BIOTRACE: A VERIFIABLE PATHOLOGY AGENT FOR TRACING MOLECULAR-TO-PHENOTYPE EVIDENCE IN WHOLE-SLIDE VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide pathology visual question answering (VQA) often requires inferring clinically relevant phenotypes from evidence distributed across tissue regions and scales. Existing methods, including recent pathology agents, mainly acquire and reason over morphological evidence, leaving gene- and pathway-level biological support largely unexposed. Moreover, agent-based methods typically treat retrieved observations as additional context, without explicitly determining whether the evidence required by the question has been sufficiently resolved. To address these limitations, we introduce **BioTrace**, an agent system that traces question-specific evidence along a biological hierarchy with iterative evidence acquisition and verification. BioTrace organizes multiscale H&E representations into a Gene–Pathway–Phenotype evidence space, exposing concept-level predictions, concept-specific spatial responses, and supporting biological associations as queryable evidence alongside morphological observations. During inference, a pathology knowledge-guided verifier assesses evidence sufficiency, while unresolved evidence requirements guide subsequent acquisition. Experiments show that the learned evidence space performs strongly across phenotype-, pathway-, and gene-level prediction tasks. Building on this evidence space, the full BioTrace system consistently outperforms the compared pathology VQA and agent baselines on WSI-VQA and SlideBench-BCNB. Code is available at: `https://anonymous.4open.science/r/BioTrace/`

1 INTRODUCTION

Visual question answering (VQA) on gigapixel H&E whole-slide images (WSIs) aims to answer clinically meaningful questions about pathological phenotypes spanning diagnostic, molecular, and clinical characteristics (Chen et al., 2024a; 2025b). Compared with conventional image VQA, pathology VQA involves substantially larger visual fields, with relevant evidence often distributed across distant tissue regions and multiple observation scales (Chen et al., 2025a; Buckley et al., 2025; Yang et al., 2026). More fundamentally, pathological phenotypes are manifested in morphology but not determined by it alone: they reflect underlying gene and pathway activity, of which tissue morphology is only a partial projection (Schmauch et al., 2020; Fu et al., 2020). Answering such questions therefore requires not only locating relevant morphology, but also relating it to the molecular programs it reflects. Thus, an effective pathology VQA system should adaptively determine (i) where to look in the slide, (ii) at which scale to observe, (iii) what kind of evidence to acquire, morphological or biological, and (iv) whether the accumulated evidence suffices to support the answer, or which gap remains to be resolved.

Existing pathology VQA methods mainly address the first two in different ways, while largely overlooking the last two. As shown in Fig. 1(a), end-to-end **VLM-based methods** typically encode supporting information implicitly in attention maps or visual tokens, making it difficult to identify which evidence actually supports the final answer (Lu et al., 2024b; Chen et al., 2025b; Liang et al., 2025; Seyfioglu et al., 2024). More recently, another line of methods, i.e., **Agent-based methods**, can actively inspect additional regions or invoke external tools, as illustrated in Fig. 1(b) (Chen et al., 2025a; Buckley et al., 2025; Yang et al., 2026; Wong et al., 2026). They could inspect more regions

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.

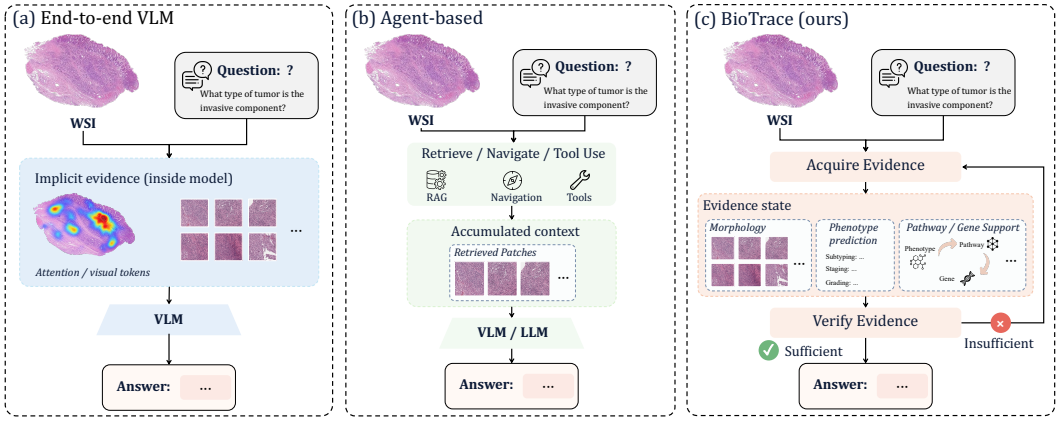

Figure 1: Comparison of evidence handling in pathology VQA. (a) End-to-end VLM methods keep supporting evidence implicit in model representations. (b) Agent-based methods progressively accumulate additional context. (c) BioTrace explicitly maintains a structured evidence state, and uses evidence verification to guide further acquisition.

across multiple scales and obtain more diverse interpretations of pathological evidence, providing stronger evidence localization capabilities compared to VLM-based methods.

However, both paradigms primarily focus on morphological evidence, leaving the biological processes underlying pathological phenotypes largely unexposed (Chen et al., 2025b; Yang et al., 2026). This suggests that the evidence space should extend beyond morphology to include pathological phenotype predictions and the underlying biological relationships linking genes, pathways, and phenotypes. Moreover, retrieved observations are often treated merely as additional context, without tracking whether the evidence required by the question has been sufficiently resolved (Chen et al., 2025a; Buckley et al., 2025; Wong et al., 2026). This calls for explicit evidence verification, so that remaining evidence gaps can guide subsequent acquisition. Together, these two limitations motivate a structured evidence space with iterative evidence acquisition and verification.

To address these challenges, we introduce **BioTrace**, an evidence-centered agent system that couples question-driven acquisition with explicit verification. As shown in Fig. 1(c), BioTrace first turns multiscale WSI representations into a queryable evidence space organized by Gene–Pathway–Phenotype relations, where prior-constrained, learned cross-level relations link phenotype predictions to their underlying pathways and genes. Built upon this evidence space, we construct an agent system that, beyond morphological observations, further queries concept-level predictions, concept-specific spatial responses, and supporting biological associations as evidence for the current question. The acquired evidence is then verified against the current requirement, and any remaining gap is fed back to guide subsequent acquisition. Through this closed loop, BioTrace progressively gathers the evidence needed for the question and grounds the final response in explicitly acquired and verified evidence. Experiments on WSI-VQA and SlideBench-BCNB demonstrate that BioTrace outperforms the compared baselines. The learned evidence space also shows strong performance across phenotype-, pathway-, and gene-level prediction tasks, supporting its use as a structured evidence interface for downstream pathology reasoning.

Our contributions are threefold:

(1) We introduce **BioTrace**, an evidence-centered agent system for pathology VQA that organizes reasoning around explicit, question-specific evidence rather than implicitly accumulated context.

(2) We propose a biologically structured multiscale evidence space organized by Gene–Pathway–Phenotype relations and their underlying cross-level biological associations. It exposes concept-level predictions, concept-specific spatial responses, and supporting biological associations as a unified, queryable interface for pathology reasoning.

(3) We develop an adaptive evidence reasoning mechanism in which unresolved evidence requirements guide acquisition across evidence sources and observation scales, while verification feeds remaining gaps back to subsequent acquisition.

2 Related Work

Structured pathology representations as evidence. Classical WSI methods aggregate patch features into slide-level representations through multiple-instance learning (Ilse et al., 2018; Lu et al., 2021), while pathology foundation models such as UNI (Chen et al., 2024b), Prov-GigaPath (Xu et al., 2024), and PRISM (Shaikovski et al., 2024) further improve the generality of tile- and slide-level representations. Multimodal pathology extends these representations with molecular or language supervision: MCAT (Chen et al., 2021) integrates histology and genomics, and CONCH (Lu et al., 2024a) advances pathology image–text alignment. These studies show that WSI representations can capture rich morphological and biological information, but such information is typically optimized for prediction or multimodal alignment and remains implicit in the learned representation. BioTrace instead organizes WSI-derived information into a multiscale Gene–Pathway–Phenotype evidence space, exposing concept-level predictions, concept-specific spatial responses, and biological associations as explicit, queryable evidence.

Evidence acquisition and reasoning in pathology VQA. Pathology VQA models such as PathChat (Lu et al., 2024b) and SlideChat (Chen et al., 2025b) perform end-to-end vision–language reasoning, while recent systems increasingly incorporate retrieval, navigation, and tool use. General agentic methods have explored retrieval (Lewis et al., 2020; Asai et al., 2024), external actions (Yao et al., 2022), feedback, and memory (Shinn et al., 2023), and pathology agents further extend these ideas to multi-round WSI analysis. Our downstream evaluation accordingly covers pathology-specific VQA models, including SlideChat, WSI-LLaVA (Liang et al., 2025), Quilt-LLaVA (Seyfioglu et al., 2024), MedDr (He et al., 2024), and MedGemma (Sellergren et al., 2026), together with agentic methods including GIANT (Buckley et al., 2025), BEACON (Wong et al., 2026), PathAgent (Chen et al., 2025a), and PathNavigate (Yang et al., 2026). Although these approaches improve pathology question answering and information acquisition, the acquired information is often accumulated as context. They do not explicitly track whether the evidence required by the current question has been satisfied. BioTrace instead verifies whether the accumulated evidence satisfies the unresolved requirement and uses the remaining evidence gap to guide subsequent acquisition.

3 Method

As illustrated in Fig. 2, we formulate pathology VQA as question-driven sequential evidence acquisition and verification. BioTrace organizes multiscale WSI representations into a structured evidence space over Gene, Pathway, and Phenotype concepts, providing concept-level predictions, concept-specific spatial responses, and cross-level biological relations as queryable evidence. Given a question, the Agent identifies the unresolved evidence requirement, acquires the corresponding predictive, visual, or biological evidence at an appropriate observation scale, and updates the evidence state. The Evidence Verifier then assesses whether the accumulated evidence is sufficient; any remaining gap is fed back to guide the next acquisition step until the evidence supports an answer.

3.1 Problem Formulation and Framework Overview

Given an H&E WSI I and a pathology question q, our goal is to acquire question-relevant patient evidence and generate an answer supported by that evidence. At each scale $s \in \mathcal{S}$, the WSI is pre-encoded into N_s patch features, $H^{(s)} \in \mathbb{R}^{N_s \times d}$, while preserving spatial correspondence across scales. The Planner converts the question into an initial structured evidence requirement, $r_0 = \mathrm{Plan}(q)$, which specifies the target concept, required evidence, preferred scale, and answer constraints. At round k, the Agent selects an acquisition action a_k according to the current requirement r_k and evidence state $\mathcal{E}_k$. Here, $a_k = (c_k, s_k)$ specifies the concept c_k to be queried and the observation scale s_k at which evidence is acquired. The resulting evidence e_k is then incorporated into the evidence state

$$\mathcal{E}_{k+1} = \mathrm{Update}\left(\mathcal{E}_k, a_k, e_k\right). \tag{1}$$

The process continues until the accumulated evidence is verified as sufficient to support the answer or the current requirement cannot be further resolved by the available evidence.

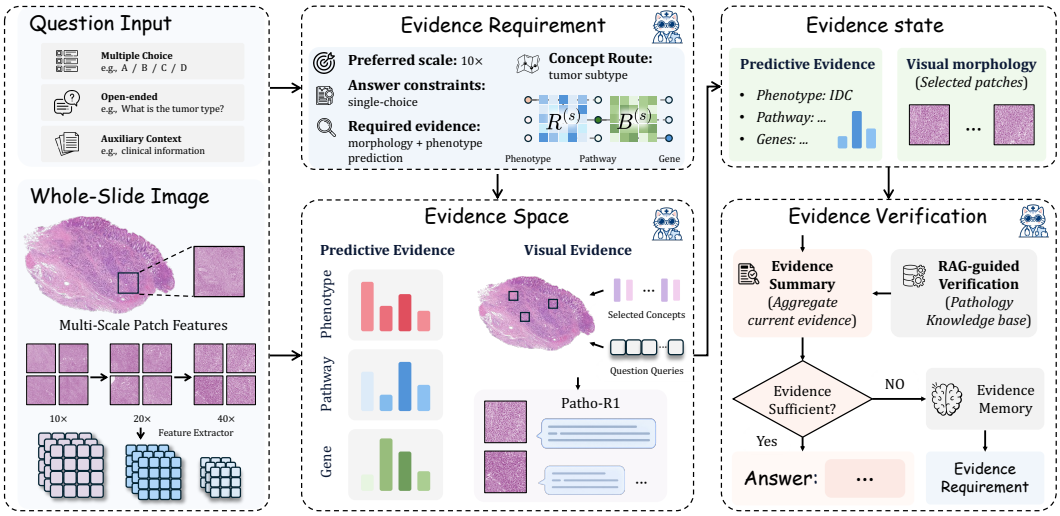

Figure 2: Overview of BioTrace

3.2 Evidence Space

We construct a biologically structured evidence space that transforms multiscale WSI features into queryable concept-level evidence. It grounds Gene, Pathway, and Phenotype concepts in the WSI and organizes them through cross-level biological relations, yielding concept-level predictions, spatial responses, and structured biological associations for subsequent evidence acquisition.

Concept-Level Evidence Grounding. As illustrated in Fig. 3, the evidence space is organized over three biological concept sets: Genes $\mathcal{G}$, Pathways $\mathcal{P}$, and Phenotypes $\mathcal{T}$. Each concept is associated with a WSI-derived representation and a spatial response over patches, allowing the same evidence representation to support subsequent prediction, localization, and evidence acquisition. At each scale s, we ground every concept directly in the WSI through a concept-specific prototype. For concept $c \in \mathcal{G} \cup \mathcal{P} \cup \mathcal{T}$ with prototype $u_c \in \mathbb{R}^d$, we compute

$$A_c^{(s)} = \mathrm{softmax}\left(\frac{u_c (H^{(s)})^\top}{\sqrt{d}}\right), \qquad \tilde{z}_c^{(s)} = A_c^{(s)} H^{(s)}. \tag{2}$$

Here, $A_c^{(s)}$ denotes the spatial response of concept c over WSI patches, while $\tilde{z}_c^{(s)}$ aggregates the corresponding patch features into a concept-specific representation. The spatial response provides localized evidence for subsequent visual inspection, whereas the concept representation supports concept-level prediction and biological relation modeling. This prototype-guided aggregation is shared across all three concept levels and preserves explicit spatial grounding at each scale.

Direct concept grounding captures the WSI evidence associated with each individual concept, but does not yet encode their biological dependencies. We therefore connect adjacent concept levels through scale-specific relation matrices

$$W_{\ell-1\to\ell}^{(s)} = \begin{cases} B^{(s)}, & \mathcal{G} \to \mathcal{P}, \\ R^{(s)}, & \mathcal{P} \to \mathcal{T}, \end{cases} \tag{3}$$

where $B^{(s)}$ denotes Gene–Pathway relations and $R^{(s)}$ denotes Pathway–Phenotype relations. Their connectivity is initialized and constrained by biological priors (Liberzon et al., 2015), while the scale-specific relation strengths are optimized during training. Details are in the **Appendix B**

For an upper-level concept $c \in \mathcal{C}_\ell$, its direct WSI representation is enhanced by relation-weighted representations from the immediately lower level

$$z_c^{(s)} = \tilde{z}_c^{(s)} + \sum_{j \in \mathcal{C}_{\ell-1}} W_{jc}^{(s)} z_j^{(s)}. \tag{4}$$

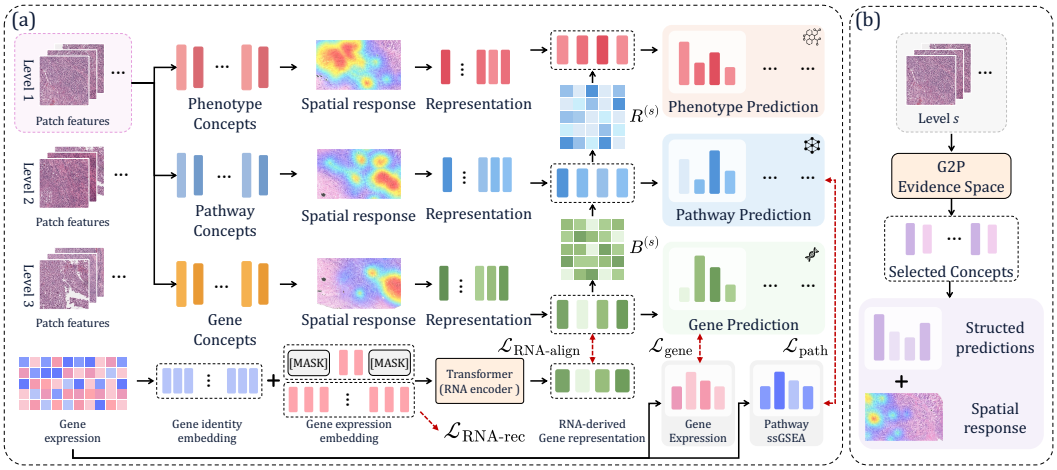

Figure 3: Overview of the biologically structured evidence space. (a) Multiscale WSI features are transformed into Gene-, Pathway-, and Phenotype-level predictions and spatial responses with structured biological relations. Red arrows denote training-only supervision from phenotype annotations and paired RNA-seq. (b) The learned evidence space provides structured and spatial evidence for downstream inference.

Genes form the bottom level and therefore use $z_g^{(s)} = \tilde{z}_g^{(s)}$. Pathways are enhanced by Gene representations through $B^{(s)}$, while Phenotypes are further enhanced by Pathway representations through $R^{(s)}$. This yields concept representations that preserve direct morphology grounding while incorporating structured biological context.

Structured Biological Evidence Learning. The resulting concept representations are mapped to structured concept-level predictions through lightweight concept-specific heads

$$\hat{y}_c^{(s)} = h_c\left(z_c^{(s)}\right), \qquad c \in \mathcal{G} \cup \mathcal{P} \cup \mathcal{T}, \tag{5}$$

where the supervision depends on the biological level. Phenotype concepts are optimized using the corresponding classification, regression, or survival objective. Gene concepts are supervised by normalized patient-specific gene-expression values, while Pathway concepts are supervised by patient-specific ssGSEA (Barbie et al., 2009) enrichment scores derived from paired RNA-seq. This yields the prediction losses $\mathcal{L}_{\text{pheno}}$, $\mathcal{L}_{\text{gene}}$, and $\mathcal{L}_{\text{path}}$.

Paired RNA-seq further provides auxiliary molecular supervision during training. We construct an RNA-derived Gene representation by combining Gene identity embeddings with expression embeddings and randomly masking a subset of expression values before an RNA encoder. The resulting representation is aligned with its WSI-derived counterpart using a cosine alignment loss $\mathcal{L}_{\text{RNA-align}}$, while the masked expression values are reconstructed with $\mathcal{L}_{\text{RNA-rec}}$. These RNA-derived signals are used only during training and are not required at inference time. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{pheno}} + \lambda_p \mathcal{L}_{\text{path}} + \lambda_g \mathcal{L}_{\text{gene}} + \lambda_{\text{align}} \mathcal{L}_{\text{RNA-align}} + \lambda_{\text{rec}} \mathcal{L}_{\text{RNA-rec}} + \mathcal{L}_{\text{reg}}, \tag{6}$$

where $\mathcal{L}_{\text{reg}}$ collects the regularization terms for structured relation learning. After training, the module exposes concept-level predictions, concept-specific spatial responses, and the learned biological relations for downstream evidence acquisition and routing.

3.3 Evidence State

At round k, the acquisition policy selects the next action according to the question, current evidence state, and unresolved requirement

$$a_k = \mathcal{P}_{\text{acq}}\left(q, \mathcal{E}_k, r_k\right). \tag{7}$$

For the concept and observation scale specified by a_k, the acquired evidence takes two complementary forms. **Predictive Evidence** provides a slide-level estimate through concept-level predictions,

while **Visual Evidence** uses concept-specific spatial responses to localize regions for morphology-grounded observation. Together, they connect global concept-level predictions with local H&E morphology across reasoning rounds.

Predictive Evidence. When predictive evidence is required, the corresponding concept-level prediction $\hat{y}_{c_k}^{(s_k)}$ is retrieved from the structured evidence space. To make reliability comparable across heterogeneous prediction tasks, we normalize held-out validation performance as

$$c_k^{\text{pred}} = \frac{1}{2} + d_{m_k}\left(\frac{v_{c_k,m_k}^{(s_k)} - L_{m_k}}{U_{m_k} - L_{m_k}} - \frac{1}{2}\right), \qquad c_k^{\text{pred}} \in [0,1]. \tag{8}$$

Here, $v_{c_k,m_k}^{(s_k)}$ denotes validation performance under metric m_k, L_{m_k} and U_{m_k} are its normalization bounds, and $d_{m_k} \in \{-1,1\}$ accounts for metric polarity. The resulting score provides a unified reliability measure for both discrete and continuous predictions, distinct from evidence sufficiency.

Visual Evidence. For morphological evidence, candidate patches are selected from concept-responsive regions and question-relevant regions

$$\mathcal{P}_k = \text{TopK}\left(A_{c_k}^{(s_k)}\right) \cup \text{Retrieve}\left(q, r_k, H^{(s_k)}\right). \tag{9}$$

The selected patches are examined by the pathology visual observer to obtain morphology-grounded evidence. The acquired predictive or visual evidence, together with its provenance (c_k, s_k) and predictive reliability when applicable, constitutes e_k and updates the evidence state through Eq. 1.

3.4 Evidence Verification

Evidence Matching and Verification. After each acquisition round, we employ retrieval-augmented verification based on RAG (Lewis et al., 2020). Our pathology RAG module retrieves question-specific knowledge and matches it against the updated evidence state

$$\mathbf{m}_k = \text{RAG}\left(q, r_k, \mathcal{E}_{k+1}\right), \tag{10}$$

where $\mathbf{m}_k$ denotes the evidence–knowledge matching scores. The retrieved knowledge is used only for verification and is not incorporated into the patient evidence state.

The Evidence Verifier then assesses evidence sufficiency

$$(\sigma_k, r_{k+1}) = \mathcal{V}\left(q, r_k, \mathcal{E}_{k+1}, \mathbf{m}_k\right), \tag{11}$$

where $\sigma_k \in \{\text{sufficient}, \text{partial}, \text{conflicting}, \text{insufficient}, \text{unavailable}\}$ denotes the verification status, and r_{k+1} specifies the remaining evidence requirement. The pair $(\hat{y}_{c_k}^{(s_k)}, c_k^{\text{pred}})$ is retained in $\mathcal{E}_{k+1}$, allowing prediction reliability to be jointly evaluated with visual and biological evidence.

Feedback Routing and Answering. When additional biological support is required, the structured hierarchy refines the unresolved requirement

$$r_{k+1} \leftarrow r_{k+1} \cup \begin{cases} \text{TopK}_{p\in\mathcal{P}}\, R_{p,c_k}^{(s_k)}, & c_k \in \mathcal{T}, \\ \text{TopK}_{g\in\mathcal{G}}\, B_{g,c_k}^{(s_k)}, & c_k \in \mathcal{P}. \end{cases} \tag{12}$$

The refined r_{k+1} is returned to the acquisition policy in Eq. 7, while previously examined evidence is retained in lightweight working memory to avoid redundant acquisition.

If σ_k indicates sufficient evidence, the final answer is generated as

$$\hat{y} = \text{Answer}\left(q, \mathcal{E}_{k+1}\right). \tag{13}$$

Otherwise, r_{k+1} guides the next acquisition round while further evidence is available. At the final round or when no additional evidence can be acquired, the accumulated evidence in $\mathcal{E}_{k+1}$ is used to produce the answer with the remaining uncertainty retained. The complete evidence-acquisition and verification protocol is provided in the **Appendix A.4**.

Table 1: Main results on WSI-VQA. All values are expressed as percentages (%). SlideChat is omitted due to training overlap. Boldface denotes the best result and underline the runner-up.

Method	Input	Total	MCQ	Open	BLEU-4	ROUGE-L
General-purpose MLLMs						
GPT-5.6 Sol	Slide-T	<u>55.91</u>	**58.97**	<u>52.46</u>	<u>76.00</u>	54.94
GPT-5.6 Luna	Slide-T	47.08	47.18	46.96	73.36	51.13
Qwen3.5-9B	Slide-T	41.74	37.63	46.38	46.54	38.03
Pathology-specific Models						
MedDr	Slide-T	35.24	45.38	23.77	49.46	22.40
Quilt-LLaVA	Patch	33.65	43.55	22.46	18.20	25.55
MedGemma-1.5-4B	Slide-T	38.01	43.55	31.74	22.58	36.82
WSI-LLaVA	Patch	35.22	38.17	31.88	46.73	41.81
Pathology Agents						
GIANT	WSI	32.78	41.40	23.04	35.75	21.45
BEACON	WSI	41.27	40.86	41.74	31.74	34.03
PathAgent	WSI	46.83	51.08	42.03	31.06	43.28
PathNavigate	WSI	50.29	53.23	46.96	66.76	<u>55.48</u>
BioTrace (Ours)	WSI	**60.93**	<u>57.14</u>	**65.22**	**82.46**	**66.30**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We evaluate BioTrace on two breast pathology benchmarks, WSI-VQA (Chen et al., 2024a) and SlideBench-BCNB (Chen et al., 2025b). WSI-VQA contains both multiple-choice and open-ended pathology questions, while SlideBench-BCNB evaluates diagnostic tasks including Tumor, ER, PR, HER2, Grading, and Molecular Subtype. The evidence space is trained using TCGA-BRCA (Network., 2012) WSIs with phenotype supervision and paired RNA-seq data. For WSI-VQA, evaluation is performed on the designated test set, with no overlap between the test WSIs and those used for training the structured evidence module. We report MCQ accuracy, open-ended accuracy, BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and overall accuracy on WSI-VQA. For SlideBench-BCNB, we report the accuracy of each diagnostic sub-task and the official weighted Overall score. Additional details are provided in the **Appendix A.1**.

Implementation Details. We use CONCH (Lu et al., 2024a) for both visual and textual feature encoding. WSI features are extracted at multiple magnification levels (40×, 20×, and 10×), and Patho-R1 (Zhang et al., 2026b) is used for visual pathology interpretation. Paired RNA-seq is used only as training-time supervision for the structured evidence module and is not required during inference. More implementation details are provided in the **Appendix A.3**.

Comparison Methods. For the structured evidence module, we compare with representative pathology foundation models, including GigaPath (Xu et al., 2024), PRISM (Shaikovski et al., 2024), TITAN (Ding et al., 2025), and CARE (Zhang et al., 2026a). For pathology VQA, we compare with three groups of representative methods. General-purpose MLLMs include GPT-5.6 Sol, GPT-5.6 Luna (OpenAI, 2026), and Qwen3.5-9B; pathology-specific models include SlideChat (Chen et al., 2025b), WSI-LLaVA (Liang et al., 2025), Quilt-LLaVA (Seyfioglu et al., 2024), MedDr (He et al., 2024), and MedGemma-1.5-4B (Sellergren et al., 2026); and pathology agents include GIANT (Buckley et al., 2025), BEACON (Wong et al., 2026), PathAgent (Chen et al., 2025a), and PathNavigate (Yang et al., 2026). Additional details are provided in the **Appendix A.2**

4.2 MAIN RESULTS

Pathology VQA performance. Tables 1 and 2 summarize the downstream VQA results. On **WSI-VQA**, BioTrace achieves the best overall performance, with particularly strong gains on open-ended questions across both accuracy and generation metrics. Since open-ended questions provide no

Table 2: Main results on SlideBench-BCNB. All values are expressed as percentages (%). Boldface denotes the best result and underline the runner-up.

Method	Input	Overall	Tumor	ER	PR	HER2	Grading	Mol. Sub.
General-purpose MLLMs								
GPT-5.6 Sol	Slide-T	53.58	83.08	67.86	54.91	44.19	49.03	<u>31.19</u>
GPT-5.6 Luna	Slide-T	51.49	**90.45**	62.48	52.46	43.15	42.76	24.86
Qwen3.5-9B	Slide-T	42.52	58.03	39.28	35.33	43.48	51.62	27.58
Pathology-specific Models								
SlideChat	WSI	55.41	89.04	71.62	<u>72.14</u>	47.45	29.55	27.39
WSI-LLaVA	WSI	52.94	80.81	61.98	56.24	<u>49.86</u>	42.79	27.79
Quilt-LLaVA	Patch	42.72	50.28	51.89	55.20	44.47	25.98	24.67
MedDr	Slide-T	41.50	54.10	39.66	45.33	44.66	39.96	22.00
MedGemma-1.5-4B	Slide-T	49.25	79.02	50.08	63.93	40.96	49.61	20.30
Pathology Agents								
GIANT	WSI	42.66	68.42	23.21	37.50	46.25	49.29	28.57
BEACON	WSI	43.31	59.09	33.33	40.91	46.03	52.31	26.67
PathAgent	WSI	53.97	84.81	61.48	63.33	40.93	<u>56.10</u>	30.53
PathNavigate	WSI	<u>57.49</u>	85.56	<u>72.63</u>	69.17	47.17	51.73	28.26
BioTrace (Ours)	WSI	**65.13**	<u>89.13</u>	**82.23**	**77.98**	**52.39**	**65.55**	**36.29**

candidate answers, they place greater demands on evidence acquisition and grounded free-form reasoning, highlighting the effectiveness of our acquisition and verification design. On **SlideBench-BCNB**, BioTrace also performs best overall and ranks first on most diagnostic categories, with notable advantages on receptor- and molecular-related tasks.

Evidence Space. We first evaluate the evidence space across phenotype-, pathway-, and gene-level prediction tasks. As shown in Figure 4, our method consistently performs strongly across all three biological levels and achieves the best overall ranking among the compared methods. These results demonstrate the effectiveness of the proposed module in providing reliable multi-level biological evidence for downstream agent reasoning. Additional results are provided in the **Appendix C**

4.3 Ablations

As shown in Table 3, we conduct ablations on the major components of BioTrace to examine their contributions. We remove the Evidence Space and further retain only phenotype-level predictions to assess the role of evidence space at different levels. We also disable the Evidence Verifier and RAG Knowledge Base, and restrict visual acquisition to the 40× scale to evaluate verification, external knowledge, and multiscale observation. Overall, all components contribute to the performance, with the Evidence Space showing the largest impact.

Table 3: Component ablations on WSI-VQA Gold. All values are expressed as percentages (%). Best results are in bold.

Method	MCQ	Open	Total
BioTrace (Full)	**57.14**	**65.22**	**60.93**
w/o Evidence Space	43.55	54.36	48.62
w/o Evidence Verifier	55.83	62.93	59.16
w/o RAG Knowledge Base	54.63	62.27	58.22
Phenotype-only Evidence	52.75	63.32	57.71
40×-only Visual Evidence	56.75	63.85	60.08

4.4 Case Study

To illustrate the reasoning process of BioTrace, we present representative case studies in Figure 5. For each case, we visualize the question, phenotype-level prediction, selected H&E regions, and supporting evidence retrieved along the Phenotype–Pathway–Gene hierarchy, together with the final answer. These examples show how BioTrace organizes heterogeneous evidence from structured biological predictions and question-guided visual acquisition into an explicit reasoning trace. The pathway- and gene-level signals shown are inferred from the WSI and serve as supporting evidence rather than direct molecular measurements. Additonal cases are provided in the **Appendix D.2**

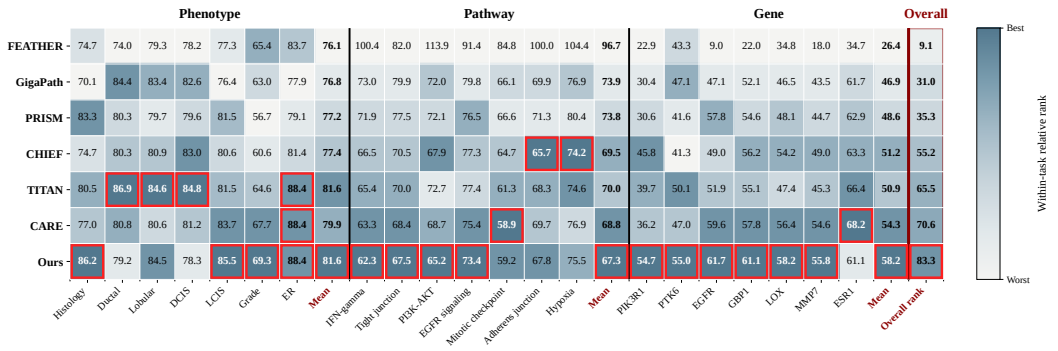

Figure 4: Comparison of evidence space across phenotype-, pathway-, and gene-level prediction tasks. Values are multiplied by 100. The Overall column shows the aggregated relative rank. Lower is better for pathway MAE, and red boxes indicate the best result for each task.

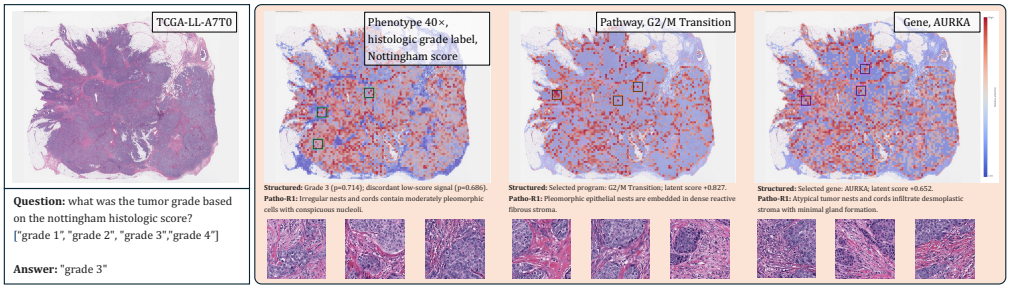

Figure 5: Evidence visualization for the histologic grading case study. The figure includes the WSI, question and answer, spatial response maps for grade-, pathway-, and gene-level concepts, Patho-R1 descriptions, and selected high-magnification H&E regions. Boxes mark the displayed regions, and the color bar denotes relative response strength.

4.5 LLM Robustness

We further evaluate the robustness of BioTrace by replacing Qwen3.5-9B with Qwen2.5-7B-Instruct and LLaMA-3.1-8B (Grattafiori et al., 2024), while keeping the remaining components unchanged. We also compare against a pathology-only Patho-R1-7B baseline operating directly on slide thumbnails without structured evidence space. As shown in Figure 6, BioTrace maintains comparable performance across different language backbones and substantially outperforms the pathology-only baseline, particularly on open-ended questions.

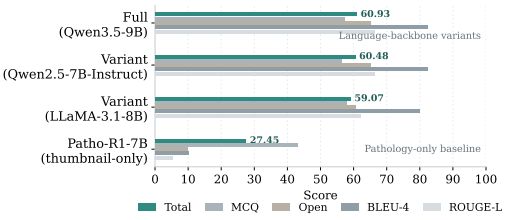

Figure 6: Robustness to different LLM backbones on WSI-VQA.

5 Conclusion

We introduced **BioTrace**, an evidence-centered agent framework for whole-slide pathology VQA that formulates question answering as iterative evidence acquisition and verification. BioTrace builds a biologically structured multiscale evidence space and enables the agent to adaptively acquire predictive, visual, and biological evidence according to unresolved question requirements. An evidence verifier further assesses evidence sufficiency and guides subsequent acquisition. Extensive experiments on WSI-VQA and SlideBench-BCNB demonstrate the effectiveness and robustness of the proposed framework. Overall, BioTrace provides an effective foundation for evidence-grounded reasoning over whole-slide pathology images.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development and manuscript editing. AI-assisted code and manuscript content were reviewed and verified by the authors. All experimental results, analyses, scientific claims, and conclusions were independently checked by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is intended for research purposes in computational pathology and does not provide clinical diagnoses or treatment recommendations. The experiments use de-identified or publicly available pathology datasets and benchmark resources according to their permitted research use. The phenotype-, pathway-, and gene-level outputs produced by our framework are model-derived predictions from H&E whole-slide images and should not be interpreted as substitutes for measured clinical assays such as RNA sequencing, immunohistochemistry, or other molecular tests. Although the proposed framework improves evidence-grounded pathology reasoning, deployment in clinical settings would require independent validation across institutions and populations, careful assessment of potential biases and failure modes, and appropriate oversight by qualified medical professionals.

REPRODUCIBILITY STATEMENT

We provide detailed experimental and implementation settings in the main paper and Appendix to support reproduction of the reported results. These include the dataset and evaluation protocols, baseline configurations, WSI feature-extraction and magnification settings, training and supervision details of the structured evidence module, Gene–Pathway–Phenotype definitions and relation priors, agent inference configurations, evidence-acquisition and verification procedures, model versions, retrieval parameters, and evaluation metrics. We also report the key hyperparameters and inference settings used throughout the experiments.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

David A Barbie, Pablo Tamayo, Jesse S Boehm, So Young Kim, Susan E Moody, Ian F Dunn, Anna C Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 2009.

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

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.

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, 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, 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, 2025b.

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.

Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature cancer*, 1(8):800–810, 2020.

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.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*, 1(3):6, 2024.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

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.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6): 417–425, 2015.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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.

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

OpenAI. Gpt-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026. Accessed: 2026-09-25.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):3877, 2020.

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.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13183–13192. IEEE, 2024.

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.

Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? *arXiv preprint arXiv:2506.09022*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

Bryan Wong, Xun Xu, Huazhu Fu, Nancy F Chen, and Mun Yong Yi. Beyond relevance: Bayesian evidence acquisition for agentic whole-slide image reasoning. *arXiv preprint arXiv:2608.05757*, 2026.

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.

Chunze Yang, Qidong Liu, Wenjie Zhao, Yue Tang, Jiusong Ge, Di Zhang, Jiashuai Liu, Lei Wu, Junbo Lu, Ni Zhang, et al. Pathnavigate: A training-free pathology agent with surprise-guided scan and shared slide memory for whole-slide image vqa. *arXiv preprint arXiv:2605.23559*, 2026.

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.

Di Zhang, Zhangpeng Gong, Xiaobo Pang, Jiashuai Liu, Junbo Lu, Hao Cui, Jiusong Ge, Zhi Zeng, Kai Yi, Yinghua Li, et al. Care: A molecular-guided foundation model with adaptive region modeling for whole slide image analysis. *arXiv preprint arXiv:2602.21637*, 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.

Supplementary Material

TABLE OF CONTENTS

A Experimental Settings and Reproduction Protocols 14
A.1 Datasets, Structured Supervision, and Metrics 14
A.2 Baseline Settings and Evaluation Scope 14
A.3 Implementation Details and Agent Configuration 14
A.4 Agent Reproduction Protocol 16

B Structured Biological Evidence 17
B.1 Gene–Pathway Prior and Pathway Supervision 17
B.2 Learned Pathway–Phenotype Relations 17

C Detailed Results of Structured Biological Evidence 17
C.1 Summary of Structured Biological Prediction 17
C.2 Phenotype-Level Results 17
C.3 Pathway-Level Results 18
C.4 Gene-Level Results 18

D Additional Analysis 22
D.1 Efficiency Analysis 22
D.2 Additional Visualization 23

A Experimental Settings and Reproduction Protocols

This section provides additional experimental and implementation details for reproducing the evaluation of BioTrace. We describe the benchmark settings, baseline configurations, structured-evidence supervision, agent inference parameters, and the evaluation protocol used in our experiments.

A.1 Datasets, Structured Supervision, and Metrics

We evaluate BioTrace on WSI-VQA (Chen et al., 2024a) and SlideBench-BCNB (Chen et al., 2025b). WSI-VQA contains an official test set of 735 questions associated with TCGA-BRCA whole-slide images, including 390 multiple-choice questions and 345 open-ended questions. SlideBench-BCNB contains 7,274 multiple-choice questions covering Tumor, ER, PR, HER2 status, HER2 expression, Grading, and Molecular Subtype. For compact reporting, HER2 status and HER2 expression are summarized in the HER2 column, while the Overall score follows the official benchmark aggregation protocol.

The structured evidence module is developed from TCGA-BRCA (Network., 2012) H&E WSIs together with phenotype annotations and paired RNA-seq data. RNA-seq is used only during training and is not required during downstream agent inference. Gene-expression targets are normalized using statistics estimated from the training cohort. For pathway supervision, we compute patient-specific ssGSEA (Barbie et al., 2009) enrichment scores using the fixed member genes of each pathway and standardize the resulting scores using training-cohort statistics. Gene–Pathway memberships are specified by biological prior knowledge (Liberzon et al., 2015) and remain fixed, whereas Pathway–Phenotype relations are optimized during training. At inference time, phenotype-, pathway-, and gene-level evidence is predicted entirely from H&E WSIs.

For WSI-VQA, we report Total accuracy, multiple-choice accuracy (MCQ), open-ended accuracy (Open), BLEU-4 (Papineni et al., 2002), and ROUGE-L (Lin, 2004). For SlideBench-BCNB, we report task-wise accuracy together with the official weighted Overall score. For structured biological prediction, phenotype endpoints are evaluated using the corresponding AUC, accuracy, or C-index; pathway prediction is evaluated using mean absolute error (MAE); and gene-expression prediction is evaluated using Pearson correlation. In the detailed structured-evidence tables, all values are multiplied by 100 for compact presentation.

A.2 Baseline Settings and Evaluation Scope

To make the comparison settings explicit, Table 4 summarizes the actual inference configuration used for each downstream VQA baseline. Depending on the model interface, methods receive whole-slide thumbnails, pre-extracted pathology features, or selected high-resolution image regions. When necessary, the original model or navigation mechanism is interfaced with our local pathology perception and language-reasoning stack while preserving the evidence-acquisition procedure used in the comparison.

For the structured biological prediction experiments, we additionally compare against FEATHER (Shao et al., 2025), GigaPath (Xu et al., 2024), PRISM (Shaikovski et al., 2024), CHIEF (Wang et al., 2024), TITAN (Ding et al., 2025), and CARE (Zhang et al., 2026a). These experiments use identical target definitions within each phenotype, pathway, and gene so that the comparison directly measures the biological information retained by the slide representations.

A.3 Implementation Details and Agent Configuration

BioTrace uses three scale-specific structured evidence models corresponding to $40\times$, $20\times$, and $10\times$ WSI representations. CONCH (Lu et al., 2024a) is used to encode visual pathology content and textual concepts. The three structured models are executed once for each patient, and the resulting phenotype-, pathway-, and gene-level predictions are cached for subsequent agent reasoning. This avoids repeatedly executing the prediction models during multi-round evidence acquisition.

Patho-R1-7B (Zhang et al., 2026b) serves as the pathology perceptor and converts selected image regions into morphology-grounded observations. Qwen3.5-9B is used for language-side question understanding, evidence verification, and final answer generation. The two models are exposed as

Table 4: Detailed inference settings of the methods used in the downstream pathology VQA experiments. The table reports the actual visual input and model configuration used in our evaluation.

Name in main tables	Source	Visual input	Backbone / actual setting
General-purpose MLLMs			
GPT-5.6 Sol	External API run	Uploaded WSI thumbnail(s)	GPT-5.6 Sol API endpoint
GPT-5.6 Luna	External API run	Uploaded WSI thumbnail(s)	GPT-5.6 Luna API endpoint
Qwen3.5-9B	Local inference	WSI overview thumbnail(s), with DX slides preferred	Local Qwen3.5-9B served with vLLM (Kwon et al., 2023); no G2P evidence or patch navigation
Pathology-specific Models			
SlideChat	Official checkpoint + adapted input	40× CONCH-v1.5 feature sequence	Official LongNet/projector with Qwen2.5-7B; fixed untrained 768-to-512 orthogonal projection
WSI-LLaVA	Official checkpoint + adapted input	40× CONCH-v1.5 features	`wsi-llava-v1.5-7b`; adaptive pooling to 576 tokens and zero-padding the feature dimension to the model-required width
MedDr	Official checkpoint	One DX WSI thumbnail	InternViT-6B + Yi-34B; official 448-pixel preprocessing with greedy decoding
MedGemma-1.5-4B	Official checkpoint	One DX WSI thumbnail	`medgemma-1.5-4b-it`; direct question answering
Quilt-LLaVA	Official checkpoint + WSI adapter	Five 20× RGB patches	`Quilt-Llava-v1.5-7b`; fixed-seed tissue sampling with deterministic majority voting
Pathology Agents			
PathAgent	Local implementation	Multiscale patch trajectory	10×/20×/40× CONCH-v1.5 retrieval; Patho-R1-7B perception with Qwen3.5-9B adjudication
PathNavigate	Official code + dataset adapter	5×/20× search-readout patches	10× and 40× CONCH-v1.5 features; CONCH-based query navigation; Patho-R1-7B + Qwen3.5-9B; RAG enabled
GIANT	Official navigation + local VLM adapter	WSI thumbnail and iterative crops	Official GIANT navigation and action space with Qwen3.5-9B as the local language backend; at most 20 navigation steps
BEACON	Local implementation	10× CONCH candidates + selected crops	Belief/EIG-based acquisition with CONCH-supported candidate selection; five patches per round; Patho-R1-7B + Qwen3.5-9B
BioTrace	Ours	Multiscale WSI evidence	CONCH-based structured evidence at 40×/20×/10×; Patho-R1-7B perception and Qwen3.5-9B evidence reasoning

independent local services, and Qwen reasoning is run without the optional thinking mode. The pathology perceptor is invoked only for image regions selected by the evidence-acquisition policy rather than exhaustively across the WSI.

Table 5 summarizes the main inference parameters used by BioTrace. The evidence-acquisition budget and redundancy-control parameters are kept fixed across the downstream evaluation.

The agent explicitly separates three sources of information: structured WSI predictions, directly observed H&E morphology, and supportive biological or retrieved knowledge. Pathway and gene outputs are continuous WSI-derived predictions and are not interpreted as measured RNA, IHC, FISH, or other molecular assays. Accordingly, pathway- and gene-level observations provide supportive context for reasoning rather than direct molecular confirmation.

Table 5: Main inference configuration of BioTrace used in the downstream VQA experiments.

Component	Setting
Visual/text encoder	CONCH
Available magnifications	$10\times$ / $20\times$ / $40\times$
Initial structured evidence	$10\times$
Adaptive visual refinement	$10\times \rightarrow 20\times \rightarrow 40\times$
Pathology perceptor	Patho-R1-7B
Language reasoner	Qwen3.5-9B
Qwen thinking mode	Disabled
Evidence acquisition	
Top pathway candidates	3
Genes per pathway	2
Top patches per evidence source	4
Question-relevant patches per slide	4
Maximum evidence groups	4
Same-scale IoU threshold	0.50
Feature cosine threshold	0.95
Global bypass candidates per magnification	2
Agent mode	Hierarchical evidence acquisition
Pathway/Gene inspection	$40\times$
Pathway/Gene evidence	Supportive WSI-derived evidence

A.4 Agent Reproduction Protocol

The inference procedure is organized around three operations: *evidence requirement*, *evidence acquisition*, and *sufficiency verification*. For each question, the agent first determines the requested target and constructs the initial evidence state from cached structured predictions. The Evidence Verifier then determines whether the current evidence is sufficient for answering or whether additional observations are required.

At Round 0, the agent first examines compact structured phenotype evidence. When further evidence is required, it adaptively acquires visual observations from the WSI. The default spatial hierarchy follows a coarse-to-fine trajectory: $10\times$ evidence may be refined to a spatially corresponding $20\times$ or $40\times$ region, and a $20\times$ observation may subsequently be refined at $40\times$. This spatial linkage prevents evidence collected at different magnifications from becoming disconnected from the same underlying tissue region.

At the highest magnification, the agent additionally inspect pathway-level evidence associated with the current phenotype target and, when necessary, the corresponding gene-level evidence. Pathway and gene predictions are treated as supportive evidence and are not allowed to replace directly observed morphology or measured clinical assays. The evidence hierarchy therefore prioritizes direct phenotype evidence and target-specific visual morphology over pathway- and gene-level support.

Each newly acquired observation is written into typed working memory together with its source, magnification, relevance, and reliability context. After every acquisition stage, the Evidence Verifier re-evaluates the current state and assigns one of five states: *sufficient*, *partial*, *conflicting*, *insufficient*, or *unavailable*. If the evidence is sufficient, the agent proceeds to final answer generation; otherwise, it selects the next evidence source allowed by the hierarchy. The agent terminates when sufficient evidence has been collected, when the requested information is judged unavailable from the accessible evidence, or when the available acquisition budget has been exhausted.

Finally, Qwen3.5-9B produces the answer from the accumulated evidence memory. Structured predictions, Patho-R1 visual observations, and supportive pathway/gene evidence remain explicitly distinguishable in the final context, allowing the answer to be traced back to its supporting evidence. For every evaluated question, the implementation stores the structured predictions, selected

Table 6: Summary of structured biological evidence prediction. All values are multiplied by 100. Higher is better for phenotype and gene prediction, while lower is better for pathway MAE.

Method	Phenotype Mean ↑	Pathway MAE ↓	Gene Pearson ↑
FEATHER	66.5	96.0	23.4
GigaPath	68.6	72.5	45.8
PRISM	66.2	74.0	43.9
CHIEF	66.9	70.1	48.2
TITAN	<u>69.5</u>	70.0	49.4
CARE	69.3	**68.8**	<u>52.6</u>
BioTrace	**70.6**	70.2	**54.0**

regions, Patho-R1 observations, verifier decisions, and final answer, enabling offline inspection of the complete reasoning trajectory.

B Structured Biological Evidence

This section provides additional details of the biological structure underlying the evidence space, including the fixed Gene–Pathway prior and the learned Pathway–Phenotype relations.

B.1 Gene–Pathway Prior and Pathway Supervision

The Gene–Pathway relations in our structured evidence module are guided by biological prior knowledge (Liberzon et al., 2015), providing a biologically constrained organization of gene-level evidence into pathways. The corresponding relation strengths are further optimized during training. Figure 7 summarizes the pathways used in our model, their functional groups, representative member genes, and the number of genes associated with each pathway.

Pathway-level representations are explicitly supervised using patient-specific RNA-derived ssGSEA (Barbie et al., 2009) enrichment scores during training. These scores are computed independently for each patient and standardized using statistics estimated from the training cohort. At inference time, RNA-seq is not required; the model predicts pathway-level evidence directly from H&E WSIs. Thus, biological priors provide structured guidance for Gene–Pathway aggregation, while pathway representations are further adapted through molecular supervision.

B.2 Learned Pathway–Phenotype Relations

Gene–Pathway relations are initialized and constrained by biological prior knowledge, while their relation strengths are further optimized during training. Pathway–Phenotype relations are likewise learned to adapt pathway-level evidence to different phenotype targets. Figure 8 visualizes the learned effective Pathway–Phenotype weights at $40\times$ magnification. Larger weights indicate stronger learned contributions of a pathway to the corresponding phenotype representation. The resulting relations exhibit distinct patterns across morphological and clinical phenotypes.

C Detailed Results of Structured Biological Evidence

C.1 Summary of Structured Biological Prediction

We further report detailed results for phenotype-, pathway-, and gene-level prediction. All values are multiplied by 100. Higher values indicate better performance except for pathway prediction, which is evaluated by MAE and therefore favors lower values.

C.2 Phenotype-Level Results

Table 7 reports the detailed phenotype-level results. BioTrace achieves the highest average performance across the evaluated phenotypes. The improvements are particularly evident for histologi-

Functional group	Pathway	Representative member genes	n
Hormone RTK	ESR signaling	ESR1 ESR2 NCOA1 NCOA3 FOXA1 GATA3	17
	Nuclear receptor transcription	ESR1 ESR2 NCOA1 NCOA3 PGR RXRA	13
	ERBB2 signaling	ERBB2 ERBB3 ERBB4 EGFR GRB7 HSP90AA1	15
	EGFR signaling	EGFR EGF TGFA GRB2 SHC1 KRAS	17
	PI3K-AKT signaling	PIK3CA PIK3R1 AKT1 PTEN MTOR PDPK1	17
Cell cycle DNA	G2/M transition	CCNB1 CDK1 CDC25C PLK1 AURKA WEE1	14
	Mitotic checkpoint	PLK1 AURKB CDC20 BUB1 BUB1B TTK	12
	p53 DNA-damage response	TP53 ATM CHEK2 MDM2 CDKN1A BAX	11
	Double-strand break repair	ATM BRCA1 BRCA2 RAD51 PALB2 PRKDC	15
Adhesion ECM	Adherens junction	CDH1 CTNNA1 CTNNB1 CTNND1 JUP PVRL1	11
	Tight junction	TJP1 OCLN CLDN1 CLDN3 CLDN4 EPCAM	12
	TGF-beta signaling	TGFB1 TGFBR1 TGFBR2 SMAD2 SMAD3 SMAD4	14
	ECM degradation	MMP1 MMP2 MMP7 MMP9 MMP14 ADAMTS4	17
	Integrin interactions	ITGA1 ITGA2 ITGA5 ITGA6 ITGAV ITGB1	24
	Collagen formation	COL1A1 COL1A2 COL3A1 COL4A1 P4HA1 LOX	14
Immune	TCR signaling	CD3D CD3E LCK ZAP70 LAT ITK	15
	BCR signaling	MS4A1 CD79A CD79B SYK BTK CD19	15
	IFN-gamma signaling	IFNG IFNGR1 JAK1 JAK2 STAT1 GBP1	13
	PD-1 co-inhibition	PDCD1 CD274 PDCD1LG2 PTPN6 PTPN11 ZAP70	11
Stress death	Hypoxia response	HIF1A EPAS1 VHL CA9 SLC2A1 LDHA	14
	RIPK1-regulated necrosis	RIPK1 RIPK3 MLKL CASP8 TNF FADD	10
	BMP signaling	BMP2 BMP4 BMP7 BMPR1A BMPR2 SMAD1	15

Figure 7: Gene–Pathway organization used in the structured evidence module. Pathways are organized into functional groups, including hormone/RTK signaling, cell-cycle/DNA processes, adhesion/ECM, immune signaling, and stress/death responses. Gene–Pathway associations are defined with biological prior knowledge. Representative member genes are shown for each pathway.

cal type, LCIS, histologic grade, mitotic score, PR status, and HER2 status, while the model also matches the best result on ER status. These results indicate that the structured evidence module captures phenotype information across both morphological and molecularly associated endpoints.

C.3 Pathway-Level Results

Table 8 presents the detailed pathway-level results. BioTrace remains competitive across diverse signaling and regulatory pathways, achieving the best results on several pathways, including EGFR signaling, PIP3–AKT signaling, tight-junction interactions, and interferon-gamma signaling. CARE obtains the lowest average MAE, which may partly reflect its molecular-guided pretraining that aligns pathology representations with RNA and protein profiles. Despite this advantage, BioTrace achieves comparable overall pathway prediction while explicitly organizing these molecular signals within the proposed Gene–Pathway–Phenotype hierarchy.

C.4 Gene-Level Results

Table 9 further evaluates the fine-grained molecular information preserved by the structured evidence module. BioTrace achieves the highest average Pearson correlation of 54.0 and obtains the best performance on a broad range of genes spanning signaling, immune, cell-cycle, and extracellular-matrix processes. The consistent gene-level performance complements the pathway-

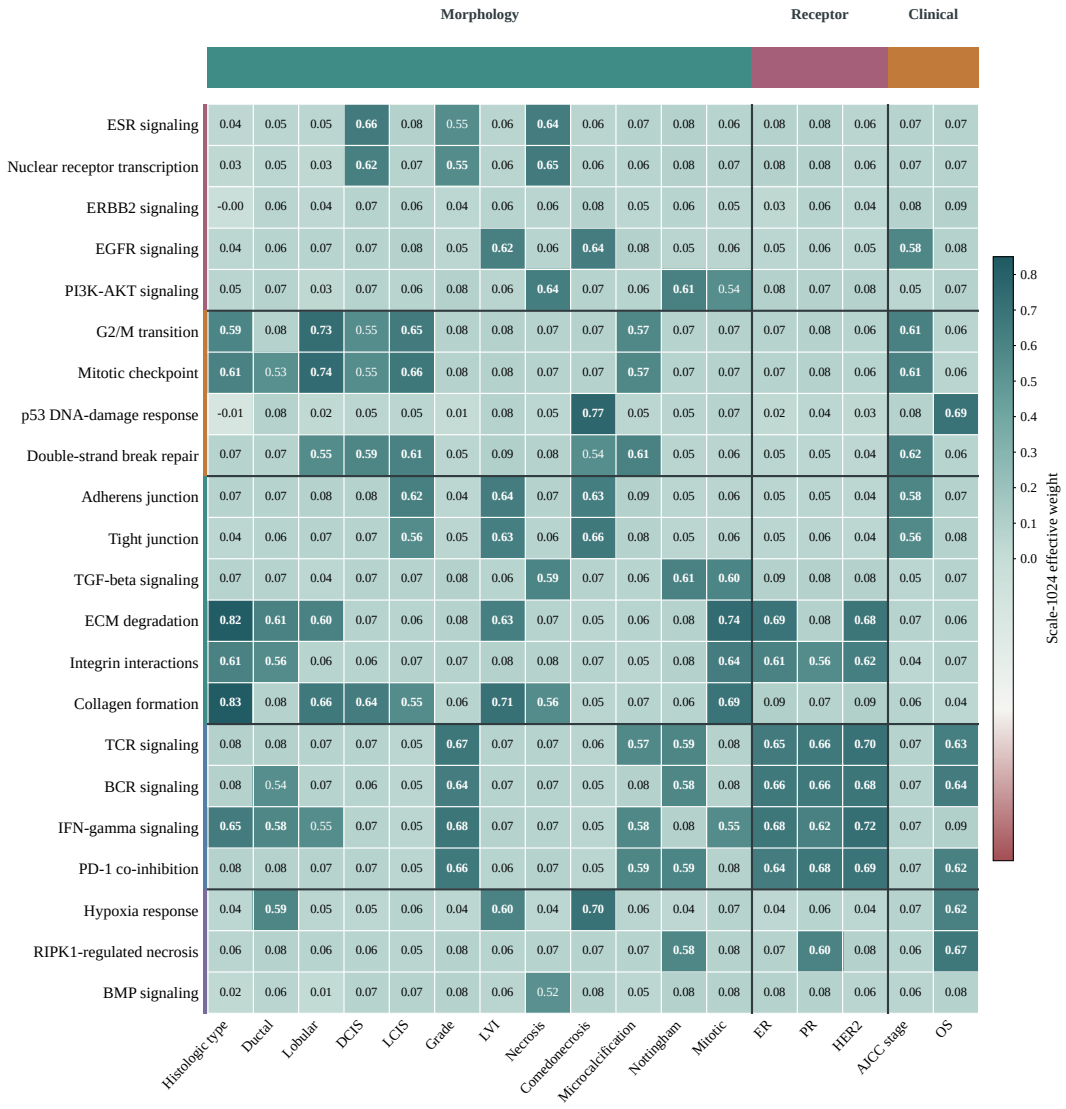

Figure 8: Learned Pathway–Phenotype relations at 40× magnification. Rows correspond to pathway-level representations and columns denote phenotype targets grouped into morphology, receptor, and clinical categories. Color intensity represents the effective learned weight between each pathway and phenotype. Unlike the fixed Gene–Pathway memberships, Pathway–Phenotype relations are optimized during training.

and phenotype-level results, suggesting that the proposed hierarchy can retain fine-grained molecular signals while organizing them into higher-level biological evidence for downstream reasoning.

Table 7: Detailed phenotype prediction results. All values are multiplied by 100 and higher is better. Boldface and underline denote the best and second-best results, respectively.

Task	Metric	FEATHER	GigaPath	PRISM	CHIEF	TITAN	CARE	BioTrace
Histological type	ACC	74.7	70.1	<u>83.3</u>	74.7	80.5	77.0	**86.2**
Ductal component	AUC	74.0	<u>84.4</u>	80.3	80.3	**86.9**	80.8	79.2
Lobular component	AUC	79.3	83.4	79.7	80.9	**84.6**	80.6	<u>84.5</u>
DCIS	AUC	78.2	82.6	79.6	<u>83.0</u>	**84.8**	81.2	78.3
LCIS	AUC	77.3	76.4	81.5	80.6	81.5	<u>83.7</u>	**85.5**
Histologic grade	ACC	65.4	63.0	56.7	60.6	64.6	<u>67.7</u>	**69.3**
Lymphovascular invasion	ACC	<u>66.0</u>	63.8	**68.1**	59.6	63.8	61.7	<u>66.0</u>
Necrosis	AUC	71.7	**79.1**	74.6	75.4	<u>78.3</u>	72.8	65.7
Comedonecrosis	AUC	77.5	**79.0**	63.8	74.3	75.5	<u>78.3</u>	70.9
Microcalcification	AUC	71.5	**79.6**	71.4	74.8	<u>78.9</u>	76.0	73.4
Nottingham score	ACC	65.7	**74.3**	62.9	71.4	60.0	71.4	<u>72.9</u>
Mitotic score	ACC	<u>58.3</u>	<u>58.3</u>	50.0	<u>58.3</u>	50.0	55.6	**63.9**
ER status	ACC	<u>83.7</u>	77.9	79.1	81.4	**88.4**	**88.4**	**88.4**
PR status	ACC	71.1	71.7	70.4	70.4	73.7	<u>76.3</u>	**78.9**
HER2 status	ACC	34.3	34.3	24.5	26.5	37.3	<u>40.2</u>	**49.0**
AJCC pathologic stage	ACC	23.4	28.1	**34.5**	30.4	26.9	26.3	<u>33.3</u>
OS	C-index	58.4	60.0	<u>64.2</u>	54.9	**66.3**	59.6	55.1
Mean	-	66.5	68.6	66.2	66.9	<u>69.5</u>	69.3	**70.6**

Table 8: Detailed pathway prediction results. All values are MAE $\times 100$ and lower is better. Boldface and underline denote the best and second-best results, respectively.

Pathway	FEATHER	GigaPath	PRISM	CHIEF	TITAN	CARE	BioTrace
ESR-mediated signaling	77.8	59.5	61.9	<u>59.1</u>	**55.5**	**55.5**	59.3
Nuclear Receptor transcription pathway	85.8	62.8	67.4	64.2	<u>62.2</u>	**59.9**	62.7
Signaling by ERBB2	114.1	**68.6**	<u>70.3</u>	**68.6**	71.1	**68.6**	70.9
Signaling by EGFR	91.4	79.8	76.5	77.3	77.4	<u>75.4</u>	**73.4**
PIP3 activates AKT signaling	113.9	72.0	72.1	<u>67.9</u>	72.7	68.7	**65.2**
G2/M Transition	87.3	67.6	67.5	65.1	63.6	**60.8**	<u>61.7</u>
Mitotic Spindle Checkpoint	84.8	66.1	66.6	64.7	61.3	**58.9**	<u>59.2</u>
p53-Dependent G1 DNA Damage Response	100.4	83.4	83.5	81.0	<u>76.4</u>	**76.0**	82.9
DNA Double-Strand Break Repair	116.1	<u>84.5</u>	88.7	84.7	85.2	**81.6**	86.7
Adherens junctions interactions	100.0	69.9	71.3	**65.7**	68.3	69.7	<u>67.8</u>
Tight junction interactions	82.0	79.9	77.5	70.5	70.0	<u>68.4</u>	**67.5**
Signaling by TGF-beta Receptor Complex	100.8	67.8	67.7	<u>65.2</u>	70.6	**64.6**	72.0
Degradation of the extracellular matrix	87.8	75.8	75.2	70.0	<u>69.3</u>	**67.1**	70.4
Integrin cell surface interactions	102.8	<u>76.3</u>	80.6	79.9	78.2	**75.5**	77.8
Collagen formation	84.8	67.5	70.9	**64.9**	<u>65.4</u>	67.4	71.3
TCR signaling	83.0	65.2	70.1	<u>62.2</u>	**62.0**	64.8	63.9
Signaling by the B Cell Receptor (BCR)	97.3	71.3	73.2	68.2	**64.8**	67.9	<u>67.6</u>
Interferon gamma signaling	100.4	73.0	71.9	66.5	65.4	<u>63.3</u>	**62.3**
Co-inhibition by PD-1	83.6	64.9	67.7	<u>64.0</u>	**62.4**	67.0	66.3
Cellular response to hypoxia	104.4	76.9	80.4	**74.2**	<u>74.6</u>	76.9	75.5
RIPK1-mediated regulated necrosis	103.0	87.9	93.3	**85.9**	92.4	88.3	<u>86.4</u>
Signaling by BMP	110.4	73.8	74.5	73.4	<u>70.9</u>	**67.3**	73.5
Mean	96.0	72.5	74.0	70.1	<u>70.0</u>	**68.8**	70.2

Table 9: Gene-level prediction results. Values are Pearson correlation $\times 100$ and higher is better. Boldface and underline denote the best and second-best results for each gene, respectively.

Gene	FEATHER	GigaPath	PRISM	CHIEF	TITAN	CARE	BioTrace
CDC25A	40.9	63.0	63.1	64.4	66.5	**70.7**	66.7
ITK	42.5	57.3	58.4	63.9	61.2	**65.8**	64.3
CDC20	52.6	61.4	60.5	58.4	63.3	**64.9**	63.7
CXCL9	20.6	53.4	47.5	59.5	60.1	60.4	**62.4**
EGFR	9.0	47.1	57.8	49.0	51.9	59.6	**61.7**
TGFBR2	22.2	54.7	46.5	59.8	53.4	**61.7**	61.6
IDO1	32.9	54.3	50.5	54.1	55.1	55.3	**61.1**
TRAT1	34.9	54.5	53.1	60.0	58.1	**61.7**	61.1
GBP1	22.0	52.1	54.6	56.2	55.1	57.8	**61.1**
FOXO1	7.9	49.3	48.2	60.1	56.0	59.2	**61.0**
CHEK1	18.7	54.8	54.6	50.9	59.9	**61.5**	60.1
PIK3CD	37.4	49.8	53.0	50.0	**61.0**	59.9	60.0
PRKCB	25.5	52.7	50.8	56.1	59.3	**61.1**	59.9
CXCL10	10.2	51.2	51.2	53.7	54.1	54.2	**59.1**
PTPRC	31.5	46.4	44.1	54.7	55.4	**60.5**	58.8
CDC25B	4.3	53.7	49.3	53.3	54.4	**60.1**	58.4
LOX	34.8	46.5	48.1	54.2	47.4	56.4	**58.2**
RAD51	29.9	51.4	52.4	49.8	53.9	54.0	**57.6**
PDCD1LG2	22.7	55.3	51.4	**62.5**	53.5	54.8	57.5
LAT	10.5	53.1	54.3	54.0	**60.1**	55.4	57.1
PRKCQ	25.9	43.2	46.9	53.4	52.0	55.9	**57.0**
SOCS1	18.1	49.7	55.7	50.8	54.9	**57.2**	56.8
MMP14	34.4	44.5	46.4	55.2	49.5	53.7	**56.6**
MS4A1	17.4	47.6	53.8	51.2	54.5	54.6	**56.5**
CARD11	44.4	51.4	47.8	51.7	56.1	**57.9**	56.3
MMP7	18.0	43.5	44.7	49.0	45.3	54.6	**55.8**
BTK	34.0	43.6	41.6	48.2	53.7	51.6	**55.0**
PTK6	43.3	47.1	41.6	41.3	50.1	47.0	**55.0**
PIK3R1	22.9	30.4	30.6	45.8	39.7	36.2	**54.7**
LYN	0.8	47.2	48.5	51.8	**56.0**	51.0	54.5
PLCG2	6.9	47.3	46.5	49.4	52.1	53.0	**53.7**
ADAMTS5	21.9	44.3	44.5	29.8	48.2	50.5	**53.6**
HLA-DRA	29.1	41.7	38.1	50.6	52.4	**55.1**	53.6
BMP4	32.3	46.7	37.1	43.3	43.9	**55.0**	52.6
ACVR1	29.6	44.6	42.8	44.1	49.0	47.8	**52.3**
MMP19	22.4	45.5	49.5	**55.5**	47.8	49.8	51.7
CTSS	21.6	42.1	36.1	44.1	**51.6**	50.0	**51.6**
CFLAR	3.1	41.5	36.9	42.7	38.1	**51.4**	50.9
AKT3	28.9	37.2	32.0	30.4	42.4	46.0	**50.5**
TGFB1	11.9	49.8	32.8	40.3	47.6	46.2	**50.3**
PAX5	19.7	40.9	30.9	40.7	48.7	46.2	**50.1**
CDK2	10.0	37.3	44.2	39.2	44.5	47.0	**49.6**
LAMB3	24.8	40.6	34.6	41.9	43.3	**54.4**	48.7
CR2	12.9	43.4	42.3	39.5	44.2	46.4	**48.0**
PIK3CG	34.3	41.1	28.9	44.5	44.7	**49.3**	47.9
CDC25C	15.8	44.9	46.7	45.4	**48.2**	44.7	47.8
MMP3	5.6	44.5	39.2	46.0	32.7	43.2	**47.7**
ITGA5	27.3	43.4	35.5	41.8	44.1	**48.2**	47.4
ITGB8	17.5	40.8	34.8	44.5	41.5	**49.7**	46.3
PDK1	14.1	43.7	32.3	45.4	40.8	**49.2**	46.0
RUNX2	22.6	34.9	34.4	44.5	33.2	44.8	**45.5**
ESR2	18.7	32.5	33.2	34.6	43.2	38.4	**45.3**
ITGB1	27.7	36.3	34.7	30.5	33.1	**45.9**	45.0
BLNK	34.8	37.2	33.2	41.9	32.6	42.6	**43.4**
ITGA4	9.4	35.9	27.7	42.5	39.7	**48.1**	42.6
ACTN1	33.3	29.0	29.8	33.9	38.1	**46.1**	41.8

Continued on next page

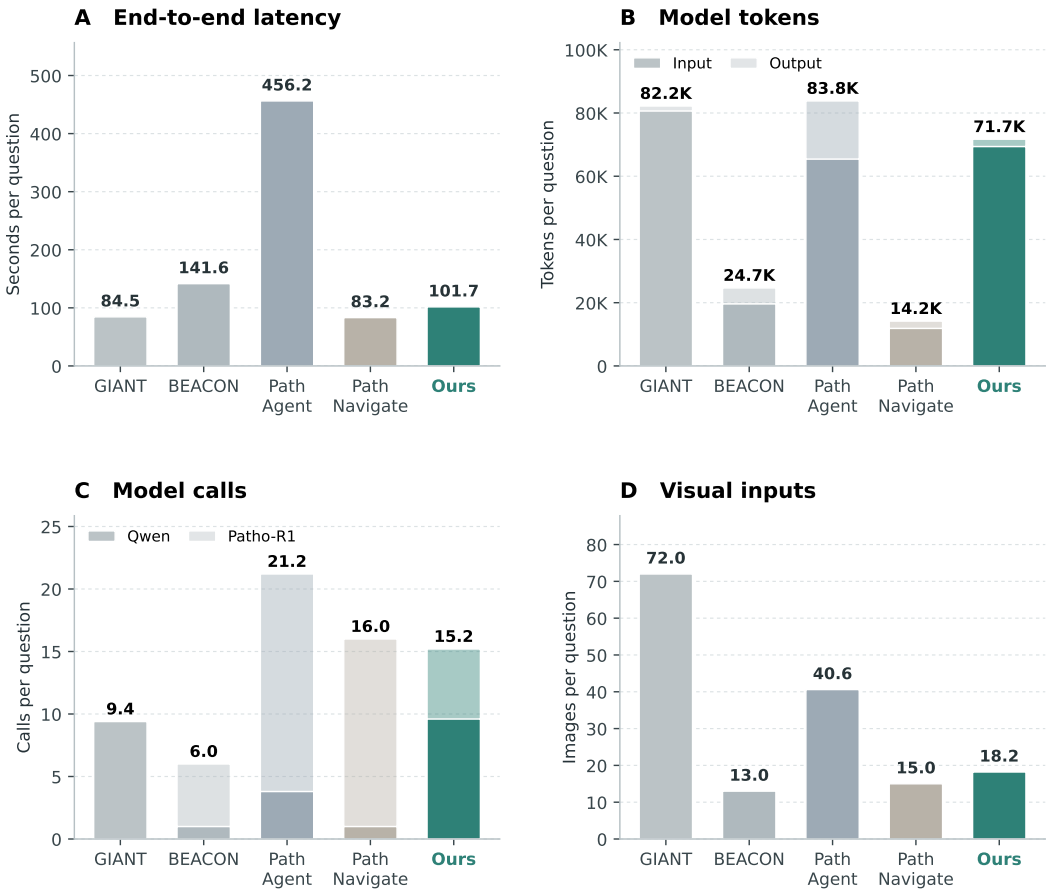

Figure 9: Efficiency comparison under a unified five-question protocol with preloaded local services. We report end-to-end latency, model tokens, model calls, and visual inputs per question; lower is better for all metrics.

Table 9: Gene-level prediction results (continued).

Gene	FEATHER	GigaPath	PRISM	CHIEF	TITAN	CARE	BioTrace
RXRA	24.4	35.5	35.7	38.8	<u>38.9</u>	33.7	**41.8**
Mean	23.4	45.8	43.9	48.2	49.4	<u>52.6</u>	**54.0**

D Additional Analysis

D.1 Efficiency Analysis

Figure 9 compares the inference cost of BioTrace with representative pathology agents under the same evaluation protocol. Despite incorporating structured biological evidence, multiscale visual acquisition, and iterative evidence verification, BioTrace maintains a moderate overall inference cost. It requires 101.7 seconds, 15.2 model calls, and 18.2 visual inputs per question, substantially reducing the corresponding costs of PathAgent, which requires 456.2 seconds, 21.2 calls, and 40.6 visual inputs. Its end-to-end latency also remains in a similar range to GIANT and PathNavigate.

The token cost of BioTrace is higher than that of the most compact agents, reflecting the additional structured and multiscale evidence passed through the reasoning process, but remains below GIANT and PathAgent. Overall, these results indicate that the additional biological reasoning and evidence-



Figure 10: Additional visualization of BioTrace. Structured evidence from phenotype, pathway, and gene levels jointly supports the final answer through complementary spatial and biological cues.

verification mechanisms do not introduce prohibitive computational overhead, yielding a balanced trade-off between reasoning capability and inference cost.

D.2 Additional visualization

Figure 10 presents additional examples of BioTrace across different pathology questions. Across these cases, BioTrace consistently combines phenotype-level predictions, pathway- and gene-level supportive evidence, and concept-specific spatial responses to form an explicit multi-level evidence trace. The selected regions further provide morphology-grounded observations that complement the structured predictions. Together, these examples illustrate how the proposed evidence space supports question-specific reasoning by integrating predictive, spatial, and biological evidence.