiments, each region $r \in \mathcal{R}$ is 8192 px, further subdivided into 512-px square tiles. The reasoning agent uses Qwen3.5 27B Qwen Team (2026a) as its underlying LLM. All reported results are averaged over five runs for each dataset. Additional implementation details are in Appendix D.

3.2 Results and Ablations

Under the zero-shot setting, XPathAgent achieves the best performance on all benchmark-task combinations spanning cancer grading, sTIL grading, tumor subtyping, and metastasis detection (Table 2 3). XPathAgent achieves the highest QWK scores for cancer grading, reaching 0.789 on SICAP and

Table 2: Zero-shot performance on **cancer grading** and **stromal tumor-infiltrating lymphocyte (sTIL) scoring** benchmarks. We report quadratic weighted kappa (QWK) together with Weighted-F1. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	SICAP		KMC-RCC		TCGA-UCEC		TCGA-BRCA		TIGER-WSITILS	
	QWK ↑	W-F1 ↑	QWK ↑	W-F1 ↑	QWK ↑	W-F1 ↑	QWK ↑	W-F1 ↑	QWK ↑	W-F1 ↑
PLIP Huang et al. (2023)	0.319	0.255	0.306	0.251	0.217	0.124	0.225	0.175	0.274	0.206
CONCH Lu et al. (2024)	0.694	0.417	0.624	0.550	0.579	0.524	0.536	0.452	0.553	0.491
MUSK Xiang et al. (2025)	0.711	0.440	0.615	0.548	0.545	0.437	0.560	0.498	0.582	0.535
PRISM Shaikovski et al. (2024)	–	–	–	–	–	–	0.442	0.414	0.472	0.467
PathAgent Chen et al. (2025a)	0.714	0.537	0.699	0.606	0.597	0.528	0.445	0.394	0.487	0.439
WSI-Agents Lyu et al. (2025)	0.527	0.398	0.476	0.334	0.479	0.403	0.500	0.436	0.519	0.522
HistoSelect Huang et al. (2026)	0.769	0.600	0.649	0.580	**0.641**	0.568	0.572	0.521	0.589	0.538
XPathAgent-Base	0.753	0.572	0.686	0.615	0.621	0.565	0.599	0.516	0.614	0.558
XPathAgent	**0.789**	**0.655**	**0.722**	**0.623**	0.638	**0.574**	**0.620**	**0.528**	**0.631**	**0.565**

Table 3: Zero-shot performance on **tumor subtyping** and **metastasis detection** benchmarks. We report balanced accuracy (BACC) together with Weighted-F1. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	TCGA-RCC		TCGA-NSCLC		CAMELYON16		CAMELYON17	
	BACC ↑	W-F1 ↑	BACC ↑	W-F1 ↑	BACC ↑	W-F1 ↑	BACC ↑	W-F1 ↑
PLIP Huang et al. (2023)	0.761	0.746	0.717	0.709	0.399	0.416	0.465	0.522
CONCH Lu et al. (2024)	0.873	0.861	0.873	0.855	0.671	0.653	0.694	0.670
MUSK Xiang et al. (2025)	0.822	0.812	0.844	0.817	0.604	0.599	0.625	0.608
PRISM Shaikovski et al. (2024)	0.615	0.774	0.876	0.838	0.800	0.789	0.811	0.814
PathAgent Chen et al. (2025a)	0.799	0.742	0.764	0.736	0.570	0.526	0.629	0.583
WSI-Agents Lyu et al. (2025)	0.898	0.833	0.880	0.862	0.522	0.561	0.577	0.595
HistoSelect Huang et al. (2026)	0.875	0.819	0.916	0.865	0.613	0.628	0.630	0.657
XPathAgent-Base	0.921	0.907	0.910	0.884	0.784	0.791	0.807	0.801
XPathAgent	**0.939**	**0.914**	**0.931**	**0.903**	**0.822**	**0.808**	**0.815**	**0.837**

0.722 on KMC-RCC. For sTIL scoring, XPATHAGENT achieves the highest QWK on both TCGA-BRCA (0.620) and TIGER-WSITILS (0.631), representing relative gains of 8.4% and 7.1% over HistoSelect. For metastasis detection, XPATHAGENT achieves the highest BACC on both CAMELYON16 and CAMELYON17 (0.822 and 0.815). For tumor subtyping, XPATHAGENT achieves the highest BACC on TCGA-RCC and TCGA-NSCLC (0.939 and 0.931), outperforming the second-best WSI-Agent and HistoSelect by relative gains of 4.6% and 1.6%, respectively.

For survival prognosis, XPATHAGENT reaches a mean C-index of 0.726, improving over the strongest specialized survival model, MDCS-MoAME (0.704), by by a 3.1% relative gain (Table 1). In table 4 Five-shot inference consistently improves performance, including QWK gains from 0.789 to 0.890

Table 4: Effect of in-context examples on XPATHAGENT across pathology benchmarks. We report QWK for grading and sTIL scoring, and BACC for metastasis detection and subtyping. Best results are in **bold**.

Dataset	Metric	0-shot	1-shot	5-shot
Grading				
SICAP (Gleason)	QWK	0.789	0.856	**0.890**
KMC-RCC Grade	QWK	0.722	0.787	**0.833**
TCGA-UCEC	QWK	0.638	0.703	**0.742**
Metastasis Detection				
CAMELYON16	BACC	0.822	0.839	**0.888**
CAMELYON17	BACC	0.815	0.852	**0.898**
Subtyping				
TCGA-RCC	BACC	0.939	0.954	**0.983**
TCGA-NSCLC	BACC	0.931	0.957	**0.991**
sTIL Scoring				
TCGA-BRCA	QWK	0.620	0.664	**0.748**
TIGER-WSITILS	QWK	0.631	0.659	**0.721**

on SICAP and BACC from 0.931 to 0.991 on TCGA-NSCLC, showing that XPATHAGENT supports adaptation without visual fine-tuning.

We further conduct ablation studies to evaluate the contribution of each proposed module, including the pre-curated Diagnostic Atlas, concept–evidence reasoning, and multi-tool routing. Progressively adding the Diagnostic Atlas, Concept-Evidence Tree, and the Invariance Calibrated Reliability improves mean survival C-index from 0.691 to 0.704, 0.714, and 0.726, respectively (Table 6).

Performance remains stable across reasoning model scales, with Qwen3.5-27B Qwen Team (2026a) and Qwen3.5-35B Qwen Team (2026b) achieving comparable results (Table 6). Finally, leave-one-tool-out experiments show that removing any specialist degrades performance, with the largest drops observed when CONCH or PRISM is removed (Table 5). Results on more tasks in Appendix D.

4 Related Work

4.1 Computational Pathology

Table 5: Leave-one-tool-out analysis of the XPathAgent tool cache. We report the mean C-index across the four survival cohorts.

Tool Configuration	C-index ↑
Full Tool Cache	**0.726**
w/o CellViT++	0.675
w/o CONCH	0.664
w/o PRISM	0.651
w/o TITAN	0.697

Recent computational pathology models have progressed from task-specific architectures toward general-purpose foundation models. Vision encoders such as UNI Chen et al. (2023) and Virchow Vorontsov et al. (2023) learn transferable histopathology representations, while PLIP Huang et al. (2023), CONCH Lu et al. (2024), and MUSK Xiang et al. (2025) incorporate vision-language supervision; slide-level and multimodal models such as PRISM Shaikovski et al. (2024), TITAN Ding et al. (2025), and SlideChat Chen et al. (2025b) further extend this modeling to whole-slide analysis . For downstream WSI prediction, patch-level representations are commonly aggregated using multiple-instance learning, with methods such as ABMIL Ilse et al. (2018), CLAM Lu et al. (2021), TransMIL Shao et al. (2021), and DTFD-MIL Zhang et al. (2022) differing in how informative regions are identified and combined. Survival-specific approaches further adapt slide-level modeling by incorporating molecular information or long-range tissue structure, including MCAT Chen et al. (2021), SurvPath Jaume et al. (2024), HFGPI Zhou et al. (2026), and MDCS-MoAME Qu et al. (2026). XPathAgent constructs task-relevant, reliability-aware evidence reusable across multiple pathology tasks.

4.2 Agentic Reasoning

Table 6: Ablation of XPathAgent components and reasoning model scale for survival prediction. Performance is measured using C-index; higher is better.

Method	BLCA ↑	BRCA ↑	LUAD ↑	UCEC ↑	Overall ↑
Component Ablation					
XPathAgent-Base	0.662	0.684	0.700	0.717	0.691
XPathAgent-Base + DA	0.683	0.695	0.707	0.729	0.704
XPathAgent-Base + DA + CET	0.698	0.709	0.716	0.734	0.714
XPathAgent-Base + DA + CET + ICR	**0.713**	**0.721**	**0.729**	**0.741**	**0.726**
Reasoning Model Scale					
Qwen 3.5 4B Qwen Team (2026c)	0.702	0.707	0.711	0.719	0.710
Qwen 3.5 27B Qwen Team (2026a)	0.713	0.721	0.729	**0.741**	0.726
Qwen 3.5 35B Qwen Team (2026b)	**0.716**	**0.722**	**0.732**	0.739	**0.727**

LLM-based agents extend language models with iterative reasoning and external tool use, as exemplified by ReAct, Toolformer, and multi-agent frameworks (Yao et al., 2022; Schick et al., 2023; Wu et al., 2023). In computational pathology, PathFinder Ghezloo et al. (2025) introduced iterative agent-based region selection, while CPathAgent Sun et al. (2026) and PathAgent Chen et al. (2025a) perform multi-scale evidence collection and navigation; WSI-Agents Lyu et al. (2025) coordinates specialized pathology models, and HistoSelect Huang et al. (2026) performs question-conditioned selection of WSI evidence. Much of this literature focuses on language-mediated WSI analysis, including VQA, diagnosis, and report generation. XPathAgent instead targets heterogeneous computational pathology endpoints and explicitly estimates input-dependent reliability of tool-derived evidence while accounting for correlated failures between measurement routes.

5 Conclusion

In this work, we present XPathAgent, a plug-and-play framework for concept-to-evidence reasoning over heterogeneous pathology tools. By grounding clinical queries in established pathology concepts, dynamically refining tool-to-region assignments, and calibrating measurement reliability through invariance and correlated-error analysis, XPathAgent enables reliable and traceable evidence integration without modifying the underlying pathology models. Extensive evaluation across 13 benchmark–task combinations spanning five pathology tasks demonstrates broad generalization and consistently strong performance under both zero-shot and in-context learning settings. These results highlight the potential of concept-driven orchestration to integrate diverse pathology models into reliable, interpretable, and extensible agents for computational pathology.

AI Use Disclosure

We used Claude (Anthropic) as a writing assistant while preparing this paper. This mainly meant rephrasing sentences for clarity, tightening wording, and catching grammatical errors and incon-

sistencies in the draft. All technical content, including the method, experiments, and results, was written by the authors. Every AI-suggested edit was read and checked by the authors before inclusion.

REFERENCES

M. B. Amin, S. B. Edge, F. L. Greene, et al. *AJCC Cancer Staging Manual*. Springer, 8th edition, 2017.

Péter Bándi, Oscar Geessink, Quirine F. Manson, Marcory van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Amit Kumar Chanchal, Shyam Lal, Ranjeet Kumar, Jin Tae Kwak, and Jyoti Kini. A novel dataset and efficient deep learning framework for automated grading of renal cell carcinoma from kidney histopathology images. *Scientific Reports*, 13(1):5728, 2023.

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. *arXiv preprint arXiv:2511.17052*, 2025a.

Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3995–4005. IEEE, 2021.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474*, 2023.

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, 2025b.

College of American Pathologists. Cancer protocol templates. `https://www.cap.org/protocols-and-guidelines/cancer-reporting-tools/cancer-protocol-templates`, 2024. Standardized cancer reporting protocols; accessed 2026.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, 31(11):3749–3761, 2025.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, 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. doi: 10.1001/jama.2017.14585.

C. W. Elston and I. O. Ellis. Pathological prognostic factors in breast cancer. I. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 19(5):403–410, 1991. doi: 10.1111/j.1365-2559.1991.tb00229.x.

BJ Erickson, D Mutch, L Lippmann, and R Jarosz. The cancer genome atlas uterine corpus endometrial carcinoma collection (tcga-ucec). *The Cancer Imaging Archive*, 15, 2016.

Rodrigo Escobar Díaz Guerrero, Lina Carvalho, Thomas Bocklitz, Juergen Popp, and José Luis Oliveira. Software tools and platforms in digital pathology: a review for clinicians and computer scientists. *Journal of Pathology Informatics*, 13:100103, 2022. doi: 10.1016/j.jpi.2022.100103.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23431–23441, 2025.

Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Fabian Hörst, Moritz Rempe, Helmut Becker, Lukas Heine, Julius Keyl, and Jens Kleesiek. Cellvit++: Energy-efficient and adaptive cell segmentation and classification using foundation models. *Computer Methods and Programs in Biomedicine*, pp. 109206, 2026.

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.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. Pmlr, 2018.

Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11579–11590. IEEE, 2024.

Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003.

P. Laurén. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. an attempt at a histo-clinical classification. *Acta Pathologica et Microbiologica Scandinavica*, 64:31–49, 1965. doi: 10.1111/apm.1965.64.1.31.

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*, 30:863–874, 2024.

A. Lugli, R. Kirsch, Y. Ajioka, F. Bosman, et al. Recommendations for reporting tumor budding in colorectal cancer based on the international tumor budding consensus conference (ITBCC) 2016. *Modern Pathology*, 30(9):1299–1311, 2017. doi: 10.1038/modpathol.2017.46.

Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. *arXiv preprint arXiv:2507.14680*, 2025.

National Cancer Institute. TCGA-BRCA: Breast Invasive Carcinoma. Genomic Data Commons Data Portal. URL `https://portal.gdc.cancer.gov/projects/TCGA-BRCA`. Accessed: 2026-09-26.

Linjie Qu, Jin Xiao, Xiangrong Liu, Changming Sun, Hui Cui, Yuqi Fang, Ran Su, Qiangguo Jin, and Leyi Wei. Mdcs-moame: Multi-directional composite scanning with mixture of attention and mamba experts for cancer survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14461–14470, 2026.

Qwen Team. Qwen3.5-27b: Native vision-language dense foundation model. https://huggingface.co/Qwen/Qwen3.5-27B, 2026a.

Qwen Team. Qwen3.5-35b: Native vision-language mixture-of-experts foundation model. https://huggingface.co/Qwen/Qwen3.5-35B, 2026b.

Qwen Team. Qwen3.5-4b: Native vision-language dense foundation model. https://huggingface.co/Qwen/Qwen3.5-4B, 2026c.

Christopher J. Ricketts et al. The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Reports*, 23(1):313–326.e5, 2018. doi: 10.1016/j.celrep.2018.03.075.

A. Gordon Robertson, Jaegil Kim, Hikmat Al-Ahmadie, Joaquim Bellmunt, Guangwu Guo, Andrew D. Cherniack, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556.e25, 2017. doi: 10.1016/j.cell.2017.09.007.

R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs working group 2014. *Annals of Oncology*, 26(2):259–271, 2015. doi: 10.1093/annonc/mdu450.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

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.

Julio Silva-Rodríguez. Sicapv2-prostate whole slide images with gleason grades annotations. *Mendeley Data*, 1:2020, 2020.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Bowen Ding, Tao Lin, and Lin Yang. Cpathagent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. *Advances in Neural Information Processing Systems*, 38:101673–101731, 2026.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–525, 2012. doi: 10.1038/nature11404.

The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511:543–550, 2014. doi: 10.1038/nature13385.

TIGER Challenge. TIGER: Training Data. Grand Challenge. URL https://tiger.grand-challenge.org/Data/. Accessed: 2026-09-26.

G. W. van Pelt, S. Kjær-Frifeldt, J. H. J. M. van Krieken, R. Al Dieri, H. Morreau, R. A. E. M. Tollenaar, F. B. Sørensen, and W. E. Mesker. Scoring the tumor-stroma ratio in colon cancer: procedure and recommendations. *Virchows Archiv*, 473(4):405–412, 2018. doi: 10.1007/s00428-018-2408-z.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023.

WHO Classification of Tumours Editorial Board. *WHO Classification of Tumours*. IARC Press, Lyon, France, 5th edition, 2019.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. *URL https://arxiv. org/abs/2308.08155*, 2308, 2023.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Chengyang Zhang, Wenchuan Zhang, Bo Li, Mengran Li, Bob Zhang, Yuhao Yi, Hong Bu, and Jiancheng Lv. Pathosage: Towards multi-source evidence adjudication in pathology via experience-aware agentic workflow. *arXiv preprint arXiv:2606.07549*, 2026.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18780–18790. IEEE, 2022.

Junjie Zhou, Bao Xue, Meiling Wang, Wei Shao, and Daoqiang Zhang. Advancing cancer prognosis with hierarchical fusion of genomic, proteomic and pathology imaging data from a systems biology perspective. *arXiv preprint arXiv:2603.13787*, 2026.

A PREPROCESSING

A WSI in `.svs` format is partitioned into non-overlapping 8192 px regions at $40\times$ (16×16 tiles of 512 px). Background is removed with a saturation-based tissue mask; a region is discarded if it holds fewer than 8 tissue tiles. Each retained region is tiled at the resolution each tool requires (e.g. 224 px at $20\times$ for CONCH). A cropped input is treated as a single region. Tool outputs are computed once per image and cached. For some benchmarks the underlying samples are whole-slide images, but the public release provides only extracted patches. For example for SICAPv2, the source material is 155 prostate biopsy WSIs from 95 patients. These are distributed as 512 × 512 px patches, downsampled from 40× to 10×, with a single Gleason score per biopsy. Moreover, inputs that are single images rather than slides (e.g. the KMC-RCC microscope images) are treated as a single region.

B MORE DETAILS ON DIAGNOSTIC ATLAS

B.1 ONTOLOGY STORE AND CONCEPT INVENTORY

The Ontology Store provides the pathology knowledge used by `XPathAgent` to decompose a task into reusable morphological concepts. These concepts are distilled from established histologic grading systems, reporting standards, and pathology literature; we do not introduce new morphological entities. Each concept is defined once and may be reused across different tasks and organs.

Concept. The `concept` field gives the name of the morphological pattern, such as *nuclear atypia*, *glandular architecture*, or *necrosis*. The same concept name and definition are reused wherever that morphology is relevant.

Definition. The `definition` field describes the tissue pattern represented by the concept in plain pathological terms. It is task-independent: for example, *nuclear atypia* has the same morphological meaning whether it is used for grading or prognosis.

Raised and lowered by. The `raised_by` and `lowered_by` fields describe the visual findings that should increase or decrease the measured concept, respectively. They provide the agent with concrete morphological cues for interpreting tool outputs. A concept need not have both types of cues when only one direction is meaningful.

Task–organ mapping. Each concept is linked to the task–organ pairs for which it is relevant. This determines which concepts are retrieved when the agent constructs the Concept–Evidence Tree for a query. A concept may therefore appear in several tasks without being redefined.

Polarity. The `polarity` field specifies how increasing evidence for a concept affects the task-level decision. A `positive` polarity means that a larger concept value increases the task score, whereas a `negative` polarity means that it decreases it. Polarity is task-specific even when the underlying concept definition is shared.

Aggregation. The `aggregation` field specifies how the regional measurements selected by the reading rule are combined into a slide-level concept value. For example, `precision-weighted` gives greater influence to regional measurements supported by more reliable evidence, rather than treating every region equally. Other aggregation rules follow the corresponding regional read, such as retaining the maximum reliable value or recording whether reliable evidence is present.

Role. The `role` field specifies how a concept is used in the task-level decision. A `driver` is a concept whose measured value directly contributes to the final task score. This field separates concepts that affect the decision itself from concepts that may instead provide supporting or contextual information.

Weight. The `weight` field specifies the relative importance of a driver concept. `core` concepts receive the primary task-level contribution, while `secondary` concepts provide additional supporting evidence with lower relative importance. These labels are converted to the task-level weights used during score propagation.

Concept	Polarity	Aggregation	Role	Weight
glandular architecture	−	precision-weighted	driver	core
nuclear atypia	+	max over reliable regions	driver	core
proliferative activity	+	max over reliable regions	driver	core
mononuclear infiltrate	−	precision-weighted	driver	core
structure tracking invasion	+	presence over reliable regions	driver	core
lesion extent	+	max over reliable regions	driver	core
necrosis	+	precision-weighted	driver	secondary
stromal extent	+	precision-weighted	driver	secondary

Table 7: Task-specific Ontology Store entry for breast survival prognosis. Polarity indicates whether increasing evidence for a concept increases (+) or decreases (−) estimated risk. The regional-read field specifies how a concept should be interpreted across tissue regions, while aggregation specifies how the corresponding regional evidence is combined. Role determines how the concept participates in task-level interpretation, and weight denotes its relative contribution.

As a concrete example, Table 7 shows the complete task-specific Ontology Store entry used for breast survival prognosis.

Table 8 lists some of the pathology concepts used in this work together with the established standard or criterion from which each concept is derived and its primary reference.

This organization makes the concept vocabulary auditable while keeping concept definition separate from task-specific use. A concept is defined once, whereas its aggregation rule, functional role, and contribution to the task are specified separately for each task–organ pair.

B.2 Tool Manifest

The Tool Manifest describes the measurement tools available to `XPathAgent` and the conditions under which each tool can be trusted. For every tool, it records what type of pathology evidence the tool can measure, the spatial scale at which it operates, the image conditions it requires, the kinds of image degradation that may affect it, and the main situations in which its output can become unreliable. These entries are used when constructing candidate measurement routes and when checking whether a route remains suitable for a given image.

Table 9 shows a representative entry for CONCH Lu et al. (2024). The remaining tools are described using the same set of fields.

Concept	Standard / defining criterion	Primary reference
glandular architecture	Tubule and gland formation, the first component of the Nottingham combined histologic grade	Elston & Ellis, 1991 Elston & Ellis (1991)
nuclear atypia	Nuclear pleomorphism, the second component of the Nottingham grade; also the nuclear-grade element of CAP reporting	Elston & Ellis, 1991 Elston & Ellis (1991); CAP College of American Pathologists (2024)
proliferative activity	Mitotic count, the third component of the Nottingham grade, assessed over a defined field area	Elston & Ellis, 1991 Elston & Ellis (1991)
cellular dyscohesion	Loss of cell-to-cell adhesion distinguishing diffuse from intestinal-type carcinoma; single cells and small clusters infiltrating the stroma	Laurén, 1965 Laurén (1965)
necrosis	Confluent tumour necrosis, including the comedo pattern; a graded and reported element in multiple organ protocols	CAP College of American Pathologists (2024); WHO WHO Classification of Tumours Editorial Board (2019)
mononuclear infiltrate	Stromal tumour-infiltrating lymphocytes (TILs), scored as mononuclear cells within the invasive tumour stroma	Salgado et al., 2015 Salgado et al. (2015)
stromal extent	Tumour–stroma ratio: the proportion of supporting connective tissue relative to epithelium	van Pelt et al., 2018 van Pelt et al. (2018)
tumor epithelial extent	Neoplastic epithelial fraction of the region; the complementary term to the tumour–stroma ratio and the invasive-tumour reference compartment for TIL scoring	Salgado et al., 2015 Salgado et al. (2015); van Pelt et al., 2018 van Pelt et al. (2018)
invasive front configuration	Configuration of the tumour–host boundary, including tumour budding at the invasive front	Lugli et al., 2017 (ITBCC 2016) Lugli et al. (2017)
structure tracking invasion	Lymphovascular and perineural invasion: tumour within a vascular, lymphatic, or perineural space; a required CAP reporting element	CAP College of American Pathologists (2024); AJCC Amin et al. (2017)
secretory differentiation	Mucinous and signet-ring differentiation, as defined in WHO histologic typing	WHO, 2019 WHO Classification of Tumours Editorial Board (2019)
lesion extent	Tumour size and component extent (areal fraction and maximum contiguous dimension); a core CAP/AJCC staging quantity	AJCC Amin et al. (2017); CAP College of American Pathologists (2024)
host tissue context	Organ and tissue site, the setting against which every other finding is interpreted; the anchor of every organ-specific reporting protocol	CAP College of American Pathologists (2024); WHO WHO Classification of Tumours Editorial Board (2019)

Table 8: Provenance of the pathology concepts used in the Ontology Store. Each concept corresponds to a morphological pattern derived from an established grading system, reporting standard, or pathology reference. Concept definitions are shared across tasks and organs; task-specific fields determine how each concept is interpreted and used in the final decision.

C Invariance-Calibrated Reliability: Implementation Details

C.1 Offline Calibration

ICR is calibrated separately for each benchmark using only unlabeled images from that benchmark. We sample tissue-rich regions and evaluate each applicable tool under controlled perturbations along three axes: resolution, focus, and stain deviation, each at multiple severity levels including the clean image. The resulting per-region Spearman correlations are averaged to form the invariance curves used in the main text.

For every benchmark (eg TCGA-BRCA), we use 20 reference slides (not used for evaluation) with five tissue-rich regions per slide and up to 24 tiles per region. Focus perturbations use Gaussian blur

Field	CONCH
What the tool returns	For each image tile, CONCH compares the visual content against text descriptions of the two opposing states of a pathology concept and returns a relative similarity score. Repeating this over all tiles in a region produces a spatial map of the concept.
Spatial scale	CONCH operates on image tiles within a region, so its output preserves spatial information and can show where in the tissue a concept is expressed more strongly or weakly.
How to interpret the output	Positive values indicate that a tile more closely matches the description associated with increased presence of the concept, while negative values favor the opposing description. The score is a relative similarity measure rather than a probability.
What it is suited to measure	CONCH is primarily used for tissue- and architecture-level patterns, such as gland formation, tumor extent, stromal composition, inflammatory infiltrate, necrosis, and other morphology that can be recognized from local tissue appearance.
What it is not suited to measure	It is not used for precise cell counting, detailed cell typing, mitotic activity, or molecular markers. Very fine cellular features may also be unreliable when the image resolution is too low to resolve them clearly.
Resolution requirements	Architectural features require at least $10\times$ magnification, while finer cellular features require approximately $20\times$. Routes are not used when the available resolution falls below the requirement of the concept being measured.
Input requirements	The model is used on H&E-stained tissue. Inputs that do not satisfy this requirement are excluded before measurement.
Image conditions that affect reliability	Measurements can degrade when tissue is out of focus or when staining differs substantially from the appearance seen under normal H&E preparation. These factors are evaluated by ICR before the measurement is trusted.
Known limitations	CONCH can give misleading measurements for fine cellular concepts when used below the required resolution. In addition, strong separation between the two text descriptions in the model's embedding space does not by itself guarantee that the corresponding image measurement will be reliable.
Primary Role	CONCH provides spatial evidence for pathology concepts within each region. Its concept maps can therefore contribute directly to regional measurements and can be compared with measurements from other tools for corroboration.

Table 9: Representative Tool Manifest entry for CONCH. The manifest describes how the tool is used, what kinds of pathology evidence it can and cannot measure, the conditions required for valid use, and the main factors that can reduce measurement reliability.

with $\sigma \in \{1, 2, 4, 8\}$ pixels; stain perturbations use increasing H&E channel shifts. The resulting curves are stored per tool and perturbation axis and evaluated at inference using linear interpolation.

C.2 Iterative Route Refinement

Algorithm 1 summarizes route refinement during inference.

We use $R_{\max} = 3$. A route is evaluated only when at least three finite tile measurements are available. Re-parameterization is tool-specific and restricted to alternatives listed in the Tool Manifest.

C.3 Discounted Cross-Tool Agreement

The co-failure estimates reuse the same per-region stability measurements collected during offline calibration. For each tool pair, we correlate their degradation profiles across reference regions, perturbation axes, and severity levels to obtain λ_{AB}; negative correlations are clipped to zero. These pairwise values are stored once per benchmark.

Algorithm 1 Iterative refinement of measurement routes

Input: Initial CET $\mathcal{T}_{t,o}$, regions $\mathcal{R}$, $a_{\min}$, $d_{\min}$, $R_{\max}$
Output: Refined measured tree

Instantiate candidate routes and remove those violating Tool Manifest requirements
for each route $\ell = (k, r, C)$ **do**
    Compute $a_k(r)$ and $d_k(r)$
end for
Prune routes with $a_k(r) < a_{\min}$
for $i = 1$ to $R_{\max}$ **do**
    $\mathcal{F} \leftarrow \{\ell : d_k(r) < d_{\min}\}$
    if $\mathcal{F} = \emptyset$ **then**
        break
    end if
    for each $\ell \in \mathcal{F}$ **do**
        $\ell \leftarrow$ REPARAMETERIZE(ℓ)
        Re-evaluate route and recompute $d_k(r)$
    end for
end for
Prune routes with $d_k(r) < d_{\min}$
Compute $w_k(r) = a_k(r)d_k(r)$ for surviving routes
return Refined measured tree

At inference, discounted agreement is computed for every pair of surviving routes measuring the same concept within a region. These pairwise values are retained alongside the corresponding measurements and reliabilities and passed to the orchestrating agent for concept-level interpretation.

Table 10: Summary of datasets used in our experiments.

Task	Dataset	Organ	Prediction target	# Samples	Metric
Grading	SICAP	Prostate	Gleason grade	155 WSIs	QWK / W-F1
	KMC-RCC	Kidney	RCC grade	722 images	QWK / W-F1
	TCGA-UCEC	Endometrium	Cancer grade	566 WSIs	QWK / W-F1
Metastasis	CAMELYON16	Lymph node	Metastasis status	399 WSIs	BACC / W-F1
	CAMELYON17	Lymph node	Metastasis status	500 WSIs	BACC / W-F1
Subtyping	TCGA-RCC	Kidney	RCC subtype	884 WSIs	BACC / W-F1
	TCGA-NSCLC	Lung	NSCLC subtype	1,041 WSIs	BACC / W-F1
sTIL scoring	TCGA-BRCA	Breast	Ordinal sTIL score	1,035 WSIs	QWK / W-F1
	TIGER-WSITILS	Breast	Ordinal sTIL score	82 WSIs	QWK / W-F1
Survival	TCGA-BLCA	Bladder	Overall survival	437 WSIs	C-index
	TCGA-BRCA	Breast	Overall survival	1,017 WSIs	C-index
	TCGA-LUAD	Lung	Overall survival	529 WSIs	C-index
	TCGA-UCEC	Endometrium	Overall survival	539 WSIs	C-index

D Additional Experimental Details

D.1 Datasets and Evaluation Protocols

We evaluate XPathAgent on 13 benchmark–task combinations spanning histologic grading, metastasis detection, tumor subtyping, stromal tumor-infiltrating lymphocyte (sTIL) scoring, and survival prognosis. Table D.1 summarizes the prediction target, number of samples, and evaluation metric for each benchmark. Unless otherwise stated, we follow the official dataset definitions and evaluation protocols.

For the zero-shot tasks, all available labeled examples in the designated evaluation cohort are used only for computing benchmark performance and are not used for task-specific training or fine-tuning. Survival prognosis is evaluated using five-fold patient-level cross-validation, with the downstream ABMIL survival head trained on four folds and evaluated on the held-out fold. The specialist tools, Diagnostic Atlas, and ICR components remain fixed across folds. For cancer grading, we retain the ordinal grading labels provided by each benchmark. Metastasis detection is evaluated as binary classification of metastasis presence versus absence using the available CAMELYON annotations. For tumor subtyping, we retain the native histologic subtype labels: KIRC, KIRP, and KICH for TCGA-RCC; and LUAD and LUSC for TCGA-NSCLC. For sTIL scoring, continuous sTIL percentages are converted into three ordinal categories: low ($< 10\%$), intermediate (10–40%), and high ($> 40\%$). The same label mapping is used for all methods evaluated on the corresponding benchmark.

D.2 Adaptation of Comparative Methods

Pathology foundation models. PLIP, CONCH, MUSK, and PRISM are evaluated with their publicly released pretrained weights using their native zero-shot vision–language interfaces. For each benchmark, class-specific text prompts are constructed from the target label space and predictions are obtained without task-specific fine-tuning. Patch-level models are aggregated to the WSI level, while PRISM uses its native slide-level representation. For tasks not included in the original evaluations, we change only the benchmark-specific class prompts.

WSI reasoning baselines. PathAgent, WSI-Agents, and HistoSelect are evaluated using their released checkpoints and original WSI reasoning pipelines. We reformulate each benchmark as a closed-ended VQA/MCQ-style task whose answer options correspond to the target label space. Their native navigation, evidence-selection, verification, and reasoning procedures are otherwise retained, and no additional task-specific fine-tuning is performed on our evaluation benchmarks.

Survival baselines. For ResNet-50+ABMIL, SurvPath, HFGPI, and MDCS-MoAME, we follow their published survival-learning setups using the same patient cohorts and five-fold patient-level cross-validation protocol wherever applicable. For pathology foundation-model survival baselines, the pretrained encoder is kept fixed and only the downstream ABMIL survival head is trained.

D.3 Clinical Review and Curation of the Diagnostic Atlas

The Diagnostic Atlas was curated with feedback from three clinicians using a structured review protocol. The review was designed to assess two aspects of the Atlas separately: (i) the clinical validity of individual pathology concepts and their definitions, and (ii) the appropriateness of the concepts assigned to each downstream pathology task.

Review protocol. Each clinician was provided with a standardized review sheet containing the candidate pathology ontology and task–concept mappings used by XPathAgent. For each pathology concept, reviewers assessed the statement *"The description for the given pathology concept in the Knowledge Base is accurate and sufficient"* on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Reviewers could add free-text comments to identify inaccurate definitions, missing morphological cues, ambiguous terminology, or additional operational details required for interpretation.

In a second stage, clinicians reviewed the mapping between pathology tasks and their supporting concepts. For each task–organ pair, reviewers assessed whether the proposed concept set was necessary and sufficient for the corresponding task, again using a five-point Likert scale. Concepts designated as *mask* or *context* variables were explicitly identified as auxiliary concepts rather than primary diagnostic drivers. Reviewers were also asked to use the free-text field to identify missing concepts, task-specific caveats, or concepts that were inappropriate for a particular organ or diagnostic setting.

Atlas refinement. Clinician feedback was used to revise both the ontology and the task–concept assignments. Rather than treating the numerical ratings as supervision for the agent, they were used exclusively as a quality-control mechanism during Atlas construction. Free-text feedback was particularly used to refine concept definitions, add task-specific morphological details, separate overly broad concepts, and correct or augment concept sets when clinically relevant evidence was missing. The resulting Atlas was fixed before the downstream evaluation of XPathAgent.

D.4 Additional Ablations and Analysis

Component contributions. Fig. 3 extends the component ablation to all non-survival benchmarks. Performance improves from the baseline to the full pipeline on every dataset. The CET provides substantial additional gains over the Diagnostic Atlas alone, particularly on KMC-RCC ($0.687 \rightarrow 0.709$), and CAMELYON16 ($0.789 \rightarrow 0.809$). Adding the complete reliability-aware pipeline further improves every benchmark. Thus, the contribution of the proposed components is not specific to survival prediction.

Contribution of specialist tools. Table 11 shows that removing any specialist reduces performance across all evaluated tasks. Removing PRISM produces the largest degradation consistently, including $0.822 \rightarrow 0.746$ on CAMELYON16 and $0.631 \rightarrow 0.553$ on TIGER-WSITILS, while removing TITAN produces smaller but consistent losses. These results indicate that the full system benefits from complementary evidence across specialists rather than depending on a single tool.

Reasoning model scale. Fig. 4 shows that the 4B reasoning model is consistently weaker, whereas the 27B and 35B models perform similarly across benchmarks. The 35B model provides only marginal improvements and is slightly worse than 27B on TCGA-RCC (0.937 vs. 0.939) and TIGER-WSITILS (0.629 vs. 0.631). We therefore use Qwen3.5-27B as the default reasoning backbone.

In-context adaptation. Table 12 compares PathAgent and XPathAgent using identical demonstrations. Both methods benefit from additional examples, but XPathAgent exhibits larger zero-to-five-shot gains on KMC-RCC ($+0.111$ vs. $+0.075$) and TCGA-BRCA ($+0.128$ vs. $+0.108$), reaching 0.833 and 0.748, respectively. This suggests that the structured evidence pipeline remains compatible with in-context adaptation rather than restricting the reasoning model's ability to use demonstrations.

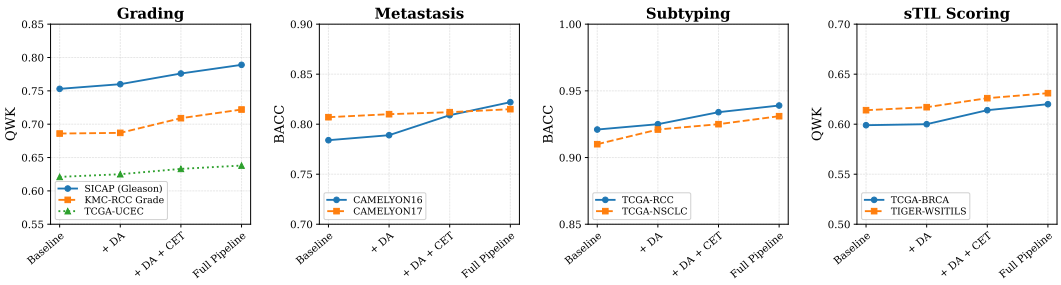

Figure 3: Component ablation of XPATHAGENT across pathology benchmarks. We progressively add the Diagnostic Atlas (DA), Concept-Evidence Tree (CET), and the complete reliability-aware pipeline. Performance is reported using QWK for cancer grading and sTIL scoring, and BAcc for metastasis detection and tumor subtyping.

Table 11: Leave-one-tool-out analysis across non-survival pathology benchmarks. We report QWK for grading and sTIL scoring, and BACC for metastasis detection and tumor subtyping. Each column removes one specialist while keeping all remaining components fixed.

Dataset	Full	w/o CellViT++	w/o CONCH	w/o PRISM	w/o TITAN
Grading (QWK)					
SICAP (Gleason)	**0.789**	0.737	0.744	0.749	0.764
KMC-RCC Grade	**0.722**	0.691	0.697	0.685	0.700
TCGA-UCEC	**0.638**	0.611	0.594	0.579	0.599
Metastasis Detection (BACC)					
CAMELYON16	**0.822**	0.780	0.759	0.746	0.774
CAMELYON17	**0.815**	0.764	0.752	0.741	0.789
Tumor Subtyping (BACC)					
TCGA-RCC	**0.939**	0.902	0.921	0.878	0.917
TCGA-NSCLC	**0.931**	0.895	0.883	0.890	0.907
sTIL Scoring (QWK)					
TCGA-BRCA	**0.620**	0.570	0.588	0.545	0.575
TIGER-WSITILS	**0.631**	0.580	0.569	0.553	0.605

Reliability and corroboration. Table 13 isolates the evidence-verification mechanism while holding the Diagnostic Atlas, CET, tools, and reasoning model fixed. Model-consensus verification is inconsistent, falling below no reliability weighting on three of the four benchmarks. In contrast, ICR without co-failure correction improves performance on all four datasets, and accounting for correlated failures provides a further gain in each case. Full ICR therefore achieves the strongest result on every evaluated task, supporting both input-conditional reliability estimation and dependence-aware corroboration.

E LIMITATIONS

While XPATHAGENT improves evidence integration across heterogeneous pathology tools, several limitations remain. First, Invariance-Calibrated Reliability is a label-free proxy that measures a tool's stability under controlled degradation rather than its correctness; a tool with a systematic semantic bias may remain stable and thus be deemed reliable. Second, the Diagnostic Atlas is expert-curated and closed-world: extending the framework to new tasks or organs requires additional curation, the concept vocabulary is bounded by established guidelines and cannot surface novel prognostic morphology, and the fixed linear, polarity-weighted aggregation does not capture context-dependent concept interactions.

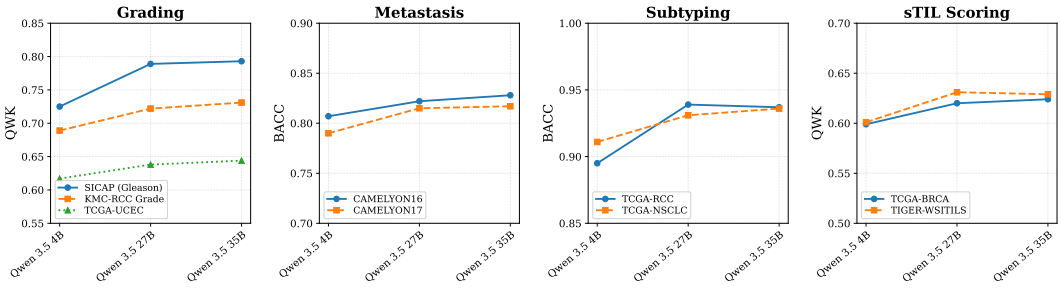

Figure 4: Effect of reasoning model scale across four pathology benchmarks. All XPATHAGENT components and specialist tools are kept fixed, and only the reasoning model is varied. We report QWK for cancer grading and sTIL scoring, and BAcc for metastasis detection and tumor subtyping.

Table 12: In-context adaptation of PathAgent and XPATHAGENT. Both methods receive the same demonstrations at each shot level. Δ denotes the improvement from zero- to five-shot inference.

Dataset	Method	0-shot	1-shot	5-shot	$\Delta_{0\to5}$
KMC-RCC	PathAgent	0.699	0.725	0.774	+0.075
	XPATHAGENT	0.722	0.787	**0.833**	**+0.111**
TCGA-BRCA	PathAgent	0.445	0.481	0.553	+0.108
	XPATHAGENT	0.620	0.664	**0.748**	**+0.128**

Table 13: Effect of the evidence-verification mechanism across representative pathology tasks. All variants use the same Diagnostic Atlas, Concept–Evidence Tree, specialist tools, and reasoning backbone; only the mechanism used to assess and combine tool-derived evidence is changed.

Verification mechanism	SICAP QWK ↑	CAMELYON16 BACC ↑	TCGA-RCC BACC ↑	TCGA-BRCA QWK ↑
No reliability weighting	0.776	0.809	0.934	0.614
Model-consensus verification Lyu et al. (2025)	0.769	0.803	0.936	0.610
ICR w/o co-failure correction	0.783	0.814	0.937	0.618
Full ICR	**0.789**	**0.822**	**0.939**	**0.620**