CAN THE EVIDENCE REACH THE ANSWER? DIAGNOSTIC REACHABILITY OVER UNCERTAIN EVIDENCE GRAPHS FOR AGENTIC WHOLE-SLIDE REASONING.

Anonymous authors
Paper under double-blind review

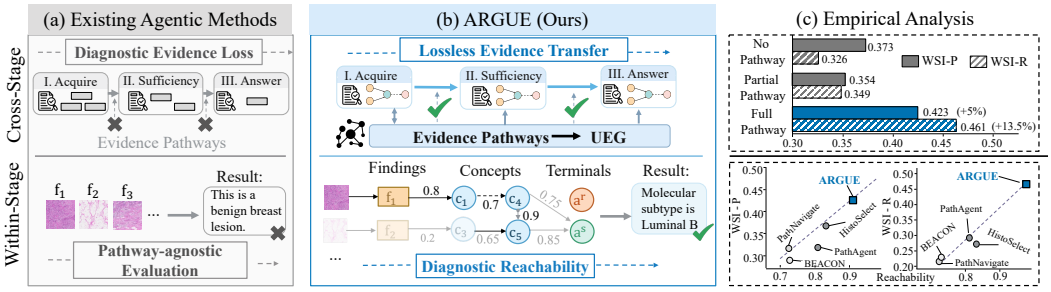

Figure 1: **Motivation and empirical analysis of diagnostic evidence pathways. (a)** Limitations of existing agentic WSI-VQA systems. **(b)** ARGUE addresses these limitations with a shared Uncertain Evidence Graph (UEG) and diagnostic reachability assessment. **(c)** The UEG with full pathway structure outperforms structurally ablated variants, and higher reachability is also associated with better performance across baselines.

ABSTRACT

Whole-slide image visual question answering (WSI-VQA) requires agentic systems to collect diagnostically relevant evidence that is often sparse and spatially dispersed across gigapixel slides. Existing agentic WSI-VQA systems typically decompose reasoning into three stages: evidence acquisition, acquisition sufficiency assessment, and answer generation, yet represent the acquired evidence as unstructured patch level information. Because they do not explicitly preserve the diagnostic evidence pathways that connect morphological observations to an answer through diagnostic concepts, they exhibit two limitations: unreliable pathway agnostic evidence evaluation within stages and diagnostic evidence loss across stages. To address these limitations, we propose ARGUE (**A**ctive **R**easoning over a **G**raph of **U**ncertain **E**vidence), a training-free agentic WSI-VQA system. Within stages, ARGUE uses diagnostic reachability to enable pathway-aware evaluation by quantifying the probability of realizing evidence pathways under uncertain diagnostic relations. Across stages, ARGUE maintains a shared Uncertain Evidence Graph (UEG) to preserve these pathways, thereby preventing the loss of structured evidence. Experiments on three benchmarks show that ARGUE achieves the best overall performance among all baselines. Ablation studies further validate the effectiveness of ARGUE's core components, while robustness analyses show consistent performance under external noise and different implementation choices.

1 INTRODUCTION

Whole-slide images (WSIs) are gigapixel digital scans used for histopathological examination, which serves as the gold standard for cancer diagnosis. In this context, whole-slide image visual question answering (WSI-VQA) aims to answer diagnostic questions grounded in WSIs (Liang et al., 2025; Lu et al., 2024; Lyu et al., 2025). Answering such questions is challenging because

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.

the system must locate and reason over diagnostic evidence that may be sparse and spatially dispersed across a gigapixel WSI, while operating under a constrained inspection budget. Existing work attempts to address this challenge using agentic systems.

Existing agentic WSI-VQA systems adopt a ***three-stages*** pipeline: selecting patches during ***evidence acquisition***, assessing whether further inspection is warranted during ***acquisition sufficiency assessment***, and integrating the evidence during ***answer generation*** (Chen et al., 2025a; Lyu et al., 2025; Weishaupt et al., 2025). Throughout this pipeline, they capture visual embeddings (Huang et al., 2026) and textual morphological observations (Chen et al., 2025a; Ghezloo et al., 2025) as evidence directly from selected patches, and pass this unstructured information to subsequent stages. However, clinical diagnostic reasoning is primarily established through structured evidence pathways Pena & de Souza Andrade-Filho (2009); Lou et al. (2025), where pathologists interpret patch level morphological observations as diagnostic concepts (Zhao et al., 2024; Jiang et al., 2026; Sun et al., 2025a) to support or refute an answer through diagnostic relations, with uncertainty in the validity of these relations. Existing methods do not explicitly preserve and evaluate these evidence pathways throughout the three stages, leading to two limitations (Figure 1).

The first limitation is ***within-stage pathway agnostic evaluation***, where existing systems (Wong et al., 2026; Chen et al., 2025a; Huang et al., 2026) evaluate evidence without tracing complete evidence pathways. For example, a patch of *irregular crowded cells* may appear suspicious, but it supports the conclusion that *the tumor is invasive* only if the tissue context shows that tumor cells have crossed their normal boundary (Bejnordi et al., 2017). If any link in this pathway is missing, treating the patch as sufficient evidence may misdirect decision making within the three stages.

The second limitation is ***cross-stage diagnostic evidence loss***, where handoffs between stages fail to preserve the evidence pathways established within a stage. In the example above, acquisition may recognize that the observed patch does not yet form a complete evidence pathway to the conclusion of *invasive tumor*. If the handoff retains only the morphological observation but omits this evidence gap, acquisition sufficiency assessment can stop prematurely, and answer generation may treat the same observation as conclusive.

To examine how these limitations relate to WSI-VQA performance, we present two complementary analyses in Figure 1(c). As a result, across recent agentic WSI-VQA baselines Yang et al. (2026); Wong et al. (2026); Chen et al. (2025a); Huang et al. (2026), lower diagnostic reachability (i.e. the probability of realizing at least one evidence pathways under uncertain diagnostic relations) is generally associated with poorer performance. A structural ablation further shows that removing diagnostic relations from the full evidence representation also harms system performance. These observations expose two fundamental **technical challenges**. First, the system must maintain a coherent and uncertainty-aware representation of complete diagnostic evidence pathways throughout the three-stages pipeline, such that their structured evidence can be preserved across stage boundaries. Second, the combinatorial explosion over uncertain diagnostic relations makes exact evaluation of pathway realizability intractable as the evidence graph grows. Yet repeated estimates of pathway realizability are still needed to guide the pipeline.

To this end, we propose **ARGUE** (**A**ctive **R**easoning over a **G**raph of **U**ncertain **E**vidence), a training-free agentic WSI-VQA framework that addresses these two challenges with two coupled components, respectively (Figure 1 (b)). First, to address ***within-stage pathway agnostic evidence evaluation***, ARGUE formalizes each evidence pathway as a directed graph path composed of uncertain diagnostic relations, connecting a morphological observation to an answer through diagnostic concepts. Based on this formulation, we further introduce ***diagnostic reachability***, which quantifies the probability that at least one complete evidence pathway is realized under uncertain diagnostic relations. This pathway level criterion enables ARGUE's three-stages decisions to depend on whether morphological observations participate in complete evidence pathways rather than on local salience alone. Since exact computation of the diagnostic reachability is #P-hard in general, we estimate it via Monte Carlo sampling of UEG realizations. We further derive a theoretical bound on the estimation error that accounts for both Monte Carlo sampling error and bounded errors in the estimated relation-validity probabilities.

Second, to address ***cross-stage diagnostic evidence loss***, ARGUE merges these directed evidence pathways through shared nodes and relations into an ***Uncertain Evidence Graph*** (**UEG**), which serves as a shared evidence state across the three stages. The graph is continuously updated

and reused across evidence acquisition, acquisition sufficiency assessment, and answer generation, thereby preserving the evidence pathway that explains how the evidence supports or refutes the answer. This shared structured evidence state strengthens cross-stage evidential consistency and prevents structured evidence from being lost during stages handoffs.

Our contributions are threefold:

- We propose ARGUE, an agentic WSI-VQA framework that preserves uncertain diagnostic relations across all three stages through a shared Uncertain Evidence Graph.
- We introduce diagnostic reachability for pathway aware evidence evaluation, with Monte Carlo estimation and a theoretical bound on the estimation error.
- Extensive evaluations across three benchmarks demonstrate ARGUE's leading performance, core-component effectiveness, and robustness to external perturbations and implementation choices.

2 Related Work

Whole slide vision language models. Early WSI-VQA works formulated whole slide image interpretation as generative visual question answering, aligning natural language questions with patch-level representations to produce slide level answers (Chen et al., 2024). Subsequent models such as SlideChat and WSI-LLaVA improved WSI encoding and pathology instruction tuning, while Quilt-LLaVA (Seyfioglu et al., 2024) extended multimodal pathology modeling to broader diagnostic and explanatory tasks (Chen et al., 2025c; Liang et al., 2025). These models primarily construct a visual context before answer generation while do not explicitly control sequential evidence acquisition or maintain a structured diagnostic state during inference.

Agentic WSI-VQA systems. Recent advanced WSI-VQA systems increasingly adopt agentic workflows that mimic the coarse to fine examination process of pathologists. These systems iteratively decide which regions to inspect, interpret the acquired observations, determine whether further inspection is needed, and synthesize a final diagnostic response (Ghezloo et al., 2025; Weishaupt et al., 2025). This workflow therefore naturally decomposes agentic WSI-VQA into three stages: evidence acquisition, acquisition sufficiency assessment, and answer generation.

Most recent progress has focus on the ***evidence acquisition***: how to acquire diagnostically useful regions under a limited inspection budget. PathFinder (Ghezloo et al., 2025) and CPathAgent (Sun et al., 2025b) model multi scale examination through explicit navigation and regional interpretation, while GIANT (Buckley et al., 2025) enables multimodal models to autonomously navigate WSIs through iterative multi scale cropping. Pathology-CoT (Wang et al., 2026) further learns such examination behavior from expert viewing trajectories. PathFound (Hua et al., 2026) uses an evolving diagnostic hypothesis to guide subsequent evidence acquisition. BEACON (Wong et al., 2026) selects patches according to expected information gain, and AdaptivePath (Chen et al., 2026) uses abnormality driven navigation. For the ***acquisition sufficiency assessment*** stage, existing methods typically adopt simple heuristics to determine whether further inspection should continue. A common strategy is to terminate exploration after a predefined number of regions or interaction steps Buckley et al. (2025); Ghezloo et al. (2025); Chen et al. (2026). Other methods rely on LLM-based evaluator to decide whether further inspection is necessary Sun et al. (2025b); Wong et al. (2026). Finally, in ***answer generation*** stage, existing WSI agents typically aggregate the collected visual evidence using an LLM or multimodal reasoner. Some methods directly synthesize an answer from the retrieved regions, while others introduce multi agent verification, cross scale aggregation, or an additional arbiter to reconcile intermediate predictions Lyu et al. (2025); Chen et al. (2026; 2025a).

Similar to ARGUE, several very recent works (Lou et al., 2025; Zhao et al., 2026; Jiang et al., 2026) also highlight the importance of structured evidence representations through concept indexed evidence banks or graph based evidence representation. However, these efforts primarily focus on organizing accumulated evidence or supervising diagnostic reasoning, without jointly addressing how uncertainty aware diagnostic pathways can be preserved across different stages, or how pathway awared evidence can be evaluated during inference. ARGUE fills this gap by maintaining an evolving Uncertain Evidence Graph as a shared cross stage evidence state and introducing diagnostic reachability to guide pathway aware decisions throughout the entire WSI-VQA pipeline. Further discussion of related works are provided in Appendix A.8.

3 METHOD

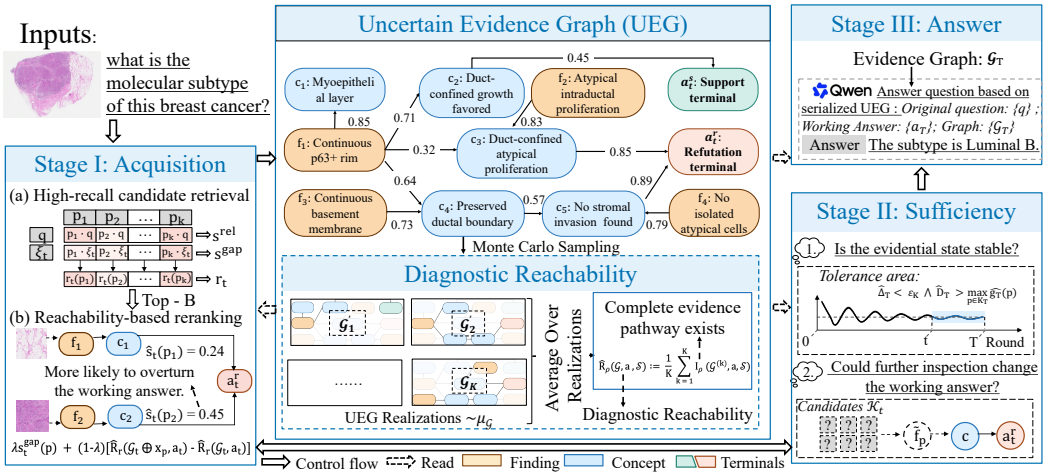

Figure 2: The overview of our proposed ARGUE.

As illustrated in Figure 2, ARGUE maintains an Uncertain Evidence Graph (UEG) that links morphological observations to the working answer through uncertain diagnostic relations (Section 3.1). It estimates diagnostic reachability via Monte Carlo sampling to assess complete evidence pathways (Section 3.2). Using this shared graph and reachability estimates, ARGUE iteratively acquires patches to address evidence gaps and challenge the working answer, updates the UEG and working answer, and assesses acquisition sufficiency. Once the sufficiency criterion is met, it generates the final answer from the evidence pathway preserved in the UEG (Section 3.3).

3.1 UNCERTAIN EVIDENCE GRAPH FOR CROSS-STAGE EVIDENCE PRESERVATION

To prevent cross-stage diagnostic evidence loss, we use a shared Uncertain Evidence Graph (UEG) to preserve evidence pathways and their diagnostic context across ARGUE's three stages: evidence acquisition, acquisition sufficiency assessment, and answer generation. Each pathway encodes evidence about how a morphological finding supports or refutes the working answer through diagnostic concepts with uncertain relations. Specifically, inspired by the probabilistic modeling of relations in uncertain graphs Ke et al. (2019), we formulate the UEG to jointly represent evidence pathways and uncertainty in their constituent relations:

Definition 1 (UEG and Evidence pathway). Given a question q and working answer a_t at round t, the UEG is defined as an uncertain directed graph:

$$\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, \boldsymbol{\pi}_t), \quad \mathcal{V}_t = \mathcal{F}_t \dot{\cup} \mathcal{C}_t \dot{\cup} \{a_t^s, a_t^r\}$$

Here, $\mathcal{F}_t$ and $\mathcal{C}_t$ contain atomic morphological finding nodes and semantic concepts nodes, respectively. a_t^s and a_t^r are terminal nodes represent support for and refutation of the working answer a_t, respectively. Each edge $e \in \mathcal{E}_t$ represents an uncertain diagnostic relation with validity probability $\pi_t(e) \in [0, 1]$. An evidence pathway is a directed path in $\mathcal{G}_t$ from a finding node to a terminal node through one or more concept nodes:

$$P := \left(f \to c_1 \to \cdots \to c_L \to a_t^\rho\right), \qquad L \in \mathbb{Z}_{\geq 1}, \quad f \in \mathcal{F}_t, \quad c_\ell \in \mathcal{C}_t \quad \text{for} \quad \ell = 1, \ldots, L,$$

where the notation $\to$ denotes a directed edge $(u, v) \in \mathcal{E}_t$.

UEG Construction and Updating. We initialize $\mathcal{C}_0$ with question-conditioned morphological concept. For each inspected patch, we decompose the visual observer's description into atomic findings and added to $\mathcal{F}_t$. An LLM then proposes sparse relations linking findings to compatible existing or new concepts, and concrete concepts to broader parents. It then proposes edges to a_t^s or a_t^r when a concept node specifically supports or contradicts a_t, respectively, and omits terminal edges otherwise. Following prior work in general medical reasoning (Chen et al., 2025b; Rezaei et al., 2025) and pathology (Lyu et al., 2025; Wong et al., 2026; Vorontsov et al., 2026; Liang et al., 2025), for each new edge e, we estimate $\pi_t(e)$ using a two-way softmax over the LLM's *Yes*/*No* validity logits. When the working answer changes, we replace its terminals and recompute concept-to-terminal relations while retaining applicable findings and concepts. Construction and probability estimation details are provided in Appendix A.9.

3.2 Diagnostic Reachability for Within-Stage Pathway Aware Evaluation

Existing **within-stage pathway agnostic evidence evaluation** lacks an explicit model of the reasoning process that links evidence to the answer. The UEG provides a probabilistic formulation in which each edge is assigned an estimated relation-validity probability, and each realized pathway constitutes a complete evidence chain supporting or refuting the working answer. This reformulates otherwise implicit reasoning as reachability estimation over the UEG, with the reachability defined as the probability that at least one complete evidence pathway is realized.

Definition 2 (Diagnostic reachability). For a UEG $\mathcal{G}_t$ with finding nodes $\mathcal{F}_t$, let $Z_t = (Z_e)_{e\in\mathcal{E}_t} \sim \mu_{\mathcal{G}_t}$ be a joint binary relation-validity assignment induced by π_t, where $Z_{t,e} = 1$ indicates a valid relation. Then the diagnostic reachability is defined as:

$$R_\rho(\mathcal{G}_t, a_t; \mathcal{S}) := \Pr_{Z_t\sim\mu_{\mathcal{G}_t}} \left(S \rightsquigarrow_{\mathcal{G}_{t,Z_t}} a_t^\rho\right), \text{with} \quad \mathcal{S} \subseteq \mathcal{F}_t, \quad \rho \in \{s, r\} \tag{1}$$

where $\mathcal{G}_{t,Z_t} := (\mathcal{V}_t, \{e \in \mathcal{E}_t : Z_{t,e} = 1\})$ is the deterministic graph induced by the assignment of the variables $Z_{t,e}$ and is referred to as a *UEG realization* (see Figure 2). $\mathcal{S} \rightsquigarrow_{\mathcal{G}_{t,Z_t}} a^\rho$ means that there exists some $P \in \mathcal{P}_\rho(\mathcal{G}_t, a_t)$ starts at a finding node in $\mathcal{S}$ and satisfies $\mathcal{E}(P) \subseteq \mathcal{E}(\mathcal{G}_{t,Z_t})$, where $P_\rho(\mathcal{G}_t, a_t)$ is the set of all evidence pathways leading to a_t.

Exact computation of diagnostic reachability is intractable. Unless otherwise specified, we set $\mathcal{S} = \mathcal{F}_t$ and write $R_\rho(\mathcal{G}_t, a_t) := R_\rho(\mathcal{G}_t, a_t; \mathcal{F}_t)$. Evaluating this probability by exhaustive enumeration involves up to 2^M UEG realizations for M uncertain relations, making it impractical as the graph grows.

Proposition 1. For each $\rho \in \{s, r\}$, exact evaluation of $R_\rho(\mathcal{G}_t, a_t; \mathcal{S})$ is #P-hard when $|\mathcal{S}| = 1$ and relation validities are independent. The proof is given in Appendix A.5.

Diagnostic reachability estimation. Given the huge complexity, we estimate diagnostic reachability by Monte Carlo sampling on *UEG realizations*. For a UEG realization $\mathcal{G}'_t = (\mathcal{V}_t, \mathcal{E}'_t)$, let $z_e = \mathbf{1}\{e \in \mathcal{E}'_t\}$. Because relation validities are modeled as mutually independent Bernoulli variables, the probability of the realization is: $\mu_{\mathcal{G}_t}(z) = \prod_{e\in\mathcal{E}_t} p_{t,e}^{z_e}(1-p_{t,e})^{1-z_e}$, where $p_{t,e} = \pi_t(e)$ is the validity probability of edge e. Each relation is sampled once per UEG realization, with its state shared across all evidence pathways containing it. Therefore, reachability reduces to a deterministic path existence query, so Definition 2 can be estimated using K i.i.d. draws $\mathcal{G}^{(1)}, \ldots, \mathcal{G}^{(K)} \sim \mu_{G_t}$, which is an unbiased estimation for $R_\rho(\mathcal{G}_t, a_t; \mathcal{S})$ with a total cost of $O(K(|\mathcal{V}_t| + |\mathcal{E}_t|))$:

$$\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S}) := \frac{1}{K}\sum_{k=1}^{K} I_\rho(\mathcal{G}_t^{(k)}, a_t; S). \qquad \text{where} \quad I_\rho(\mathcal{G}'_t, a_t; \mathcal{S}) \coloneqq \mathbf{1}\left\{\mathcal{S} \rightsquigarrow_{\mathcal{G}'_t} a_t^\rho\right\}. \tag{2}$$

Bounding Monte Carlo sampling and LLM probability estimation deviations. The reachability estimation in Equation 2 introduces two types of deviation: the finite sample Monte Carlo deviation and the deviation in the relation-validity probability estimation. Let $R^\star_\rho(\mathcal{G}_t, a_t; \mathcal{S})$ denote reachability obtained by replacing the estimated probabilities with their reference counterparts, we quantify their combined effect on diagnostic reachability and establish a unified bound:

Proposition 2 (Combined deviation bound for diagnostic reachability). Fix a UEG $\mathcal{G}_t$, a working answer a_t, a source set $\mathcal{S} \subseteq \mathcal{F}_t$, and a direction $\rho \in \{s, r\}$ before Monte Carlo sampling. Suppose that the estimated validity probability p_j of the j-th uncertain relation differs from its reference value $p_j^\star$ by at most $\varepsilon_j^{\mathrm{LLM}} \in [0, 1]$. Let $\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S})$ be the reachability estimation obtained from K i.i.d. UEG realizations from $\mu_{\mathcal{G}_t}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have:

$$\left|\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R^\star_\rho(\mathcal{G}_t, a_t; \mathcal{S})\right| \le \sqrt{\frac{\log(2/\delta)}{2K}} + \beta_{\mathcal{G}_t} \tag{3}$$

where $\beta_{\mathcal{G}_t} := 1 - \prod_{j=1}^{M}(1 - \varepsilon_j^{\mathrm{LLM}})$, and M is the number of uncertain relations in $\mathcal{G}_t$. The first term decreases with the sample count K, whereas the second captures the effect of the assumed relation probability errors. The proof is provided in Appendix A.6.

3.3 ARGUE's three-stage specific strategies

Stage I: Evidence Gap Aware Evidence Acquisition. ARGUE formulates *evidence acquisition* as a prospective graph extension problem, seeking patches that address unresolved evidence needs and challenge the working answer. To do so, we evaluate candidate extensions using two complementary signals: *evidence gap relevance* measures how well a patch matches the evidence currently

needed, while *predicted gains in refutation reachability* assess its potential to strengthen a complete evidence pathway against the working answer. Candidate selection proceeds through high recall retrieval followed by reranking with these two signals.

Step 1: high recall candidate retrieval. At round t, let $\mathcal{U}_t$ be the set of uninspected patches. Given $(q, a_t, \mathcal{G}_t)$, a structured LLM call (prompt is given in Appendix A.10) produces a set of evidence gap descriptions $\xi_{t,1}, \cdots, \xi_{t,M_t}$, where each description specifies an observable cue worth investigating to verify uncertain judgments or supply missing details. For each $x_p \in \mathcal{U}_t$, we compute:

$$s^{\mathrm{rel}}(p) = \cos(\phi_{\mathrm{text}}(q), \phi_{\mathrm{img}}(x_p)), \qquad s_t^{\mathrm{gap}}(p) = \max_{m \in \{1,\cdots,M_t\}} \cos(\phi_{\mathrm{text}}(\xi_{t,m}), \phi_{\mathrm{img}}(x_p)). \quad (4)$$

where ϕ_{text} and ϕ_{img} are frozen text and image encoders, respectively. We retain the stronger retrieval signal: $r_t(p) := \max\{s^{\mathrm{rel}}(p), s_t^{\mathrm{gap}}(p)\}$and keep up to B top ranked patches in $\mathcal{K}_t \subseteq \mathcal{U}_t$. Doing so prevents exploration from being anchored to an initially incorrect working hypothesis, allowing discriminative evidence for alternative diagnoses not yet represented in the current graph to enter the candidate pool.

Step 2: diagnostic reachability reranking. For each $x_p \in \mathcal{K}_t$, we treat the evidence gap description maximizing Equation 4 as a hypothetical finding. Integrating it into $\mathcal{G}_t$ using the same relation construction procedure yields $\mathcal{G}_t \oplus x_p$ (Appendix A.9). The gain of refutation pathway is therefore: $\widehat{g}_t(p) := \widehat{R}_r(\mathcal{G}_t \oplus x_p, a_t) - \widehat{R}_r(\mathcal{G}_t, a_t)$. Using the estimator in Section 3.2, we compute its estimated gain in refutation reachability: $\widehat{s}_t(p) := \lambda s_t^{\mathrm{gap}}(p) + (1-\lambda)\widehat{g}_t(p)$. The first term measures how well a candidate patch matches the current evidence gap. The second estimates how much confirming that gap would increase the probability of realizing a complete evidence pathway against the working answer a_t. With λ balancing these signals, $\widehat{s}_t(p)$ prioritizes patches that address unresolved evidence needs and have the potential to challenge the working answer through evidence pathways.

Step 3: Working answer update. After selection, the visual observer inspects patches and generates textual morphological observation, which is then incorporated into the UEG to obtain $\mathcal{G}_{t+1}$. It then evaluates whether the updated graph still favors the working answer a_t using $\widehat{D}(\mathcal{G}_t, a_t) := \widehat{R}_s(\mathcal{G}_t, a_t) - \widehat{R}_r(\mathcal{G}_t, a_t)$.When $\widehat{D}(\mathcal{G}_{t+1}, a_t) > 0$, the support reachability remains greater than the refute reachability. The overall UEG therefore continues to favor the working answer, and ARGUE leaves the working answer unchanged: $a_{t+1} = a_t$. Otherwise, the updated UEG no longer favors a_t. ARGUE then produce an updated answer a_{t+1} based on q and $\mathcal{G}_{t+1}$. The two terminals a_{t+1}^{ρ} are also updated accordingly. Previously morphological observations and concepts that remain applicable are retained when the answer changes. Details are provided in Appendix A.9.

Stage II: Evidence Acquisition Sufficiency Assessment. ARGUE stops the acquisition loop when two signals agree: first, the evidential state encoded by the UEG must be operationally stable across successive rounds, second, there should be no evaluated candidate in $\mathcal{K}_t$ to appreciably strengthen a refutation pathway. At each post update state $t \geq 2$ with $\mathcal{U}_t \neq \varnothing$, define $\widehat{\Delta}_t := \max_{\rho \in \{s,r\}} \left| \widehat{R}_\rho(\mathcal{G}_t, a_t) - \widehat{R}_\rho(\mathcal{G}_{t-1}, a_{t-1}) \right|$. ARGUE applies the joint criterion:

$$\max\left\{\widehat{\Delta}_t, \max_{p \in \mathcal{K}_t} \widehat{g}_t(p)\right\} \leq \varepsilon_K \quad \wedge \quad \widehat{D}_t > \max_{p \in \mathcal{K}_t} \widehat{g}_t(p), \quad \varepsilon_K := \frac{2}{\sqrt{K}} \quad (5)$$

where K is the number of UEG realizations. Both terms of reachability changes are computed using the same Monte Carlo samples to reduce sampling noise. Furthermore, ARGUE additionally imposes a maximum number of acquisition rounds to prevent noisy or slowly varying signals from prolonging acquisition indefinitely. This boundary is reached once five evidence acquisition round is completed. Acquisition terminates as soon as either the joint criterion in Equation 5 or this finite exploration boundary is reached. Appendix A.7 bounds the effect of the two deviations in Proposition 2 on this stopping criterion, and quantifies the resulting risk of premature stopping.

Stage III: UEG serialization and Answer Generation. At termination round T, ARGUE constructs a structured evidence context from the final UEG $\mathcal{G}_T = (\mathcal{V}_T, \mathcal{E}_T, \boldsymbol{\pi}_T)$. A serializer is then applied to preserve findings, diagnostic concepts, and relation-validity probabilities. Using the estimator in Equation 2, we annotate each finding node with its support and refutation reachability relative to the working answer a_T, yielding:

$$\mathcal{X}_T = \mathrm{Serialize}\big[\mathcal{G}_T, \{(f, \widehat{R}_\rho(\mathcal{G}_T, a_T; \{f\})) : f \in \mathcal{F}_T\}\big], \quad \rho \in \{r, s\}. \quad (6)$$

As a result, the final LLM receives the structured evidence context. $(q, a_T, \mathcal{X}_T)$ and generates the response by integrating findings with their diagnostic relations. The working answer a_T serves as an advisory hypothesis that the model may confirm or revise based on the evidence.

Table 1: Performance comparison on WSI-Bench, **best** and second best are highlighted.

Method	B-1 ↑	R-L ↑	METEOR ↑	Morph. P/R ↑	Diagnosis P/R ↑	Treatment P/R ↑	Avg P/R ↑
Qwen2.5-VL-Instruct	0.295	0.259	0.275	0.209 / 0.169	0.263 / 0.198	0.523/ 0.245	0.232 / 0.178
Patho-R1	0.332	0.239	0.286	0.366 / 0.219	0.437 / 0.246	0.750 / 0.501	0.396 / 0.235
LLaVA-Med	0.135	0.153	0.119	0.225 / 0.300	0.405 / 0.430	0.356 / 0.443	0.269 / 0.333
Quilt-LLaVA	0.276	0.195	0.232	0.299 / 0.381	0.327 / 0.305	0.632 / 0.378	0.318 / 0.364
HistoSelect	0.194	0.182	0.166	0.325 / 0.241	0.432 / 0.343	**0.861** / 0.428	0.368 / 0.270
PathAgent	0.305	0.252	0.298	0.279 / 0.268	0.419 / 0.281	0.444 / 0.798	0.316 / 0.290
PathNavigate	0.239	0.164	0.207	0.300 / 0.199	0.296 / 0.219	0.743 / 0.487	0.316 / 0.214
BEACON	0.221	0.156	0.150	0.253 / 0.199	0.336 / 0.265	0.750 / 0.451	0.289 / 0.223
ARGUE (Ours)	0.346	**0.290**	**0.315**	**0.375** / **0.429**	**0.536** / **0.509**	0.717 / **0.819**	**0.423** / **0.461**

4 EXPERIMENTS

We conduct experiments to evaluate ARGUE through the following research questions:

- **RQ1:** How does the overall performance of ARGUE compare against SOTA baselines?
- **RQ2:** How do ARGUE's core components contribute to its performance?
- **RQ3:** How robust is ARGUE to relation-validity probability noise and implementation choices?

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate ARGUE on three benchmarks: WSI-Bench (Liang et al., 2025), WSI-VQA (Chen et al., 2024), and SlideBench-BCNB (Chen et al., 2025c). These benchmarks cover 30+ cancer types, with WSI-Bench spanning three broad task categories and SlideBench-BCNB providing seven closed-ended question types. Details are provided in Appendix A.1.

Baselines and Metrics. Baselines including *MLLM baselines*: Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Patho-R1 (Zhang et al., 2026), LLaVA-Med (Li et al., 2023), and Quilt-LLaVA (Seyfioglu et al., 2024) and *WSI-VQA baselines*: PathAgent (Chen et al., 2025a), PathNavigate (Yang et al., 2026), BEACON (Wong et al., 2026), and HistoSelect (Huang et al., 2026). We report accuracy for closed-ended questions; BLEU-1 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) for report generation and open-ended VQA; and WSI-Precision (WSI-P) and WSI-Relevance (WSI-R), following WSI-LLaVA (Liang et al., 2025), for open-ended WSI-Bench responses. Overall accuracy, WSI-P, and WSI-R are micro-averaged over their applicable evaluation instances. Details are provided in Appendix A.2.

Implementation details. ARGUE and all agentic baselines use the same backbone models configuration and prompt templates, with Qwen2.5-VL-Instruct for visual observation, Qwen2.5-7B-Instruct (Yang et al., 2024) for text reasoning, and frozen CONCH encoders (Lu et al., 2023) for patch retrieval. Unless otherwise specified, all experiments use serialized contexts of the same length and identical prompts. Alternative configurations using Qwen3-14B (Yang et al., 2025) and Qwen3-8B (Bai et al., 2025a) are detailed in Table 4 of Appendix A.4. Further implementation details are provided in Appendix A.3.

4.2 **RQ1**: OVERALL PERFORMANCE COMPARISON AND COST ANALYSIS

We first examine whether preserving and evaluating complete evidence pathways improves WSI-VQA performance without increasing inference cost. As shown in Tables 1 and 2, ARGUE achieves the best overall performance among the evaluated methods across all three benchmarks, notably improving WSI-VQA accuracy by 9.0% over the strongest baseline: HistoSelect. Despite using general purpose Qwen models for visual observation and text reasoning, ARGUE even outperforms biomedical and pathology specialized pretrained baselines, including LLaVA-Med, Quilt-LLaVA and Patho-R1 in overall performance. More so, these gains come without a substantial increase in inference cost. ARGUE uses slightly more tokens than PathAgent and fewer than BEACON (Figure 3), while maintaining comparable inference time (Figure 9 in the Appendix A.4). Together, ***these results support the practical value of our proposed framework***.

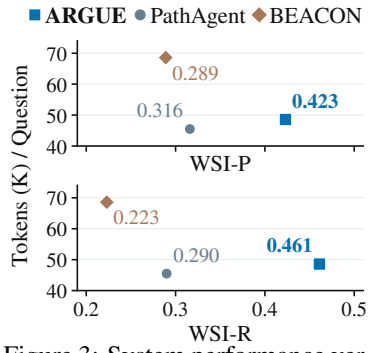

Figure 3: System performance versus token usage on WSI-Bench.

Table 2: Results on WSI-VQA and SlideBench-BCNB, **best** and second best are highlighted.

Method	WSI-VQA			SlideBench-BCNB							
	B-1 ↑	R-L ↑	ACC ↑	Tumor ↑	ER ↑	PR ↑	HER2 ↑	HER2 Expression ↑	Grading ↑	Molecular Subtype ↑	Overall Accuracy ↑
Qwen2.5-VL	0.062	0.099	0.403	0.458	0.417	0.412	0.431	0.214	0.207	0.192	0.335
Patho-R1	0.198	0.229	0.405	0.661	**0.752**	0.601	0.280	0.275	0.424	0.262	0.466
LLaVA-Med	0.124	0.179	0.385	0.239	0.450	0.294	0.262	0.179	0.440	0.124	0.281
Quilt-LLaVA	0.180	0.251	0.438	0.486	0.454	0.427	0.344	0.229	0.367	0.252	0.366
HistoSelect	0.157	0.214	0.459	0.812	0.739	0.555	0.303	0.248	0.435	0.174	0.467
PathAgent	0.116	0.157	0.413	0.784	0.450	0.537	0.578	0.188	0.346	0.170	0.438
PathNavigate	0.159	0.318	0.410	0.376	0.390	0.404	0.716	0.147	0.367	0.248	0.378
BEACON	0.169	0.184	0.413	0.775	0.284	0.312	**0.743**	0.161	0.419	0.142	0.405
ARGUE (Ours)	**0.228**	**0.328**	**0.549**	**0.835**	0.693	**0.656**	0.633	**0.307**	**0.497**	**0.312**	**0.563**

Table 3: Strategy substitutions in ARGUE on WSI-Bench. Bold blue cells mark the best P or R value. Blue arrows show drops from ARGUE (Default) in percentage points. Higher is better.

Stage	Variant	Morph. P/R↑	Diagnosis P/R↑	Treatment P/R↑	Overall P/R↑
(a) Acquisition	Question Relevance	0.360 / 0.420	0.427 / 0.375	0.580 / 0.696	0.383↓4.0 / 0.420↓4.1
	Random Patches	0.232 / 0.289	0.485 / 0.443	0.439 / 0.749	0.294↓12.9 / 0.339↓12.2
	LLM Relevance	0.345 / 0.406	0.450 / 0.465	0.569 / 0.696	0.376↓4.7 / 0.429↓3.2
(b) Sufficiency	LLM Stop	0.314 / 0.403	0.372 / 0.356	0.620 / 0.815	0.338↓8.5 / 0.408↓5.3
(c) Generation	High Reachability	0.367 / 0.419	0.518 / 0.496	0.635 / 0.782	0.409↓1.4 / 0.449↓1.2
	Low Reachability	0.317 / 0.414	0.480 / 0.489	0.643 / 0.709	0.365↓5.8 / 0.441↓2.0
	Random Evidence	0.333 / 0.411	0.495 / 0.506	0.631 / 0.799	0.379↓4.4 / 0.446↓1.5
ARGUE (Default)		**0.375** / **0.429**	**0.536** / **0.509**	**0.717** / **0.819**	**0.423** / **0.461**

4.3 RQ2: Ablation Studies and Design Analysis

Effectiveness of the UEG and diagnostic reachability. Figure 1(c, top) shows that removing diagnostic relations or retaining only partial evidence pathway reduces ARGUE's performance. Complementarily, Figure 1 (c, bottom) shows that methods with higher diagnostic reachability tend to achieve better performance. This association is consistent with our concern that individually plausible observations may provide insufficient support for the final answer when the underlying diagnostic evidence pathways are incomplete. Here, reachability is estimated on UEGs constructed post hoc from each baseline's observations using the same protocol as ARGUE. Beyond pathway structure, we examine the roles of relation-validity probability. Replacing ARGUE's LLM estimated relation-validity probabilities with randomly generated values or CONCH text similarity scores reduces both WSI-P and WSI-R (Figure 4). ***Taken together, these results support the need to preserve evidence pathways across stages to reduce evidence loss, and assess the diagnostic reachability of evidence within each stage to avoid misdirected decisions***.

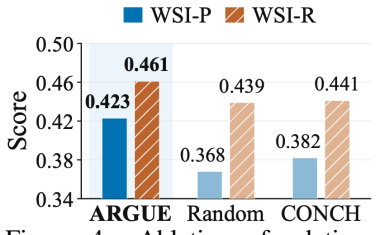

Figure 4: Ablation of relation-validity probability assignment.

Effectiveness of ARGUE's specific strategies in three stages. We replace one of ARGUE's stage specific strategy at a time (Table 3). For ***evidence acquisition***, *Question Relevance* ranks patches by image-text similarity to the question, *Random Patches* randomly select patches, and *LLM Relevance* uses LLM based selection (Weishaupt et al., 2025; Buckley et al., 2025). For ***acquisition sufficiency assessment***, *LLM Stop* refers to an LLM judgment of whether further inspection is needed (Chen et al., 2025a; Wong et al., 2026). For ***answer generation***, *High/Low Reachability* retaining pathways with highest and lowest reachability, respectively. *Random Evidence* randomly selecting pathways in UEG. Under the same inspection budget, all three acquisition alternatives yield lower overall performance than ARGUE (Table 3). Furthermore, replacing the ARGUE's acquisition sufficiency assessment (Section 3.3) with LLM autonomously determine when to stop decreases the WSI-P by 8.5%. In addition, both answer generation variants also underperform ARGUE, with *Low Reachability* performing worse than *Random Evidence* overall. Taken together, ***these ablation studies help rule out alternative explanations for the observed performance gains, supporting the effectiveness of the shared UEG, diagnostic reachability, and the stage-specific strategies built on them***.

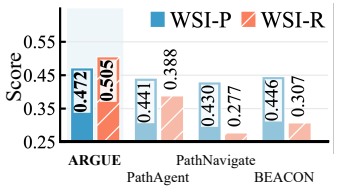

Figure 5: Performance comparison with Patho-R1 as the observer.

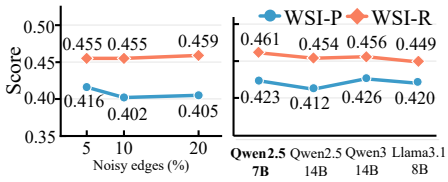

Figure 6: ARGUE's sensitivity to relation-validity probability estimators and noise in their estimates.

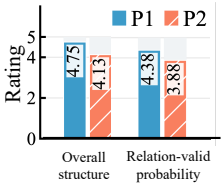

Figure 7: Pathologist ratings.

4.4 **RQ3**: Robustness Analysis

Finally, we test whether ARGUE's pathway-aware reasoning remains effective when relation-validity probabilities contains noise or implementation choices change. On WSI-Bench, varying the candidate pool size B, acquisition weight λ, and Monte Carlo sample count K produces only modest changes in ARGUE's performance (Figure 10 in the Appendix), suggesting limited sensitivity to hyperparameters. Furthermore, replacing the visual observer in all methods with the pathology pretrained Patho-R1 Zhang et al. (2026) preserves ARGUE's lead on both metrics (Figure 5), suggesting that its performance advantage is not tied to the original choice of visual observer. In addition, perturbing $5\%, 10\%, 20\%$ of randomly selected edges with Gaussian noise of variance 0.5 results in only modest performance degradation (Figure 6, left). Using alternative LLMs to estimate relation-validity probability also yields comparable performance across model families and sizes, with no consistent gains from larger models (Figure 6, right). Together, these sensitivity analyses show that **ARGUE remains effective across a broad range of hyperparameter settings and model choices**, while remaining robust to perturbations in relation-validity probabilities.

4.5 Case Study and Pathologist Evaluation

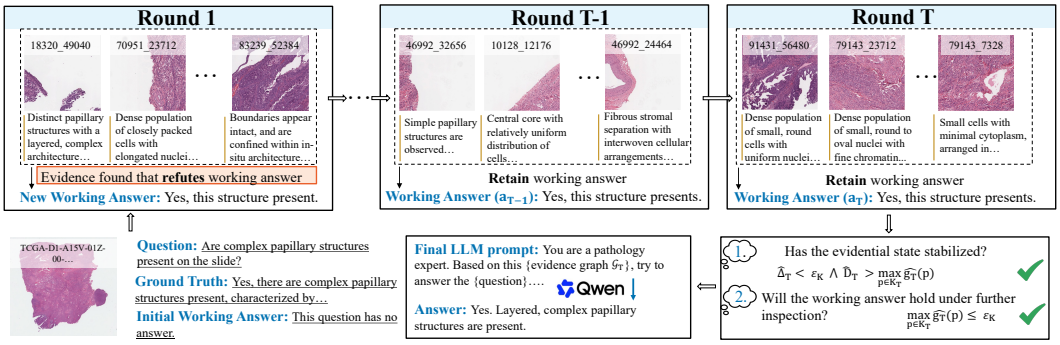

Figure 8: Case study of ARGUE's evidence pathway aware reasoning in a whole-slide image.

The case study in Figure 8 illustrates how ARGUE revises an initially incorrect answer in response to newly acquired counterevidence and subsequently terminates acquisition when the reachability based sufficiency criterion is met. More casestudies are provided in A.4. In addition, the pathologist assessments (Figure 7) provide complementary qualitative support for the plausibility of UEG structure and relation-validity probabilities. Details of the pathologist assessments are provided in Appendix A.4. Furthermore, the lower probability ratings also underscore the importance of representing uncertainty in diagnostic relations rather than assuming their validity. Taken together, **these results provide domain expert evidence for the reliability of the UEG** constructed by our method.

5 Conclusion

To address the limitations of existing approaches, ARGUE coordinates WSI-VQA through a shared Uncertain Evidence Graph and diagnostic reachability. As a result, ARGUE achieves the best overall performance among the evaluated methods across three benchmarks, including a 9% accuracy improvement over the strongest WSI-VQA baseline. Ablation studies validate the effectiveness of its core design components, while robustness analyses demonstrate stable performance under noisy relation probabilities and across different hyperparameter settings and model choices, thereby clearly isolating the sources of the observed performance gains. Blinded assessments by two independent pathologists further support the clinical plausibility of the learned graph structure and the estimated relation validity.

AI USE STATEMENT

In this work, we used generative AI tools to support several aspects of the research and manuscript preparation, including identifying and summarizing relevant literature, assisting with parts of the experimental workflow and the interpretation of experimental results, and conducting manuscript-wide proofreading and language refinement. All AI-assisted text, code, and experimental outputs were reviewed by the authors. The authors further verified cited sources against the original literature and cross-checked reported experimental results against the corresponding experimental records.

REFERENCES

Naheed Anjum Arafat, Ehsan Bonabi Mobaraki, Arijit Khan, Yllka Velaj, and Francesco Bonchi. Estimate and reduce uncertainty in uncertain graphs. In *11th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2024, San Diego, CA, USA, October 6-10, 2024*, pp. 1–12. IEEE, 2024. doi: 10.1109/DSAA61799.2024.10722821. URL https://doi.org/10.1109/DSAA61799.2024.10722821.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Babak Ehteshami Bejnordi, Guido Zuidhof, Maschenka Balkenhol, Meyke Hermsen, Peter Bult, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, and Jeroen van der Laak. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of medical Imaging*, 4(4):044504–044504, 2017.

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. URL https://arxiv.org/abs/2511.17052.

Jingyun Chen, Fengchun Liu, Linghan Cai, Songhan Jiang, Shenjin Huang, Hongpeng Wang, Lequan Yu, and Yongbing Zhang. Agentic visual reasoning in whole-slide pathology images via active perception. *arXiv preprint arXiv:2608.08648*, 2026. URL https://arxiv.org/abs/2608.08648.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. Cod, towards an interpretable medical agent using chain of diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14345–14368, 2025b.

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, 2024. doi: 10.1007/978-3-031-72764-1_23.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5134–5143, 2025c. doi: 10.1109/CVPR52734.2025.00484.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom Oluchi Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G. Elmore, Ranjay Krishna, and Linda G. Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In *IEEE/CVF International Conference on Computer Vision*, pp. 23431–23441, 2025. doi: 10.1109/ICCV51701.2025.02175.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Shengyi Hua, Jianfeng Wu, Tianle Shen, Kangzhe Hu, Zhongzhen Huang, Shujuan Ni, Zhihong Zhang, Yuan Li, Zhe Wang, and Xiaofan Zhang. Pathfound: An agentic multimodal model activating evidence-seeking pathological diagnosis. *Medical Image Analysis*, pp. 104200, 2026.

Wentao Huang, Weimin Lyu, Peiliang Lou, Qingqiao Hu, Xiaoling Hu, Shahira Abousamra, Wenchao Han, Ruifeng Guo, Jiawei Zhou, Chao Chen, and Chen Wang. Act like a pathologist: Tissue-aware whole slide image reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.00667.

Songhan Jiang, Fengchun Liu, Ziyue Wang, Linghan Cai, and Yongbing Zhang. Pathreasoner-r1: Instilling structured reasoning into pathology vision-language model via knowledge-guided policy optimization. *arXiv preprint arXiv:2601.21617*, 2026.

Ruoming Jin, Lin Liu, Bolin Ding, and Haixun Wang. Distance-constraint reachability computation in uncertain graphs. *Proc. VLDB Endow.*, 4(9):551–562, 2011. doi: 10.14778/2002938.2002941. URL http://www.vldb.org/pvldb/vol4/p551-jin.pdf.

Xiangyu Ke, Arijit Khan, and Leroy Lim Hong Quan. An in-depth comparison of s-t reliability algorithms over uncertain graphs. *Proc. VLDB Endow.*, 12(8):864–876, 2019. doi: 10.14778/3324301.3324304. URL http://www.vldb.org/pvldb/vol12/p864-ke.pdf.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Jinyue Li, Yuci Liang, Qiankun Li, Xinheng Lyu, Jiayu Qian, Huabao Chen, Kun Wang, Zhigang Zeng, Anil Anthony Bharath, and Yang Liu. Pathmem: Toward cognition-aligned memory transformation for pathology mllms. *arXiv preprint arXiv:2603.09943*, 2026.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. WSI-LLaVA: A multimodal large language model for whole slide image. In *IEEE/CVF International Conference on Computer Vision*, pp. 22718–22727, 2025. doi: 10.1109/ICCV51701.2025.02109.

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

Peiliang Lou, Yuxin Dong, Caixia Ding, Chunbao Wang, Ruifeng Guo, XiaoBo Pang, Chen Wang, and Chen Li. Pathograph: A graph-based method for standardized representation of pathology knowledge. *Scientific Data*, 12(1):872, 2025.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *CoRR*, abs/2004.09666, 2020. URL https://arxiv.org/abs/2004.09666.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, Georg Gerber, Anil V. Parwani, and Faisal Mahmood. Towards a visual-language foundation model for computational pathology. *CoRR*, abs/2307.12914, 2023. doi: 10.48550/ARXIV.2307.12914. URL https://doi.org/10.48550/arXiv.2307.12914.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush U Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg K. Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative ai copilot for human pathology. *Nature*, 634:466 – 473, 2024. URL https://api.semanticscholar.org/CorpusID:270437603.

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. In *Medical Image Computing and Computer Assisted Intervention*, pp. 680–690, 2025. doi: 10.1007/978-3-032-04971-1_64.

Silviu Maniu, Reynold Cheng, and Pierre Senellart. An indexing framework for queries on probabilistic graphs. *ACM Trans. Database Syst.*, 42(2):13:1–13:34, 2017. doi: 10.1145/3044713. URL https://doi.org/10.1145/3044713.

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.

Gil Patrus Pena and José de Souza Andrade-Filho. How does a pathologist make a diagnosis? *Archives of Pathology & Laboratory Medicine*, 133(1):124–132, 2009. doi: 10.5858/133.1.124.

Michalis Potamias, Francesco Bonchi, Aristides Gionis, and George Kollios. k-nearest neighbors in uncertain graphs. *Proc. VLDB Endow.*, 3(1):997–1008, 2010. doi: 10.14778/1920841.1920967. URL https://vldb.org/pvldb/vol3/R89.pdf.

J. Scott Provan. The complexity of reliability computations in planar and acyclic graphs. *SIAM Journal on Computing*, 15(3):694–702, 1986. doi: 10.1137/0215050.

Mohammad R Rezaei, Reza Saadati Fard, Jayson Parker, Rahul G Krishnan, and Milad Lankarany. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between llms and evolving medical knowledge. In *EMNLP (Findings)*, pp. 12682–12701, 2025.

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.

Susu Sun, Leslie Tessier, Frédérique Meeuwsen, Clément Grisi, Dominique van Midden, Geert Litjens, and Christian F Baumgartner. Label-free concept based multiple instance learning for gigapixel histopathology. *arXiv preprint arXiv:2501.02922*, 2025a.

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. In *Advances in Neural Information Processing Systems*, 2025b. URL https://papers.nips.cc/paper_files/paper/2025/hash/933b5d002cf251b3e854d586e55ac58c-Abstract-Conference.html.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H Bernhard, Ran A Godrich, Juan A Retamero, et al. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, pp. 1–11, 2026.

Sheng Wang, Ruiming Wu, Charles Herndon, Songhao Li, Yihang Liu, Shunsuke Koga, Xiaowei Xu, David E Elder, Jonathan Alex Miles, Annie Jin, et al. Pathology-cot: learning visual chain-of-thought agents from expert whole-slide image diagnosis behaviour. *Nature Biomedical Engineering*, pp. 1–15, 2026.

Luca L. Weishaupt, Chengkuan Chen, Drew F. K. Williamson, Richard J. Chen, Guillaume Jaume, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Ming Y. Lu, and Faisal Mahmood. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964*, 2025. URL https://arxiv.org/abs/2506.20964.

Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10370–10388, 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. URL https://arxiv.org/abs/2608.05757.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

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 Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 28418–28426. AAAI Press, 2026. doi: 10.1609/AAAI.V40I33.40071. URL https://doi.org/10.1609/aaai.v40i33.40071.

Beidi Zhao, Gexin Huang, Ciro Zhang, Anqi Li, Yusheng Tan, Chen Zhou, Gang Wang, Zu-hua Gao, and Xiaoxiao Li. Slidebank: A persistent hierarchical evidence bank for consistent whole-slide reasoning. *arXiv preprint arXiv:2609.00342*, 2026.

Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus CF Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. *npj Digital Medicine*, 7(1):383, 2024.

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pp. 4442–4457, 2025.

A Appendix

A.1 Datasets Details

We evaluate ARGUE on three complementary WSI-VQA benchmarks that differ in scale, disease coverage and question type. For example, WSI-Bench emphasizes broad morphology grounded analysis across diverse cancer types, WSI-VQA formulates heterogeneous breast-cancer prediction tasks as generative question answering, and SlideBench-BCNB provides a focused evaluation of finegrained breastbiopsy assessment on a non-TCGA cohort. In the experiments, we use the official test/train split of all datasets across all methods.

WSI-Bench. WSI-Bench (Liang et al., 2025) is a large scale benchmark containing 179,569 question-answer pairs from 9,850 WSIs spanning 30 cancer types. Its task design follows the progression of pathological interpretation from lower level visual observation to higherlevel clinical conclusions. For example, morphology questions cover global slide appearance, key diagnostic findings, regional tissue organization, and specific microscopic features. Diagnostic questions address histological typing, grading, molecular subtyping, and staging. Clinically oriented tasks further include treatment recommendation, prognosis, and report generation. The benchmark contains both open-ended and closed-ended questions, requiring models to produce detailed slide grounded descriptions in addition to discrete predictions. We therefore use WSI-Bench to assess whether an agent can collect spatially distributed morphological evidence and connect local observations to diagnostic and clinically relevant outputs.

WSI-VQA. WSI-VQA (Chen et al., 2024) contains 8,672 slide level question answer pairs associated with 977 WSIs from the TCGA-BRCA cohort. It reformulates a heterogeneous set of breast cancer assessment problems into a unified question answering format. The benchmark combines 4,535 closed-ended pairs generated from WSI captions with 4,137 open-ended pairs constructed from TCGA clinical variables, resulting in questions that range from categorical biomarker queries to free form slide level responses. Compared with the broader disease coverage of WSI-Bench, WSI-VQA provides a more concentrated evaluation of whether an agent can adapt its evidence acquisition and answer synthesis to different clinical targets.

SlideBench-BCNB. SlideBench-BCNB (Chen et al., 2025c) is originally derived from the Early Breast Cancer Core Needle Biopsy (BCNB) cohort containing 7,274 question answer pairs from 1058 WSIs. The benchmark converts the original classification objectives and labels into closed-ended VQA questions with explicit answer options, enabling a standardized evaluation of seven clinically relevant tasks: tumor typing, ER status, PR status, HER2 status, HER2 expression level, histological grading, and molecular subtyping. Because it is constructed from a non TCGA biopsy cohort, SlideBench-BCNB complements the other two benchmarks by evaluating fine grained diagnostic discrimination under a distinct data distribution.

A.2 Details of Baselines

We organize the baseline methods into two categories according to whether they explicitly perform WSI-level evidence acquisition on purpose. The first category consists of **multimodal large language model (MLLM) baselines**, which generate answers from a fixed visual context and primarily evaluate the benefits of general purpose, biomedical, or pathology specific visual language pretraining and reasoning. These models are pretrained and do not explicitly maintain an iterative whole slide search process. The second category consists of **WSI-VQA baselines**, which are specifically designed for gigapixel slide analysis and introduce dedicated mechanisms for region selection, multi scale navigation, or multi agent verification. This categorization separates improvements arising from stronger multimodal representations and domain specific training from those arising from explicit whole slide evidence acquisition and control.

MLLM Baselines. Qwen2.5-VL (Bai et al., 2025b) is a general purpose multimodal large language model composed of native dynamic-resolution perception and broad multimodal reasoning. In our comparison, it measures how far a non pathology specific MLLM can reason from a WSI thumbnail without an explicit mechanism for controlling evidence acquisition on WSI. Moving from general-purpose modeling to medical domain adaptation, LLaVA-Med (Li et al., 2023) adapts the LLaVA architecture to the biomedical domain through a two stage training curriculum. It first aligns visual representations with biomedical concepts using image–caption pairs from PubMed

Central and subsequently performs end to end instruction tuning on biomedical visual conversations, enabling conversational question answering over medical images. Quilt-LLaVA (Seyfioglu et al., 2024) further specializes this adaptation to histopathology by addressing the limited spatial grounding of existing pathology instruction data. The instruction data are constructed from spatially localized narratives extracted from educational histopathology videos. This supervision grounds diagnostic language in localized tissue appearances and adapts the LLaVA framework to pathology specific visual understanding and reasoning. Patho-R1 (Zhang et al., 2026) further shifts the research focus from domain alignment to diagnostic reasoning. By combining reasoning oriented pathology data with multimodal reinforcement learning, it encourages the model to produce explicit diagnostic rationales. Its training combines continued pretraining on pathology image text data, supervised finetuning on diagnostic chain of thought examples, and reinforcement learning to improve both answer accuracy and reasoning quality.

WSI-VQA Baselines. HistoSelect (Huang et al., 2026) identifies visual evidence selection as the principal bottleneck in WSI-VQA reasoning. It performs question guided, coarse to fine visual selection, thereby reducing the amount of irrelevant visual context presented to the answer model. PathAgent (Chen et al., 2025a) is a training free iterative WSI agentic system organized around a Navigator, a Perceptor, and an Executor. These components expose PathAgent's iterative process of slide exploration, morphological evidence accumulation, hypothesis reflection, and self correction, making the evidential basis of its final answer traceable. PathNavigate (Yang et al., 2026) targets another limitation of training free agents: question first retrieval is more likely to overlook diagnostically important morphology that is not explicitly mentioned in the query. Its main contribution is a slide first navigation strategy that identifies visually surprising regions before question matching and preserves the resulting global slide context in a shared online memory. BEACON (Wong et al., 2026) further moves beyond the semantic relevance between question and patch by formulating patch selection as Bayesian experimental design. It selects evidence according to its expected ability to reduce uncertainty among competing diagnostic hypotheses, establishing uncertainty reduction as the acquisition objective.

Details of Baseline Reproduction. We present more details of baseline reproduction here.

For ***MLLM baselines*** such as LLaVA-Med, Quilt-LLaVA, and Patho-R1, we retain their original checkpoints and visual encoders, while give them the same WSI thumbnail input. For Qwen2.5-VL, we retain its checkpoint and replace only the language backbone with Qwen2.5-7B-Instruct, while use the same WSI thumbnail.

For ***WSI-VQA baselines***, we preserve the components that define their algorithmic contributions and standardize only interchangeable pretrained models. For HistoSelect, we retain its LongNet (Ding et al., 2023), the multimodal projection layer, the tissue group sampler, and the patch selector, but disable LoRA (Hu et al., 2021) adaptation of its Qwen2.5 answer model so to ensure a fair comparison with other baselines. Such that answer generation does not benefit from dataset specific language model finetuning. Similarly, PathAgent, PathNavigate, and BEACON retain their original workflows but share a standardized pretrained model configuration: CONCH for retrieval or navigation, Qwen2.5-VL for patch observation, including both the generic descriptor and Perceptor in PathAgent, and Qwen2.5-7B-Instruct for language side control and answer generation. Finally, all methods are evaluated under identical resource budgets.

A.3 Implementation Details.

Data Processing. We extract 4096×4096 patches at $40\times$ magnification, using the CLAM (Lu et al., 2020) preprocessing pipeline to filter out background regions. Frozen CONCH image and text encoders (Lu et al., 2023) compute cosine similarities between patches and the question or evidence gap description, with patch embeddings precomputed once per slide.

Hyperparameters Used. At each acquisition round, we retrieve up to $B = 12$ candidate patches for diagnostic-reachability reranking and set the acquisition weight to $\lambda = 0.5$. Each diagnostic reachability estimate uses $K = 1600$ Monte Carlo UEG realizations.

Evaluation protocal Recent WSI-VQA studies (Yang et al., 2026; Chen et al., 2025a; Huang et al., 2026; Lyu et al., 2025) typically report results from a single execution of the full evaluation pipeline rather than aggregating statistics over multiple independent runs. This is because of the substantial computational overhead of repeatedly executing the full agentic pipeline, which would pro-

portionally increase LLM inference costs, particularly in terms of token consumption. We therefore follow the same evaluation protocol and report results without averaging over multiple independent runs.

Details of the effectiveness of evidence pathway. Figure 1(c, top) compares the full UEG with *No Pathway*, which retains only finding nodes, and *Partial Pathway*, which retains all nodes but removes diagnostic relations. In Figure 1 (c, bottom), we set $S = \{f\}$ in Equation 1 for each finding node $f \in \mathcal{F}_t$ and estimate its reachability to the corresponding answer terminal using Equation 2. For each baseline, we construct a UEG post hoc from its observations using the same protocol and configuration as ARGUE solely to estimate the reachability of their selected patches. We then average these estimates over all observations in the sampled cases.

A.4 Additional Experiments

Table 4: Comparison on WSI-Bench with shared alternative backbones: Qwen3-14B for text reasoning and Qwen3-VL-8B-Instruct for visual observation, **best** and second best are highlighted.

Method	B-1 ↑	R-L ↑	METEOR ↑	Morph. P/R ↑	Diagnosis P/R ↑	Treatment P/R ↑	Avg P/R ↑
PathAgent	0.256	0.176	0.199	0.375 / 0.330	0.442 / **0.474**	0.471 / **0.722**	0.393 / 0.375
PathNavigate	0.226	0.166	0.237	0.316 / 0.291	0.300 / 0.405	0.485 / 0.687	0.319 / 0.329
BEACON	0.268	0.175	0.195	0.362 / 0.253	0.465 / 0.322	0.679 / 0.434	0.396 / 0.275
ARGUE (Ours)	**0.332**	**0.271**	**0.354**	**0.407** / **0.471**	**0.510** / 0.461	**0.707** / 0.639	**0.440** / **0.475**

Domain expert assessment of the UEG. We randomly sampled 10 WSI-Bench questions from distinct WSIs. Under a blinded evaluation protocol, two independent pathologists rated the clinical coherence of the overall graph structure and the plausibility of relation-validity probabilities on a 1–5 rating scale, allowing decimal valued scores (1 = strongly disagree, 5 = strongly agree). For the relation-validity probability assessment, the pathologists randomly sampled 30% of the relations from the UEGs, covering three relation types: finding-to-concept, concept-to-concept, and concept-to-terminal. Each sampled relation was presented alongside the corresponding question and the relevant context needed to evaluate the relation and the plausibility of its assigned validity probability. As shown in Figure 7, the two pathologists provided favorable and broadly consistent ratings for both UEG structure and the relation-validity probability.

Evaluation with alternative backbone models. To evaluate the effectiveness of ARGUE with stronger LLM backbone, we re-evaluate ARGUE and three agentic baselines: PathAgent, PathNavigate, and BEACON on WSI-Bench using Qwen3-instruct-14B for all text reasoning and Qwen3-VL-8B-Instruct for visual observation. All compared methods use this shared backbone configuration. As shown in Table 5 4, ARGUE achieves the highest overall WSI-P and WSI-R of 0.440 and 0.475, respectively, while also obtaining the best BLEU-1, ROUGE-L, and METEOR scores. These results further support the effectiveness of preserving evidence pathways through the UEG and guiding reasoning with diagnostic reachability, suggesting that ARGUE's performance advantage is not restricted to the default backbone configuration.

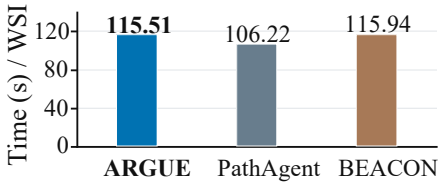

Figure 9: Inference time comparison of ARGUE, PathAGent, and BEACON.

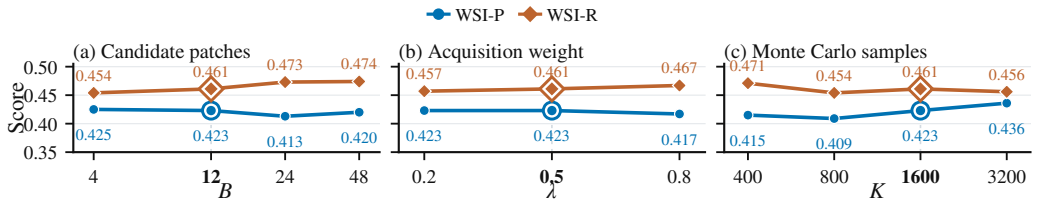

Figure 10: Hyperparameter sensitivity of ARGUE on WSI-Bench. We report WSI-P and WSI-R of ARGUE for different values of (a) the candidate pool size B, (b) the acquisition weight λ, and (c) the number of Monte Carlo UEG realizations K. Highlighted markers denote the default settings: $B = 12$, $\lambda = 0.5$, and $K = 1600$.

More Case Studies. We provide more case studies of ARGUE on WSI-Bench here.

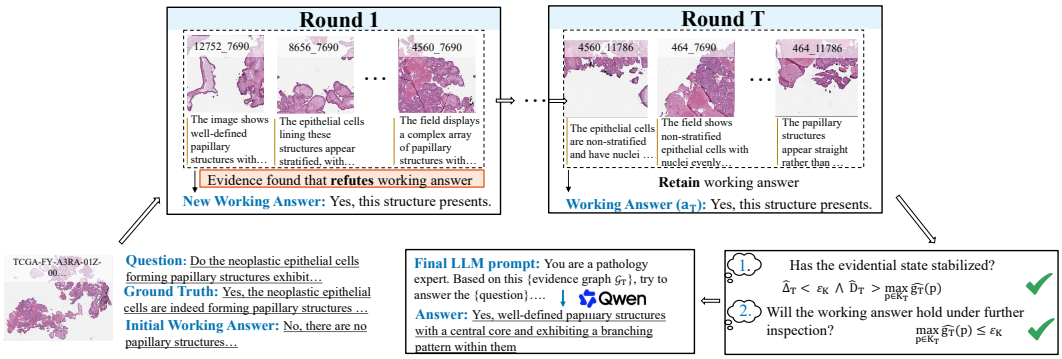

Figure 11: **Case study of identifying architectural features of papillary structures.** Newly acquired observations lead ARGUE to revise an initially negative working answer. The final response describes the papillary architecture, including a central core and branching pattern.

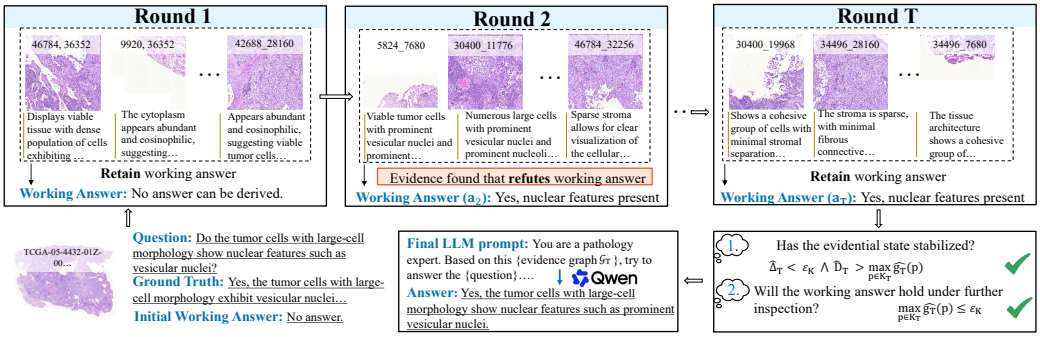

Figure 12: **Case study of identifying vesicular nuclei in tumor cells with large-cell morphology**. After the first acquisition round leaves the question unresolved, observations in the second round support an affirmative working answer, which is retained at termination.

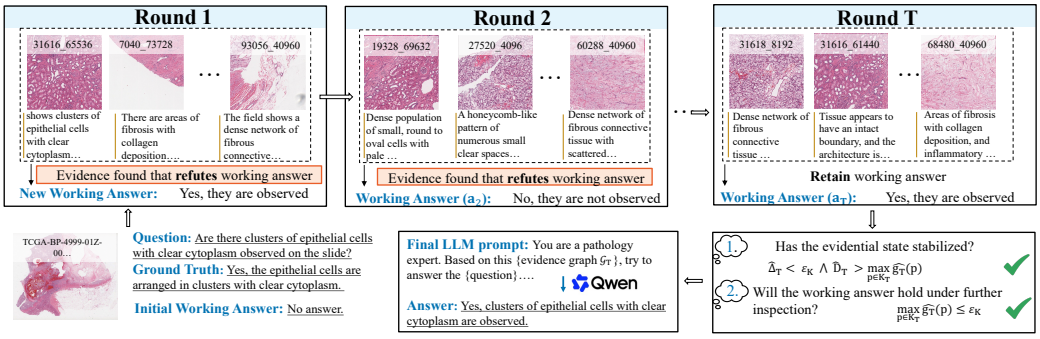

Figure 13: **Case study of assessing the presence of epithelial cell clusters with clear cytoplasm.**

A.5 Proof of Proposition 1

Proposition 1. For each $\rho \in \{s, r\}$, exact evaluation of $R_\rho(\mathcal{G}_t, a_t; \mathcal{S})$ is #P-hard when $|\mathcal{S}| = 1$ and relation validities are independent.

Proof. Diagnostic reachability is the probability that a UEG realization contains a complete evidence pathway to an answer terminal. We construct a UEG whose evidence pathway exists exactly when a selected edge subset of a directed graph connects its source to its sink. Consider a directed acyclic graph $H = (V_H, E_H)$ with source s_H and sink t_H. Let $m = |E_H|$, and let $N_H(s_H, t_H)$ count the edge subsets in which s_H reaches t_H. Computing this count is #P-complete (Provan, 1986).

Fix a round t. Construct a UEG $\mathcal{G}_t$ with $\mathcal{S} = \{f\}$, one concept c_v for every $v \in V_H$, and terminals a_t^s and a_t^r. Add $f \to c_{s_H}$, $c_{t_H} \to a_t^s$, and $c_u \to c_v$ for every $(u, v) \in E_H$. Give all $m + 2$ relations independent validity probability $1/2$. The terminal a_t^r has no incoming relation. Each realization has probability $2^{-(m+2)}$. It contains an evidence pathway from f to a_t^s exactly when $f \to c_{s_H}$ and $c_{t_H} \to a_t^s$ are valid and the valid concept relations connect s_H to t_H. Hence,

$$R_s(\mathcal{G}_t, a_t; \{f\}) = 2^{-(m+2)} N_H(s_H, t_H).$$

Therefore, an exact value of $R_s(\mathcal{G}_t, a_t; \{f\})$ determines the #P-complete count $N_H(s_H, t_H)$. The reduction uses a single finding and independent relation validities, proving the claimed #P-hardness for $\rho = s$. Connecting c_{t_H} to a_t^r instead proves it for $\rho = r$. As a result, we prove Proposition 1 holds. □

A.6 Proof of the Combined Deviation Bound for Diagnostic Reachability

Proposition 2 (Combined deviation bound for diagnostic reachability). Fix a UEG $\mathcal{G}_t$, a working answer a_t, a source set $\mathcal{S} \subseteq \mathcal{F}_t$, and a direction $\rho \in \{s, r\}$ before Monte Carlo sampling. For each of the M uncertain relations, suppose the estimated validity probability p_j and its reference value $p_j^\star$ satisfy $|p_j - p_j^\star| \le \varepsilon_j^{\mathrm{LLM}} \in [0, 1]$. Let $\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S})$ use K i.i.d. UEG realizations from $\mu_{\mathcal{G}_t}$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left| \widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R_\rho^\star(\mathcal{G}_t, a_t; \mathcal{S}) \right| \le \sqrt{\frac{\log(2/\delta)}{2K}} + \beta_{\mathcal{G}_t},$$

where $R_\rho^\star$ uses the reference probabilities and $\beta_{\mathcal{G}_t} := 1 - \prod_{j=1}^{M} (1 - \varepsilon_j^{\mathrm{LLM}})$.

Proof. The estimated and reference reachabilities concern the same nodes, relations, and reachability event. They differ only in the validity probabilities assigned to those relations. Therefore, we use one uniform draw per relation to compare the two distributions.

For relation e_j, draw U_j uniformly from $[0, 1]$, independently across j, and set $Z_{t,e_j} = \mathbf{1}\{U_j \le p_j\}$ and $Z_{t,e_j}^\star = \mathbf{1}\{U_j \le p_j^\star\}$. These coordinates form the assignments $Z_t = (Z_{t,e_j})_{j=1}^M$ and $Z_t^\star = (Z_{t,e_j}^\star)_{j=1}^M$. The two assignments then have the estimated and reference distributions, respectively. They differ at relation e_j precisely when U_j lies between p_j and $p_j^\star$, an event of probability $|p_j - p_j^\star|$. Independence across relations gives

$$\Pr(Z_t \ne Z_t^\star) = 1 - \prod_{j=1}^{M} (1 - |p_j - p_j^\star|) \le 1 - \prod_{j=1}^{M} (1 - \varepsilon_j^{\mathrm{LLM}}) = \beta_{\mathcal{G}_t}. \tag{7}$$

When the assignments agree, they produce the same UEG realization and hence the same reachability indicator for $\mathcal{S}$ and a_t^ρ. The difference between the two reachability probabilities is therefore no greater than the probability that the assignments disagree:

$$\left| R_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R_\rho^\star(\mathcal{G}_t, a_t; \mathcal{S}) \right| \le \beta_{\mathcal{G}_t}. \tag{8}$$

For $M = 0$, the empty product gives $\beta_{\mathcal{G}_t} = 0$.

We now consider the Monte Carlo estimate under the estimated probabilities. Each of its K sampled UEG realizations gives an indicator $I_\rho(\mathcal{G}_t^{(k)}, a_t; \mathcal{S})$. These indicators are i.i.d. Bernoulli variables with mean $R_\rho(\mathcal{G}_t, a_t; \mathcal{S})$. Since their average is $\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S})$, Hoeffding's inequality gives, for $\eta > 0$,

$$\Pr\Big(\Big|\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R_\rho(\mathcal{G}_t, a_t; \mathcal{S})\Big| > \eta\Big) \le 2e^{-2K\eta^2}. \tag{9}$$

Set $\eta = \sqrt{\log(2/\delta)/(2K)}$. The sampling deviation is then at most η with probability at least $1-\delta$. The relation-probability deviation in Equation 8 is fixed before sampling. On the same event, the triangle inequality yields

$$\begin{aligned}\Big|\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R^\star_\rho(\mathcal{G}_t, a_t; \mathcal{S})\Big| &\le \Big|\widehat{R}_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R_\rho(\mathcal{G}_t, a_t; \mathcal{S})\Big| \\ &\quad + \Big|R_\rho(\mathcal{G}_t, a_t; \mathcal{S}) - R^\star_\rho(\mathcal{G}_t, a_t; \mathcal{S})\Big| \\ &\le \sqrt{\frac{\log(2/\delta)}{2K}} + \beta_{\mathcal{G}_t}.\end{aligned}$$

Since this event has probability at least $1-\delta$, the displayed inequality proves Proposition 2. □

A.7 Deviation Bound for Evidence Acquisition Sufficiency Assessment

Equation 5 compares reachability across rounds, candidate gains in refutation reachability, and the current support and refutation reachability difference. In this section, we bound the effect of Monte Carlo sampling and relation-validity probability errors on these comparisons, where leaves the maximum number of acquisition rounds is a separate termination condition.

For $\rho \in \{s, r\}$ and $p \in \mathcal{K}_t$, we write the reachability change and candidate gain as

$$\begin{aligned} d_{t,\rho} &:= R_\rho(\mathcal{G}_t, a_t) - R_\rho(\mathcal{G}_{t-1}, a_{t-1}), \\ g_t(p) &:= R_r(\mathcal{G}_t \oplus x_p, a_t) - R_r(\mathcal{G}_t, a_t). \end{aligned} \tag{10}$$

Let $\widehat{\Delta}_t = \max_{\rho\in\{s,r\}} |\widehat{d}_{t,\rho}|$, then the reference value of the left-hand side of the first condition in Equation 5 is

$$\Theta^\star_t := \max\left\{\max_{\rho\in\{s,r\}} |d^\star_{t,\rho}|,\ \max_{p\in\mathcal{K}_t} g^\star_t(p)\right\}. \tag{11}$$

Applying the $\beta_{\mathcal{G}}$ bound in Equation 8 to both UEGs in each comparison gives

$$\overline{\beta}_t := \max\left\{\beta_{\mathcal{G}_t} + \beta_{\mathcal{G}_{t-1}},\ \max_{p\in\mathcal{K}_t}\left(\beta_{\mathcal{G}_t\oplus x_p} + \beta_{\mathcal{G}_t}\right)\right\}. \tag{12}$$

For the second condition, write $D^\star_t := R^\star_s(\mathcal{G}_t, a_t) - R^\star_r(\mathcal{G}_t, a_t)$ and use $\widehat{D}_t$ from Stage I.

To account for at most T tests and B evaluated candidates at each test, set

$$r_{K,B,T}(\delta) := \sqrt{\frac{2}{K}\log\frac{T(B+5)}{\delta}}. \tag{13}$$

We can then obtain the following proposition.

Proposition 3 (Premature-stop bound). Suppose the joint criterion in Equation 5 is tested at $2 \leq t \leq T$, with $1 \leq |\mathcal{K}_t| \leq B$, and the assumptions of Proposition 2 hold for every UEG in these tests. Each comparison uses K pairs of realizations that are independent conditional on the history before sampling. The UEGs, working answers, and any candidate patch are fixed before sampling. A Stage I estimate may be reused if it satisfies these conditions for the same comparison. Let Stop_t denote that the joint criterion in Equation 5 is met at round t. For any $\delta \in (0, 1)$, we have

$$\Pr\big[\exists\, 2 \leq t \leq T : \mathrm{Stop}_t \text{ and } \Theta_t^\star > \varepsilon_K + r_{K,B,T}(\delta) + \overline{\beta}_t\big] \leq \delta. \tag{14}$$

On the same event of probability at least $1 - \delta$, every such stop also satisfies

$$\begin{aligned} D_t^\star - \max_{p \in \mathcal{K}_t} g_t^\star(p) \geq \widehat{D}_t - \max_{p \in \mathcal{K}_t} \widehat{g}_t(p) - 2r_{K,B,T}(\delta) \\ - 2\beta_{\mathcal{G}_t} - \max_{p \in \mathcal{K}_t} \left(\beta_{\mathcal{G}_t \oplus x_p} + \beta_{\mathcal{G}_t}\right). \end{aligned} \tag{15}$$

Proof. Fix a comparison used at a stopping test and condition on the history before its samples are drawn. Each sample contributes the difference of two reachability indicators, one from each side of the comparison. Every difference lies in $[-1, 1]$. Its mean is respectively $d_{t,\rho}$, $g_t(p)$, or $D_t := R_s(\mathcal{G}_t, a_t) - R_r(\mathcal{G}_t, a_t)$, and its average over K samples is the corresponding quantity.

The paired draws are conditionally independent, so for any such mean u, its estimate $\widehat{u}$, and $r > 0$, Hoeffding's inequality gives

$$\Pr(\widehat{u} - u > r) \leq e^{-Kr^2/2}, \qquad \Pr(u - \widehat{u} > r) \leq e^{-Kr^2/2}.$$

The absolute temporal change requires both tails for each direction. A candidate gain requires only $g_t(p) \leq \widehat{g}_t(p) + r$, while the support and refutation comparison requires $D_t \geq \widehat{D}_t - r$. Thus the two temporal changes contribute four tails, the candidate gains contribute at most B, and the last comparison contributes one. These bounds remain valid when a Stage I estimate is reused, since each is conditional on the history before its own samples were drawn. Taking expectations and summing over at most T tests gives a failure probability of at most $T(B + 5)e^{-Kr^2/2}$. Setting $r = r_{K,B,T}(\delta)$, we obtain

$$|\widehat{d}_{t,\rho} - d_{t,\rho}| \leq r, \quad g_t(p) \leq \widehat{g}_t(p) + r, \quad D_t \geq \widehat{D}_t - r$$

simultaneously for every tested round, direction, and evaluated candidate with probability at least $1 - \delta$.

These bounds concern sampling from the estimated probabilities. To move to the reference probabilities, we could apply Equation 8 separately to the UEGs in each comparison. The two reachability errors in a temporal change add, as do those in a candidate gain. Therefore,

$$\begin{aligned} |d_{t,\rho} - d_{t,\rho}^\star| &\leq \beta_{\mathcal{G}_t} + \beta_{\mathcal{G}_{t-1}}, \\ |g_t(p) - g_t^\star(p)| &\leq \beta_{\mathcal{G}_t \oplus x_p} + \beta_{\mathcal{G}_t}. \end{aligned} \tag{16}$$

The support and refutation reachabilities defining D_t each contribute an error from $\mathcal{G}_t$, giving $|D_t - D_t^\star| \leq 2\beta_{\mathcal{G}_t}$.

If Stop_t occurs, the first condition in Equation 5 gives $|\widehat{d}_{t,\rho}| \leq \varepsilon_K$ for both directions and $\widehat{g}_t(p) \leq \varepsilon_K$ for every $p \in \mathcal{K}_t$. On the simultaneous sampling event, the preceding bounds imply

$$\begin{aligned} |d_{t,\rho}^\star| &\leq |\widehat{d}_{t,\rho}| + r + \beta_{\mathcal{G}_t} + \beta_{\mathcal{G}_{t-1}}, \\ g_t^\star(p) &\leq \widehat{g}_t(p) + r + \beta_{\mathcal{G}_t \oplus x_p} + \beta_{\mathcal{G}_t}. \end{aligned}$$

Taking the maximum over both directions and all evaluated candidates gives $\Theta_t^\star \leq \varepsilon_K + r_{K,B,T}(\delta) + \overline{\beta}_t$. The sampling event holds for all tests with probability at least $1 - \delta$, which proves Equation 14.

For the second condition of the stopping rule, the sampling event and the bound on $D_t - D_t^\star$ give $D_t^\star \geq \widehat{D}_t - r - 2\beta_{\mathcal{G}_t}$. The bound on $g_t^\star(p)$ above holds for every candidate. Subtracting the largest candidate gain yields Equation 15. This completes the proof of Proposition 3. □

A.8 Additional Related Work Discussions

From Uncertain Graph Querying to Agentic Evidence Reasoning Uncertain graphs represent noisy relational data by assigning probabilities to candidate edges and interpreting the graph through possible world semantics. They have been studied in social, biological, and communication networks, with representative tasks including k-nearest neighbor search (Potamias et al., 2010), distance constrained reachability (Jin et al., 2011), and source to target reliability estimation through sampling and indexing (Maniu et al., 2017; Ke et al., 2019). Beyond query evaluation, recent work also investigates selecting existing edges for probability updates or verification to reduce uncertainty under a limited budget (Arafat et al., 2024).

These formulations advance probabilistic querying over given graphs, but they do not by themselves address the coupled requirements of agentic WSI-VQA: observation driven graph construction, reachability estimation under imperfect relation validity probabilities, and cross-stage evidence coordination. First, standard uncertain graph methods operate on a pre defined topology graph (Maniu et al., 2017; Ke et al., 2019). However, in our setting, patch inspection introduces previously unrepresented morphological findings, diagnostic concepts, and relations. The evidence graph must therefore be progressively constructed from observations and updated as the working answer evolves, rather than merely instantiated from a predefined candidate structure.

Second, the approximation guarantees of standard edge reliability estimators are conditional on the supplied relation-validity probabilities (Jin et al., 2011; Potamias et al., 2010). However, accurately evaluating the entire probabilistic graph does not account for errors in the probabilities themselves. Instead, we explicitly model bounded deviations from reference probabilities and analyze their combined effect with finite sample Monte Carlo error. Furthermore, through extensive experiments, we demonstrate that estimation errors in relation-validity probabilities have only a minor impact on ARGUE's performance in practice.

Third, the WSI-VQA task requires coordinating evidence use across multiple reasoning stages, rather than evaluating graph queries in isolation. Therefore, uncertain graph alone do not specify how evidence should guide further acquisition, justify stopping, and remain available for answer generation. ARGUE therefore integrates the evolving UEG and diagnostic reachability into a unified cross-stage control mechanism, with the graph preserving evidence pathways and their uncertainty as a shared evidential state. Reachability gains guide evidence acquisition, while reachability stability and remaining opportunities for refutation inform acquisition sufficiency assessment. The same graph, augmented with reachability annotations, then grounds answer generation. This coupling preserves the evidential basis of acquisition and stopping decisions through the final response.

Graph-augmented medical QA and evidence reasoning. Recent medical QA systems increasingly use knowledge graphs to ground and organize diagnostic reasoning. MindMap mines query-conditioned evidence subgraphs from an external knowledge graph, aggregates them into a reasoning graph, and prompts an LLM to generate an answer together with a graph structured reasoning trace (Wen et al., 2024). MedRAG constructs a hierarchical diagnostic knowledge graph over diseases and manifestations, retrieves disease discriminative features together with relevant EHRs, and uses graph derived discriminability to generate targeted follow up questions (Zhao et al., 2025). AMG-RAG dynamically constructs a medical knowledge graph with confidence scored relations and propagates confidence during graph traversal to retrieve external evidence for medical question answering (Rezaei et al., 2025). These approaches demonstrate the value of structured knowledge and relational evidence for medical reasoning.

Related pathology studies further investigate structured diagnostic knowledge and persistent evidence representations. PathoGraph (Lou et al., 2025) standardizes pathology knowledge through interconnected entity, phenotype, and diagnosis graphs, explicitly representing relations between pathological phenotypes and diagnostic possibilities. PathMem (Li et al., 2026) constructs a pathology knowledge graph from medical literature as long term memory, consolidates multi source relation evidence, and introduces a Memory Transformer to activate and refine context relevant knowledge into working memory. SlideBank (Zhao et al., 2026) instead organizes slide derived observations into a persistent, concept indexed evidence bank. It links pathology signals to WSI coordinates, retains signal status and enable evidence reuse across question.

ARGUE shares these efforts' emphasis on structured evidence, but focuses on the joint use of uncertain diagnostic pathways for inference time evaluation and control. Specifically, ARGUE constructs an evolving, slide specific uncertain evidence graph directly from acquired morphological observations, preserves it across evidence acquisition, acquisition sufficiency assessment, and answer generation, and estimates the diagnostic reachability of evidence. Unlike confidence propagation or structural discriminability scores, diagnostic reachability is defined over uncertain graph realizations and explicitly accounts for alternative evidence pathways sharing diagnostic relations. Consequently, the graph serves as a shared probabilistic control state for the full WSI-VQA loop, rather than only as retrieved context or an explanation interface.

LLM-based Determination of Clinical Relations. LLMs are recently broadly used in medical reasoning and computational pathology to assess relations between medical entities, observations, and clinical claims. These applications can be organized around two complementary directions: determining whether a clinical relation holds and assigning a numerical probability to its validity.

Determining whether a relation holds. The first direction is to assess whether a proposed association relation is justified by the available context. In medical knowledge graph construction, AMG-RAG (Rezaei et al., 2025) provides an LLM relation agent with descriptions of two medical entities and retrieved supporting evidence, and asks it to determine the relationship type. In diagnostic reasoning, CoD (Chen et al., 2025b) first abstracts patient symptoms and recalls candidate diseases, then uses an LLM to analyze whether each candidate disease corresponds to the observed symptoms.

A similar verification function appears in WSI-Agents (Lyu et al., 2025): its logic agent extracts claims and associated evidence from generated responses, checks whether claims are mutually compatible, and evaluates whether the evidence supports each claim. Its fact agent additionally retrieves pathology knowledge using diagnostic keywords and checks whether the claims violate the retrieved reference information. Across these settings, the shared operation is contextual assessment of whether medical concepts are related or whether clinical statements are supported.

Assigning validity probability to clinical relatons. Beyond verifying the existence of relations, existing methods also quantify their confidence or support strength. AMG-RAG (Rezaei et al., 2025) explicitly asks its relation agent to assign an evidence strength score, formalized as $s_{ij} \in [0, 1]$, alongside the relationship type. These scores are then stored on graph edges and used for multiplicative confidence propagation and threshold based pruning during graph traversal. CoD (Chen et al., 2025b) similarly generates a normalized confidence distribution over candidate diseases, and uses a confidence threshold to decide whether to diagnose or request further symptoms.

In pathology, WSI-Agents (Lyu et al., 2025) assigns evidence validity scores $V_i \in [0, 1]$ to evidence–claim pairs and compatibility scores $G_{ij} \in [0, 1]$ to claim pairs. It then averages the two sets separately and combines them into an internal consistency score used in response evaluation. The WSI-LLaVA (Liang et al., 2025) evaluation framework also performs claim level scoring: claims extracted from reference answers are compared against generated responses for WSI-Precision, while claims extracted from generated responses are compared against reference answers for WSI-Relevance. Individual comparisons receive scores from $\{0, 0.3, 0.7, 1\}$ according to clinical agreement, and the scores are averaged to obtain the corresponding evaluation metric.

The numerical relation-validity probability can also be obtained from answer token likelihoods. BEACON (Wong et al., 2026) uses a frozen text reranker to assess whether a patch description supports a diagnostic hypothesis given the question. These support scores are subsequently normalized across hypotheses and used for belief construction and evidence acquisition in following steps. PRISM2 (Vorontsov et al., 2026) likewise formulates pathological findings and diagnostic classes as Yes/No questions conditioned on WSI embeddings. It scores candidate completions using the difference between image conditioned and text only log likelihoods, and derives prediction probabilities from the answer alternatives. This provides a particularly close precedent for our two-way Yes/No-logit relation-validity probability estimation.

Together, these works establish multiple precedents for using LLMs both to determine clinical relations and to assign numerical validity probability. ARGUE applies this recurring assessment pattern to proposed directed diagnostic relations, using normalized Yes/No logits to estimate their validity

probability. Our focus is on preserving these uncertain relations in a shared UEG and evaluating evidence pathways through diagnostic reachability.

A.9 UEG Construction Details

Finding nodes. For an inspected patch, the visual observer first produces a textual observation o. The observation is then decomposed into a set of atomic morphological findings: $\{f_1, \ldots, f_k\}, \quad 1 \leq k \leq 3$, such as "small round cells", "hyperchromatic nuclei", or "fibrous stroma". Importantly, only findings derived from patches that have actually been selected and inspected are committed to the persistent UEG. Hypothetical findings (Equation **??**) used during candidate reranking are instantiated only in temporary graph copies and are not added to $\mathcal{F}_t$.

Concept nodes. The initial concept set $\mathcal{C}_0$ is generated from the question, start from six neutral dimensions: architecture, cytology, stroma and vasculature, differentiation, proliferative activity and necrosis, and invasion. For each newly added finding, the LLM either identifies a compatible existing concept or proposes a new concrete concept, optionally linked to a parent concept node. Concepts therefore represent reusable morphological abstractions, while findings retain the actual observations from individual patches. Below we provide the prompt used throughout this process:

Concept Initialization

```
Given the question of the WSI, identify visible morphological
concepts that identify plausible answers.  Start from six neutral
dimensions:  architecture, cytology, stroma and vasculature,
differentiation, proliferative activity and necrosis, and invasion.
Describe what each concept means without assuming that it is
present in the slide.

{
  "Question":  "Do the neoplastic epithelial cells forming papillary
structures exhibit specific architectural features related to
papillary architecture?"
}
```

Finding Generation and Concept Expansion

```
Decompose the patch description into 1-3 atomic, observed
morphological findings.  For each finding, reuse compatible
existing concrete concepts or create new concepts.  Optionally
identify a broader parent concept.

{
  "Patch Description":  "The image shows well defined papillary
structures with a central core, exhibiting a branching pattern.
The neoplastic epithelial cells form a confined in-situ
architecture, maintaining an intact boundary."
  "Existing concepts:":  {Central core presence, Nuclei concentrated
towards the base, Stratified epithelial cells, ...}
}
```

Node merging and canonicalization. Every finding and concept string is canonicalized by trimming leading and trailing whitespace, collapsing consecutive whitespace characters, and converting the string to lowercase. Finding strings are mapped to the same node if and only if their canonicalized forms are identical. The same criterion applies to concept strings. When a finding matches an existing node, that node is reused together with its previously assigned concept and relation-validity

probability. Likewise, a concept matching an existing canonicalized concept reuses the concept node and its corresponding relation-validity probability.

Relation construction: finding-to-concept and concept-to-concept. To keep the UEG sparse, we do not construct all possible pairs between findings and concepts. As aforementioned, for each finding f_i, the LLM selects compatible concrete concepts, either by reusing an existing concept or by proposing a new concepts, yielding a candidate finding-to-concept edges. A concrete concept may additionally be linked to a broader parent concept c_k when it is a valid instance of that parent. Furthermore, non cyclic relations are retained. We keep sparse relations in the UEG for two reasons. First, sparsity reduces the computational complexity of reachability analysis by limiting the number of edges and candidate pathways that need to be traversed. Second, these sparsity constraints limit M, the number of unique uncertain relations entering the deviation term $\beta_{\mathcal{G}}$ in Proposition 2, thereby restricting error accumulation as the UEG grows. Moreover, the validity probability of each directed edge is estimated only once, with its validity state shared across all pathways containing it. Hence, the estimation error of a shared edge is accounted for once at the graph level rather than repeatedly multiplied by the number of overlapping pathways.

General Relation Construction

```
You are assessing one proposed directed relationship in a pathology
evidence graph.  Judge only the specified relationship.  For a
finding-to-concept edge, the finding must instantiate the concept
with the same subject, polarity, and relevant attributes.  For a
concept-to-concept edge, the concrete concept must validly support
the parent concept.  Answer Yes if the relationship is justified
and No otherwise.  Return exactly Yes or No.
```

```
{
  "Question":  "question."
  "Relationship:":  {finding-to-concept, concept-to-concept}
  "Premise:":  {Source finding/concept}
  "Target:":  {Target concept}
}
```

Relation construction: concept-to-terminal. Once a tentative answer is available, for each candidate concept, the model evaluates whether it should connect to the support or refute terminal of this answer. Support requires morphology that specifically strengthens the exact answer, whereas refute requires a concrete incompatible observation or a defining alternative pattern. Notably, concepts may receive no terminal edge if it belongs to broad neutral concepts.

Terminal Relation Construction

```
Given the fixed current answer and the observed ancestor evidence
of each candidate concept, propose a sparse set of terminal
relations.  Add an edge to the support terminal only when the
morphology specifically strengthens the exact answer.  Add an edge
to the refute terminal only when the morphology is incompatible
with the answer or supports a defining alternative.  Omit concepts
that have no valid terminal relation.  Answer Yes when the proposed
relationship IS justified; answer No when it is NOT justified.
Return exactly Yes or No.
```

```
{
  "Question":  "question."
  "Premise:":  {Source concept}
  "Target:":  {Target terminal}
}
```

Relation-validity probability estimation. The probability of relation-validity of each newly created edge is estimated independently. For an edge e, LLM evaluates the validity question as given in previous prompts with the two answer tokens "Yes" and "No". Let $\ell_{\text{yes}}(e)$ and $\ell_{\text{no}}(e)$ denote the corresponding logits. The estimated validity probability is given by:

$$p_e = \frac{\exp\left(\ell_{\text{yes}}(e)\right)}{\exp\left(\ell_{\text{yes}}(e)\right) + \exp\left(\ell_{\text{no}}(e)\right)}. \tag{17}$$

If an identical directed edge already exists in the UEG, its stored probability is retained and the validity probability of this edge is not reestimated. Consequently, each directed edge receives at most one *Yes*/*No* logit probability during a run.

Rationale of ARGUE's Relation-Validity Probabilities. In this section, we discuss the rationale of our relation-validity probability design. As reviewed in Section A.8, using LLMs to assess clinical relations and assign numerical confidence scores is common practice in medical reasoning and computational pathology. Following this practice, ARGUE estimates the validity probability of each proposed diagnostic relation through a two way softmax over the LLM's Yes/No logits, explicitly representing uncertainty in the UEG rather than treating every relation as certainly valid.

The importance of this design is supported by the probability assignment ablation in Figure 4. Replacing our LLM estimated relation-validity probabilities with randomly generated values or CONCH text similarity scores reduces both WSI-P and WSI-R. These results demonstrate the contribution of LLM based relation-validity estimation to downstream performance, supporting its use over arbitrary or similarity based probability assignments.

We further assess the reliability of these estimates through independent pathologist evaluation. We randomly sample 10 WSI-Bench questions from distinct WSIs and ask two independent pathologists to rate the plausibility of the relation-validity probabilities. As reported in Figure 7, the two pathologists assign scores of 4.38 and 3.88, respectively. These favorable ratings provide domain expert evidence for the clinical plausibility of our probability estimates.

Finally, recognizing that LLM estimated probabilities may still contain errors, we explicitly evaluate ARGUE's sensitivity to relation-validity probability perturbations. In Figure 6 (right), we perturb the probabilities of 5%, 10%, and 20% of randomly selected edges with Gaussian noise of variance 0.5. These perturbations produce only modest performance degradation: even when 20% of edges are perturbed, ARGUE retains WSI-P/WSI-R of 0.405/0.459, compared with 0.423/0.461 without perturbation. Thus, although individual relation-validity estimates can be imperfect, ARGUE remains robust to the tested probability perturbations without a substantial loss of downstream performance.

Answer updates and terminal refresh. When the working answer changes from a_t to a_{t+1}, applicable finding and concept nodes together with their corresponding evidence pathways are retained. The two answer relative terminals are replaced by a^s_{t+1} and a^r_{t+1} respectively, and the concept-to-terminal classification is recomputed for the new answer. This sequential revision allows alternative plausible answers to emerge as evidence accumulates, without requiring candidate diagnoses to be enumerated in advance, which may not always be readily available in all scenarios.

Hypothetical candidate extensions. At each acquisition round, the structured LLM call produces a set of evidence gap descriptions $\xi_{t,1}, \ldots, \xi_{t,M_t}$. For each candidate $x_p \in \mathcal{K}_t$, the description attaining the maximum image–text similarity in Equation 4 is taken as the evidence gap associated with x_p, where the frozen CONCH embedding of x_p is used only for this matching and retrieval step, the raw candidate image is not passed to the visual observer during reranking. Only after a candidate is selected does the visual observer process its raw image.

To construct $\mathcal{G}_t \oplus x_p$, ARGUE temporarily treats the matched evidence gap description as a hypothetical finding and integrates it into a copy of the current UEG. The system applies the same procedure (sparse relation proposal, relation-validity probability estimation, etc.) used for ordinary UEG updates. The resulting $\mathcal{G}_t \oplus x_p$ is evaluated using the Monte Carlo estimator in Section 3.2 to compute the predicted refutation reachability gain, $\hat{g}_t(p) = \hat{R}_r(\mathcal{G}_t \oplus x_p, a_t) - \hat{R}_r(\mathcal{G}_t, a_t)$, conditional on confirming the matched evidence gap. This gain is combined with evidence gap relevance to rerank candidate patches in Equation **??**. Its maximum over the evaluated candidate pool is also used, together with inter round reachability stability, to assess acquisition sufficiency in Equation 5,

checking whether further inspection could appreciably strengthen a refutation pathway. These extensions remain temporary: only findings derived from actual visual inspection of selected patches are incorporated into the persistent UEG.

A.10 Additional Prompts

Evidence gap description generation

```
You are the supervisor of a pathology search agent.  Identify the
missing visible morphological features, especially those indicate
an evidence gap that could provide the evidence needed to answer
the current question.  Return exactly one JSON object containing
concise histopathology image-text search queries.

{
  "Question":  "question"
  "Working answer:" {"working_answer"}
  "Evidence Graph:" {"evidence_graph_json"}
}
```