DON'T TRUST, VERIFY: MULTI-AGENT FRAMEWORK WITH ADVERSARIAL EVIDENCE CROSS-EXAMINATION FOR WHOLE SLIDE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole slide image (WSI) analysis is increasingly approached by agentic systems that emulate a pathologist's workflow—surveying a slide and zooming into suspicious regions to accumulate evidence. We argue that this exploratory paradigm shares a fundamental blind spot: it over-trusts the perceptions of underlying multimodal large language models (MLLMs) and seeks to improve performance by gathering additional visual observations, even though navigating to different regions often cannot recover features that the model cannot reliably resolve. We introduce ReMIL-Agent, a training-free multi-agent framework operating along an orthogonal verification axis based on multiple instance learning (MIL), in which each MLLM-reported feature is treated as evidence to be actively cross-examined rather than accepted at face value. A cascade mechanism then revisits an initial prediction, subjects its supporting morphological cues to an adversarial Defender–Critic process grounded in a pathology knowledge base, and selectively re-examines key evidence across scales only when warranted, ultimately retreating to uncertain at the limits of vision-language model (VLM) capability. Finally, a reviewer module performs multi-scale reasoning to produce interpretable and auditable pathological reports. Empirical results demonstrate that ReMIL-Agent is a generalizable workflow agent framework that can be built on top of existing MIL methods, enabling more transparent reasoning with explicit capability boundaries and further advancing trustworthiness for clinical pathologists.

1 INTRODUCTION

Whole-slide images (WSIs) are the gold standard for cancer subtyping, yet their gigapixel scale forces models to reason from sparse, locally-sampled evidence. Two paradigms dominate. The first treats WSI classification as weakly-supervised multiple instance learning (MIL), aggregating patch features into a slide-level prediction Vaswani et al. (2017); Shao et al. (2021); Lu et al. (2021), often atop self-supervised foundation encoders Chen et al. (2024). MIL is accurate but opaque: it emits a label without verifiable evidence. The second equips vision-language models (VLMs) pretrained on pathology corpora Huang et al. (2023); Lu et al. (2024a) to describe morphology and generate diagnoses directly. However, their perceptions are unreliable at the resolutions WSIs demand, and a fluent description is not a verified one.

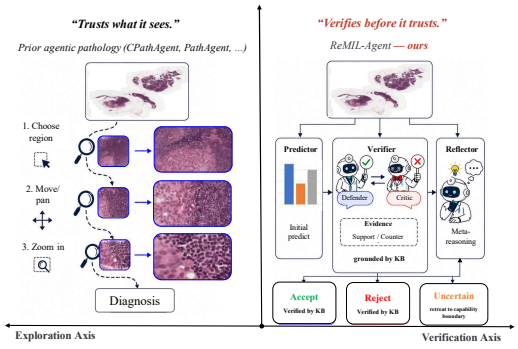

Figure 1: Two orthogonal axes for pathology agents. The histopathology images shown in both panels are sampled from the CAMELYON16 WSI `tumor_018` Ehteshami Bejnordi et al. (2017).

A recent line of work casts diagnosis as an agentic process that emulates how pathologists examine a slide—surveying tissue at low magnification, then zooming into suspicious regions and integrating

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.

evidence across scales Sun et al. (2026); Wang et al. (2025); Chen et al. (2026). We refer to this paradigm as the **exploration axis**, in which an agent determines where to look next and accumulates evidence by acquiring additional observations. However, exploration cannot fundamentally remedy perceptual failures. When the underlying VLM is unable to reliably resolve a discriminative feature at the available level of detail, repeatedly revisiting the same region does not make that feature any more observable; instead, the agent may confidently report evidence that it cannot genuinely perceive.

We argue that what agentic pathology lacks is not better navigation, but a **verification axis**, as illustrated in Fig. 1. We introduce ReMIL-Agent, a training-free multi-agent framework that operates along an axis orthogonal to exploration. Rather than trusting the VLM's perceptions and seeking additional observations, ReMIL-Agent treats every VLM-reported feature as evidence to be rigorously verified. ReMIL-Agent instantiates this principle through three sequential stages: Predictor, Verifier, and Reviewer. The Predictor first produces an initial slide-level diagnosis together with a confidence estimate, reusing any off-the-shelf MIL model without retraining. Rather than uniformly escalating every slide to costly agentic reasoning, ReMIL-Agent invokes verification only when necessary: predictions with low confidence are forwarded to the Verifier, which stages a pathology knowledge-guided dual-agent debate between a Defender and a Critic, cross-examining the supporting and contradictory evidence for the initial prediction under the grounding constraints of Pathology Knowledge Base (PathKB). This adversarial cross-examination yields one of three deterministic verdicts—ACCEPT, REJECT, or UNCERTAIN—rather than a forced binary decision. Cases that remain UNCERTAIN are passed to the Reviewer, which performs targeted, evidence-guided multi-scale re-examination to resolve the residual ambiguity, ultimately producing a final diagnosis together with an interpretable, auditable report. This verification-first paradigm keeps ReMIL-Agent training-free and model-agnostic with respect to the underlying MIL predictor, while substantially improving diagnostic accuracy, efficiency, and interpretability over prior exploration-oriented agentic pathology systems.

- We reframe agentic pathology along a verification axis and instantiate it as a training-free multi-agent cascade with adversarial, KB-grounded evidence cross-examination.
- We introduce explicit capability-boundary modeling with retreat-to-UNCERTAIN, and a planner-guided Reviewer resolves only those cases by iteratively re-examining the Defender's and Critic's evidence across scales, revising the prediction only when the updated beliefs warrant it.
- Experimental results demonstrate that our framework is readily applicable to existing instance-based MIL methods, improving predictive performance while producing validated multi-scale evidence chains for trustworthy diagnosis.

2 RELATED WORK

Pathology-specific Multimodal Large Language Models. Emerging multimodal large language models (MLLMs) enable more flexible pathology image understanding by integrating visual perception with language reasoning. PathChat Lu et al. (2024b) and PathChat+ Weishaupt et al. (2025) couple pathology foundation encoders with large language models (LLMs) for diagnostic assistance; Quilt-LLaVA Seyfioglu et al. (2024) explores pathology-oriented instruction tuning, while SlideChat Chen et al. (2025) extends multimodal reasoning to WSI-level analysis. PathGen-LLaVA Sun et al. (2025) investigates pathology report generation, and KEEP Zhou et al. (2026) incorporates disease ontologies for knowledge-enhanced subtype recognition. More recently, Patho-R1 Zhang et al. (2026) integrates explicit diagnostic reasoning into pathology MLLMs. Despite this growing capability to combine perception with language-guided reasoning, existing methods typically treat the generated diagnosis or report as final, with little effort devoted to verifying whether the underlying visual evidence actually supports the conclusions.

Agentic Pathology AI. Agentic systems have recently emerged as a promising paradigm for computational pathology, where LLMs act as planners, tool users, and decision-making agents rather than mere predictors. Early efforts decomposed diagnostic workflows into specialized agents: PathFinder Ghezloo et al. (2025) emulates pathologists' workflows via coordinated triage, navigation, description, and diagnosis agents, while SlideSeek Weishaupt et al. (2025) and WSI-Agents Lyu et al. (2026) extend this with collaborative multi-agent reasoning and hierarchical, multimodal evidence collection for gigapixel WSIs. More recent methods emphasize adaptive exploration—PathAgent Chen et al.

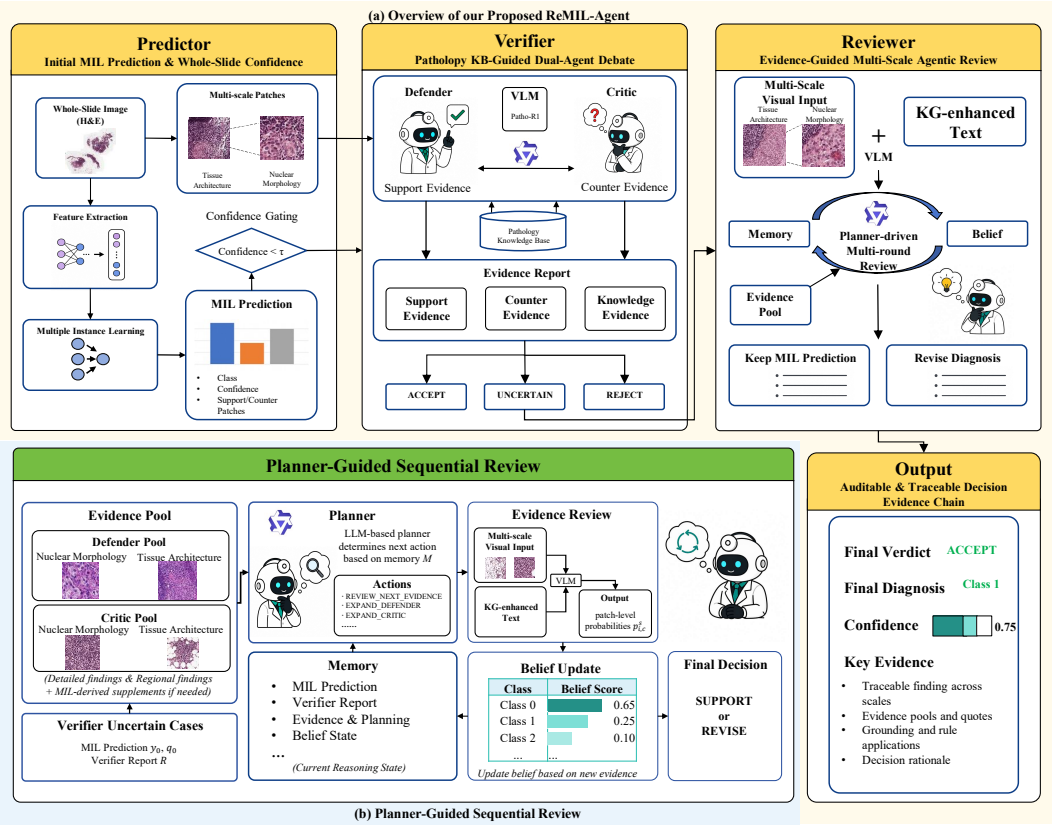

Figure 2: Overview of ReMIL-Agent. (a) Overall workflow consisting of the Predictor, PathKB-grounded Defender–Critic Verifier, and evidence-guided Reviewer. (b) The Reviewer performs planner-guided multi-round review over multi-scale evidence and updates class beliefs to support or revise the initial prediction.

(2026) enables training-free, interpretable WSI analysis through iterative navigation and reasoning, while PathFound Hua et al. (2025) and CPathAgent Sun et al. (2026) actively acquire additional evidence via multi-scale exploration to refine diagnostic hypotheses. These methods share a common exploration-oriented premise: better diagnosis follows from seeing more. ReMIL-Agent departs from this premise—rather than asking "where else should I look?", it first asks "can I trust what has already been seen?", gating further review behind verification so that additional computation is allocated only to diagnostically ambiguous cases.

3 METHOD

As shown in Fig. 2(a), ReMIL-Agent comprises three sequential stages: Predictor, Verifier, and Reviewer. The Predictor generates an initial diagnosis with a confidence estimate. Low-confidence predictions are then examined by the Verifier through a pathology knowledge-guided dual-agent debate over the supporting evidence. Cases that remain UNCERTAIN are passed to the Reviewer, which performs evidence-guided, multi-scale agentic review to refine the diagnosis and produce a final prediction with an interpretable, auditable report.

3.1 PREDICTOR: STRUCTURED EVIDENCE CONSTRUCTION

Given a WSI $\mathcal{W}$, the Predictor produces an initial diagnosis together with structured candidate evidence for downstream verification. Patch-level representations are extracted using the pathology foundation model UNI Chen et al. (2024) and aggregated via an instance-based MIL classifier, yielding a class distribution $\mathbf{p} \in \mathbb{R}^K$, prediction $\hat{y} = \arg\max_c p_c$, and confidence $p = p_{\hat{y}}$, which is

passed to the Verifier as the prior setting its decision regime. The MIL model additionally assigns each patch an importance score $\mathbf{a} = \{a_1, \ldots, a_N\}$, from which we build three complementary evidence pools: the highest-importance patches form the *support* set $\mathcal{S}$ (regions favoring $\hat{y}$), the lowest-importance patches form the *counter* set $\mathcal{C}$ (weakly supported regions offering alternative cues), and farthest-point sampling (FPS) over the remaining patches in UNI feature space yields the *context* set $\mathcal{X}$, preserving surrounding morphology and broadening visual coverage:

$$(\mathcal{S}, \mathcal{C}, \mathcal{X}) = \big(\text{Top-}k_{\mathcal{S}}(\mathbf{a}),\ \text{Bot-}k_{\mathcal{C}}(\mathbf{a}),\ \text{FPS}_{k_{\mathcal{X}}}(\mathcal{F} \setminus (\mathcal{S} \cup \mathcal{C}))\big). \tag{1}$$

Coupling importance ranking with FPS diversity retains discriminative regions while avoiding redundant concentration on similar patches; each selected patch retains both its feature and spatial coordinates for later multi-scale re-materialization. Together, $(\hat{y}, p, \mathcal{S}, \mathcal{C}, \mathcal{X})$ form the evidence pool consumed by the Verifier, enabling evidence-centered reasoning rather than reliance on the raw MIL prediction.

3.2 Verifier: Adversarial Cross-examination of Multi-scale Evidence

Given the Predictor output $(\hat{y}, p, \mathcal{S}, \mathcal{C}, \mathcal{X})$, the Verifier re-examines the flagged regions and returns $v \in \{\text{ACCEPT}, \text{REJECT}, \text{UNCERTAIN}\}$, overturning $\hat{y}$ only on positive evidence for a competing class. Each flagged patch is materialized at two scales—region (tissue architecture) and detail (cellular cytology)—and converted into structured, quote-backed evidence: a pathology VLM answers a (K+1)-way multiple-choice question and provides a free-text description under scale-specific prompts, which a text LLM parses into a per-feature dictionary, with unmentioned features labeled UNCLEAR and every positive finding backed by a verbatim quote. The Defender and Critic each summarize evidence for their respective positions and assign an overall evidence strength, with the final verdict determined by their relative difference. The framework rests on two components: a capability-aware PathKB that grounds observations and suppresses VLM hallucinations, and an adversarial dual-agent cross-examination mechanism that contrasts supporting and competing evidence.

Grounded PathKB Construction. PathKB is automatically generated by a frontier LLM (e.g., Gemini-3.5-Flash Team et al. (2023)) from the diagnostic criteria in the WHO Classification of Tumours Kleihues et al. (2000), then refined via capability and consistency constraints so every assertion is both clinically valid and VLM-observable. It comprises two components. The *pathognomonic feature block* defines each subtype's positive morphological features, specifying both what each feature is and the scale at which it is reliably observed, grounded by two mechanisms:

- **Cross-scale Consistency.** Each subtype is defined by a region-level architectural pattern confirmed by a detail-level cytological cue (e.g., glandular architecture in lung adenocarcinoma (LUAD) confirmed by intracellular mucin, and keratinizing nests in lung squamous cell carcinoma (LUSC) confirmed by keratinization); region-level evidence without a matched detail-scale confirmer is treated as unconfirmed.
- **Same-scale Exclusivity.** Mutually incompatible features (e.g., mucinous differentiation and keratinization) reported within the same field of view invalidate one another and are excluded from decision making.

These constraints prevent isolated, cross-scale-inconsistent, or contradictory VLM observations from being treated as evidence. The complementary *rule block* encodes diagnostic contrasts, morphological mimics, and evidence adjudication rules.

Adversarial Dual-agent Verification. The verdict results from an adversarial confrontation between two sub-agents reading disjoint evidence: a Defender inspects the support set $\mathcal{S}$ and argues for $\hat{y}$ via pathognomonic features of the predicted class, while a Critic inspects the counter and context sets $\mathcal{C} \cup \mathcal{X}$ and argues against it via pathognomonic features of a competing class. Reported evidence is first purified against PathKB, removing unconfirmed cross-scale observations, incompatible co-occurrences, and merge/split inconsistencies, then aggregated into support and opposition strengths, from which deterministic, regime-aware rules determine the final verdict. This confront-then-adjudicate design enables complementary evidence discovery while ensuring that the final decision remains transparent, deterministic, and fully auditable. Finally, the Verifier output the auditable report $\mathcal{R}$ that contains complementary evidence pools at each scale s:

- Defender Pool ($\mathcal{P}_D^s$): Pathognomonic and supporting findings favoring the initial prediction y_0.
- Critic Pool ($\mathcal{P}_C^s$): Prioritized strong contradictory findings challenging y_0.

3.3 REVIEWER: PROGRESSIVE MULTI-SCALE EVIDENCE REASONING

The Reviewer leverages the Verifier's structured evidence to progressively re-examine localized observations across scales, revising the original MIL prediction only when accumulated evidence consistently supports an alternative diagnosis. As shown in Fig. 2(b), it comprises two components: a *planner-guided review* that prioritizes evidence pools and determines whether the sequential review should continue, expand, or stop, and an *all-class belief arbitration* that integrates reviewed observations and performs conservative prediction revision. Algorithmic details are provided in Appendix.

Planner-guided Review. To preserve spatial correspondence, region-level evidence is retained directly while detail-level findings are mapped to their associated regions; when $\mathcal{R}$ provides insufficient localized evidence, the pools are supplemented using MIL-derived coordinates $\mathcal{C}_{\text{MIL}}$ to ensure visual coverage. A compact review memory is initialized to capture the diagnostic context,

$$\mathcal{M}_0 = \big(y_0, p_0, \mathrm{Stat}(\mathcal{R}), \mathrm{Cons}(\mathcal{R})\big), \tag{2}$$

where p_0 is the initial prediction confidence and $\mathrm{Stat}(\cdot)$, $\mathrm{Cons}(\cdot)$ denote localized evidence statistics and multi-scale consistency, respectively. An LLM-based planner Π_{plan} reads $\mathcal{M}_0$ to formulate initial diagnostic hypotheses $\mathcal{H}_0$ and set the review priority order $\mathcal{O}_0 \in \{\langle \mathcal{P}_D, \mathcal{P}_C \rangle, \langle \mathcal{P}_C, \mathcal{P}_D \rangle\}$:

$$(\mathcal{H}_0, \mathcal{O}_0) = \Pi_{\text{plan}}(\mathcal{M}_0), \tag{3}$$

examining the prioritized pool first, followed by its complement, to maintain a consistent diagnostic focus.

The review then proceeds over progressively enlarged evidence batches. Let $\mathcal{E}_t$ denote the accumulated evidence state at round t and $B_t \in \mathbb{R}^{|\mathcal{Y}|}$ the updated all-class belief table; the compact planning state is

$$\mathcal{S}_t = \big(t, B_t, \mathcal{E}_t, N_{\text{budget}}\big), \tag{4}$$

where N_{budget} is the remaining evidence budget. Guided by $\mathcal{S}_t$, the planner outputs a discrete decision

$$a_t = \Pi_{\text{plan}}(\mathcal{S}_t), \quad a_t \in \{\textsc{Continue}, \textsc{Expand}, \textsc{Terminate}\}, \tag{5}$$

determining whether the Reviewer proceeds with additional ranked evidence, expands the Defender or Critic pool, or halts the sequential process. Under expansion, spatial neighbors $\mathcal{N}(\mathcal{E}_t)$ are introduced around the current evidence anchors to distinguish isolated artifacts from consistent pathological patterns:

$$\mathcal{E}_{t+1} \leftarrow \mathcal{E}_t \cup \mathcal{N}(\mathcal{E}_t). \tag{6}$$

For each reviewed observation—an individual patch or region within an evidence pool—we use KEEP Zhou et al. (2026) to perform multi-scale evidence assessment under task-specific textual prompts $\mathcal{T}$, evaluating multiple prompt configurations and aggregating the top-ranked ones to reduce sensitivity to prompt variation. Let $P_{\text{det}}(c \mid v, \mathcal{T})$ and $P_{\text{reg}}(c \mid v, \mathcal{T})$ denote the detail- and region-scale prediction distributions for observation v; these are fused via confidence-aware weights w_s into a per-observation evidence score,

$$r_v(c) = \sum_{s \in \{\text{det,reg}\}} w_s \cdot P_s(c \mid v, \mathcal{T}), \tag{7}$$

the class-c evidence contributed by observation v to the belief update below.

All-class Belief Arbitration. To avoid limiting the review to a single challenger, we accumulate evidence across all candidate classes $c \in \mathcal{Y}$. Let $\mathcal{O}_t$ denote the observation set at round t, drawn from the evidence state $\mathcal{E}_t$ and partitioned into primary Verifier-guided observations (Defender, Critic, and their expansions). The accumulated support score, which directly updates the belief table $B_t(c)$, is

$$S_t(c) = \sum_{v \in \mathcal{O}_t} \rho_v \, r_v(c) + \lambda K^{+} \mathbb{I}(c = y_0), \tag{8}$$

Method	BRCA	NSCLC	RCC	ESCA	BLCA	THCA	Avg.
Conventional MIL Methods							
DSMIL Li et al. (2021)†	84.70	87.20	88.90	89.90	52.20	57.80	76.78
ABMIL Ilse et al. (2018)	88.38	93.76	94.39	97.62	53.87	70.67	83.12
Flow-MIL Ma et al. (2025)	86.93	93.61	91.81	95.24	65.43	74.69	84.62
General-purpose & Medical MLLMs							
Gemini-2.5-Pro-Preview-03-25 Comanici et al. (2025)†	72.40	89.20	69.20	97.10	59.80	44.90	72.10
Quilt-LLaVA Seyfioglu et al. (2024)†	55.30	54.00	53.30	58.60	61.90	45.80	54.82
PathGen-LLaVA Sun et al. (2025)†	66.50	53.50	59.00	78.60	54.70	49.20	60.25
KEEP Zhou et al. (2026)	71.96	85.05	93.46	88.10	49.21	45.54	72.22
MedGemma-1.5-4B Sellergren et al. (2025)	50.01	53.36	33.04	55.02	47.13	39.71	46.38
MedGemma-1.5-27B Sellergren et al. (2025)	47.33	54.40	40.49	62.50	49.20	33.47	47.90
Patho-R1-7B Zhang et al. (2026)	61.84	66.79	49.88	80.71	46.47	31.68	56.23
Qwen3.5-4B Yang et al. (2025)	61.93	56.33	46.30	80.56	48.18	48.30	56.93
Qwen3.5-9B Yang et al. (2025)	66.77	58.97	61.81	68.33	50.05	37.59	57.25
Qwen3.6-27B Yang et al. (2025)	56.19	63.17	56.65	85.00	55.32	43.83	60.03
Agent-based Methods							
PathAgent Chen et al. (2026)‡	61.08	62.23	48.22	68.10	52.13	40.04	55.30
CPathAgent Sun et al. (2026)†	88.50	90.80	94.60	97.10	62.70	63.20	82.82
ReMIL-Agent (ABMIL)	**92.83**	94.63	**95.64**	**100.00**	57.79	72.71	85.60
ReMIL-Agent (Flow-MIL)	88.41	**94.68**	94.20	97.62	**71.78**	**76.31**	**87.17**

Table 1: Balanced accuracy (BACC, %) on six TCGA subtyping benchmarks Weinstein et al. (2013). †: reported from CPathAgent Sun et al. (2026). ‡: results reproduced by us using the publicly available open-source code. MedGemma-1.5 Sellergren et al. (2025) uses majority voting; Patho-R1 Zhang et al. (2026) and Qwen3.x Yang et al. (2025) use multi-image inference. Best/second-best in bold/underline.

where K^+ is PathKB-based evidence from Verifier supporting y_0, and ρ_v is a reliability weight.

From the accumulated scores, we identify the strongest alternative hypothesis and its margin over the original prediction:

$$\tilde{y} = \underset{c \in \mathcal{Y},\, c \neq y_0}{\arg\max}\, S_t(c), \qquad m_t(\tilde{y}, y_0) = S_t(\tilde{y}) + \lambda K^- - S_t(y_0), \tag{9}$$

where K^- is PathKB-based evidence from Verifier contradicting y_0. Maintaining beliefs over the complete label space, rather than only y_0 and the Verifier's initial challenger, lets us recover alternative hypotheses that verification alone may have overlooked.

The belief state is evaluated after every round: if evidence remains insufficient, the arbiter returns UNCERTAIN and Π_{plan} continues or expands the review; once sufficient, it returns SUPPORT or REVISE and terminates. If the primary review stays inconclusive, we optionally search for discriminative regions of alternative classes and fold the resulting observations into S_t. The original prediction is replaced by $\tilde{y}$ only when $m_t(\tilde{y}, y_0)$ exceeds a margin threshold with cross-scale-consistent support; otherwise y_0 is conservatively preserved.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

4.1.1 DATASETS

We evaluate the proposed framework on seven public WSI benchmarks, including six TCGA cancer subtype classification tasks Weinstein et al. (2013): TCGA-BRCA (breast cancer), TCGA-NSCLC (non-small cell lung cancer), TCGA-RCC (renal cell carcinoma), TCGA-ESCA (esophageal cancer), TCGA-BLCA (bladder cancer), and TCGA-THCA (thyroid cancer), as well as the CAMELYON16

Type	Method	BACC (%)
MIL	Flow-MIL Ma et al. (2025)	91.18
	ABMIL Ilse et al. (2018)	<u>93.59</u>
MLLM	KEEP Zhou et al. (2026)	75.60
	Patho-R1-7B† Zhang et al. (2026)	51.92
Agent	PathAgent Chen et al. (2026)	58.22
	ReMIL-Agent (ABMIL)	**96.79**

Table 2: BACC (%) on CAMELYON16. † Multi-image inference. Best and second-best in **bold** and <u>underline</u>.

Method	BRCA	NSCLC	RCC	ESCA	BLCA	THCA	Avg.
ABMIL Ilse et al. (2018)							
Predictor	88.4	93.8	94.4	97.6	53.9	70.7	83.1
w/o Rev.	89.4	94.1	94.4	**100.0**	56.2	71.1	84.2
w/o Ver.	85.3	92.1	93.7	85.7	49.2	70.1	79.4
Full	**92.8**	**94.6**	**95.6**	**100.0**	**57.8**	**72.7**	**85.6**

Table 3: Component ablation with ABMIL on six TCGA subtyping benchmarks (BACC, %). Best results are shown in bold.

benchmark Ehteshami Bejnordi et al. (2017). Following previous studies, all TCGA WSIs are obtained from the HistGen benchmark Guo et al. (2024), which provides standardized training and testing splits for slide-level pathology classification. For CAMELYON16 Ehteshami Bejnordi et al. (2017), we follow the dataset split and evaluation setting adopted in DSMIL Li et al. (2021). More details are provided in Appendix.

4.1.2 Implementation Details

Our framework is implemented in PyTorch and performs training-free inference without retraining the underlying MIL classifiers. We use Flow-MIL Ma et al. (2025) and ABMIL Ilse et al. (2018) as the MIL backbones to provide the initial slide-level predictions. For the Verifier, we employ Patho-R1 Zhang et al. (2026) and Qwen3-4B Yang et al. (2025) to perform pathology-oriented evidence verification and uncertainty assessment. For the Reviewer, we utilize KEEP Zhou et al. (2026) for multi-scale visual evidence review and Qwen3-4B Yang et al. (2025) as the LLM planner for sequential evidence scheduling. The multi-scale review is performed at $20\times$ magnification, using 4096×4096 patches for region-level tissue architecture analysis and 224×224 patches for detail-level cellular morphology analysis. To improve inference efficiency, visual features are extracted offline and stored, while the task-specific text representations derived from the textual prompts are precomputed. All inference experiments are conducted on a single GPU with 80 GB of memory. Following CPathAgent Sun et al. (2026), we employ Balanced Accuracy (BACC) as the evaluation metric to mitigate class imbalance in specific datasets.

4.2 Results Analysis

Table 1 reports BACC across six TCGA benchmarks Weinstein et al. (2013) against conventional MIL methods, general-purpose/medical MLLMs, and agent-based baselines. ReMIL-Agent consistently improves both underlying MIL predictors without retraining, with gains spread across nearly all datasets rather than concentrated in a single cohort, confirming that it complements rather than replaces the underlying MIL architecture. This pattern follows directly from the framework's selective-reasoning design: predictions already well-supported by evidence (e.g., TCGA-ESCA) largely retain their strong baseline performance, while diagnostically ambiguous cohorts (e.g., TCGA-BLCA, TCGA-BRCA) see the largest gains, indicating that the benefit stems from improved evidence reliability rather than a specific backbone. Standalone general-purpose and medical MLLMs remain substantially behind MIL-based methods, even the strongest such baseline, KEEP Zhou et al. (2026), motivating our division of labor between MIL for robust slide-level prediction and VLMs for knowledge-guided verification. Against the representative agent-based baseline CPathAgent Sun et al. (2026), ReMIL-Agent achieves the best overall performance among all methods compared. Additionally, Table 2 reports BACC on CAMELYON16 Ehteshami Bejnordi et al. (2017), where ReMIL-Agent achieves the best performance of 96.79%, outperforming conventional MIL methods, medical MLLMs, and the agent-based baseline PathAgent Chen et al. (2026). Overall, these results support our central claim: verification-first reasoning, rather than further exploration, offers a more reliable and generalizable strategy for both WSI cancer subtyping and metastasis detection.

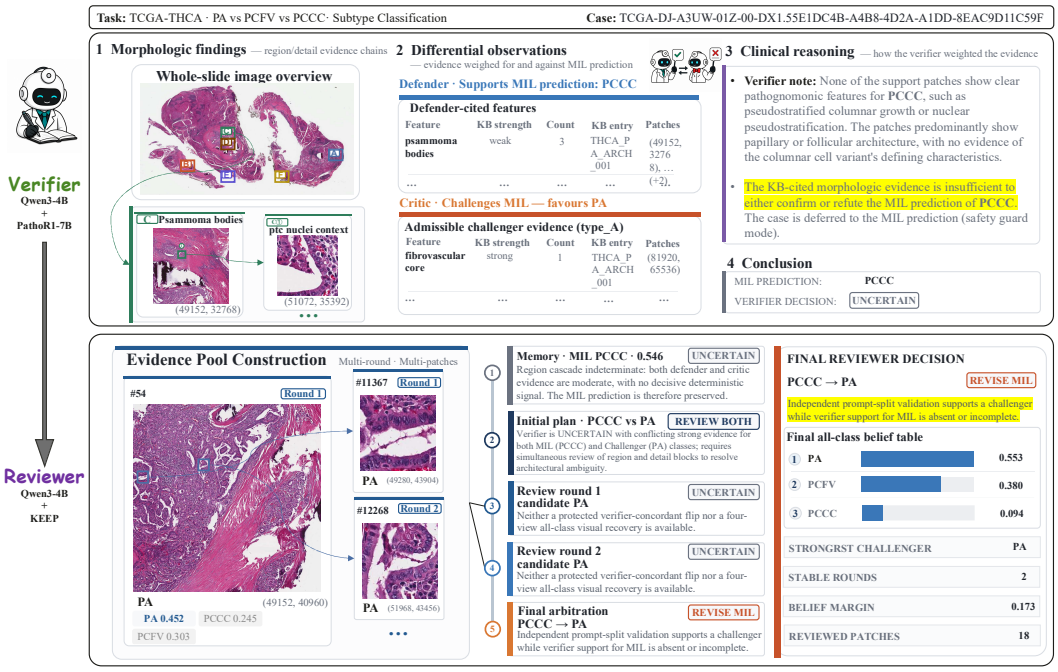

Figure 3: Case study on the TCGA-THCA subtype classification task. The Verifier returns UNCERTAIN after multi-scale morphologic verification due to insufficient PCCC-specific evidence, while the Reviewer performs sequential evidence review and revises the initial PCCC prediction to PA.

Verifier	BRCA	NSCLC	RCC	ESCA	BLCA	THCA	Avg.
w/o PathKB	88.38	93.36	94.39	72.86	53.87	68.53	78.57
w/ PathKB	89.42	94.13	94.39	100.00	56.20	71.09	84.21

Table 4: Ablation study of PathKB in the Verifier module.

4.3 Ablation Studies

Ablation of the Verifier and Reviewer. Table 3 examines the contributions of the Verifier and Reviewer using ABMIL as the Predictor; corresponding results for Flow-MIL are reported in Appendix Table 10. Removing the Reviewer generally reduces performance across both backbones, supporting the value of progressive multi-scale review for cases that remain UNCERTAIN after verification. The degradation is particularly evident on TCGA-BLCA, where the Reviewer provides substantial gains by assessing the structured evidence produced by the Verifier. Removing the Verifier leads to a larger performance drop, especially on TCGA-BLCA and TCGA-ESCA. This suggests that applying agentic review directly to unverified MIL-selected evidence may amplify inaccurate or incomplete observations. Removing both modules recovers the original MIL predictor. The complete ReMIL-Agent achieves the highest average performance with both backbones, highlighting the complementary contributions of the two modules and supporting verification before review. Additional ablation studies are provided in the appendix.

Ablation of PathKB. We further evaluate the contribution of PathKB within the Verifier. All experiments use ABMIL Ilse et al. (2018) as the Predictor and Patho-R1 Zhang et al. (2026) with Qwen3-4B Yang et al. (2025) as the VLM–LLM configuration, with the Reviewer disabled to isolate the effect of PathKB. As shown in Table 4, removing PathKB decreases the average BACC from 84.21% to 78.57%, a drop of 5.64 percentage points. The degradation is particularly pronounced on TCGA-ESCA, suggesting that PathKB provides useful domain-specific guidance for filtering inconsistent observations and grounding VLM-generated evidence.

Dataset	Method	BACC	Avg. Time	Total Time
BLCA	PathAgent	52.13	141.06	11160.22
	ReMIL-Agent	**57.79**	**20.67**	**1633.27**
THCA	PathAgent	40.04	140.68	16761.19
	ReMIL-Agent	**72.71**	**45.06**	**5362.08**

Table 5: System-level performance and efficiency comparison between PathAgent Chen et al. (2026) and our proposed ReMIL-Agent. BACC is reported as a percentage (%), while average and total time are reported in seconds (s). Average time is computed over all WSIs in each dataset.

4.4 Efficiency Analysis

We compare ReMIL-Agent (with ABMIL Ilse et al. (2018) as the default predictor) against PathAgent Chen et al. (2026) on TCGA-BLCA (binary) and TCGA-THCA (three-class), covering both binary and multi-class settings. As shown in Table 5, ReMIL-Agent consistently achieves higher accuracy while substantially reducing inference cost on both datasets. We further break down per-stage cost in Appendix. The Predictor runs on every WSI; only low-confidence cases reach the Verifier, and the Reviewer is activated only for Verifier outputs labeled UNCERTAIN. We report average time per WSI (each stage's overall system contribution) and per processed WSI (cost of a single invocation). Since only a small, further-shrinking fraction of WSIs is escalated at each stage, the Reviewer adds limited dataset-level overhead while substantially improving accuracy—confirming that confidence-gated, verification-first routing improves efficiency without sacrificing performance.

4.5 Case Studies

Figure 3 shows a representative case from TCGA-THCA subtyping, which separates papillary adenocarcinoma (PA), papillary carcinoma follicular variant (PCFV), and papillary carcinoma columnar cell (PCCC). The MIL predictor assigns PCCC at a confidence of 0.546. Examining region- and detail-scale patches, the Verifier finds papillary fronds with fibrovascular cores and follicular components, but none of the features that define PCCC–pseudostratified columnar growth or nuclear pseudostratification. PA emerges as the leading challenger, yet the accumulated evidence remains too thin to overturn the prediction outright, so the Verifier declines to force a call and returns UNCERTAIN. The Reviewer then arbitrates PCCC against PA over the evidence pools, confirms morphology consistently favoring PA, recovers no PCCC-specific support, and, once the conservative confirmation policy is satisfied, emits REVISE_MIL and corrects the prediction to PA. The correction turns on what the evidence fails to establish: PCCC is overturned not because further looking revealed PA, but because the features that would license PCCC were never confirmed. Additional cases illustrating other verification and correction patterns are provided in Appendix.

5 Conclusion

In this work, we presented ReMIL-Agent, a training-free multi-agent framework that introduces a verification-oriented paradigm for WSI analysis. ReMIL-Agent integrates a MIL-based Predictor, a PathKB-grounded Defender–Critic Verifier, and a planner-guided multi-scale Reviewer. Instead of directly trusting MLLM-generated observations, the framework explicitly verifies supporting and contradictory evidence, conservatively revises the initial prediction only when consistent alternative evidence is identified, and preserves uncertainty when the available visual evidence is insufficient. Experiments on six TCGA cancer subtyping benchmarks Weinstein et al. (2013) and CAMELYON Ehteshami Bejnordi et al. (2017) demonstrate that ReMIL-Agent consistently improves MIL without task-specific retraining, while outperforming representative MLLMs and pathology agents. The ablation results further confirm that verification should precede review. The efficiency analysis further shows that the conditional cascade substantially reduces inference cost compared with exploration-oriented agents. Case studies illustrate how ReMIL-Agent corrects erroneous MIL predictions through consistent multi-scale evidence.

AI USE STATEMENT

We used GPT to polish selected passages and GPT-Image 2 to generate the robot icons in Figures 1 and 2; all remaining elements of these figures were created manually in Microsoft PowerPoint without AI assistance. PathoKB was constructed through a draft–audit–refine process: Gemini-3.1-Flash drafted entries grounded in the diagnostic criteria of the WHO Classification of Tumours and revised them following two rounds of automated auditing by Claude Opus 4.8. Claims not traceable to these criteria were revised or discarded. Both models were used exclusively for knowledge-base construction, not inference. The authors take responsibility for the final manuscript and all AI-assisted content.

REPRODUCIBILITY STATEMENT

Our experiments use publicly available TCGA and CAMELYON datasets and open model weights of Patho-R1 and Qwen3.5. The PathoKB construction procedure, including its drafting, auditing, and refinement stages, is documented in the appendix. To facilitate reproducibility, we will release the source code publicly on GitHub upon acceptance.

REFERENCES

Jingyun Chen, Linghan Cai, Zhikang Wang, Yi Huang, Songhan Jiang, Shenjin Huang, Hongpeng Wang, and Yongbing Zhang. Toward interpretable analysis of whole-slide pathology images via large language model-based agentic reasoning. In *European Conference on Computer Vision*. Springer, 2026.

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

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5134–5143, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag J Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pp. 1–13, 2025.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, CAMELYON16 consortium, Meyke Hermsen, Quirine F Manson, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *MedRxiv*, pp. 2023–07, 2023.

Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

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

Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 189–199. Springer, 2024.

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. *arXiv preprint arXiv:2512.23545*, 2025.

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

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

Paul Kleihues, WK Cavenee, et al. Who classification of tumours. *Pathology & genetics. Tumors of the nervous system. Lyon, France: IARCpress*, 2000.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021.

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

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

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

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 *28th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025)*, pp. 680–690. Springer, 2026.

Yingfan Ma, Bohan An, Ao Shen, Mingzhi Yuan, Minghong Duan, and Manning Wang. Flow-mil: Constructing highly-expressive latent feature space for whole slide image classification using normalizing flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23561–23570, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13183–13192, 2024.

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

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Jingxiong Li, Xuan Gong, Xinheng Lyu, Tao Lin, et al. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 94611–94653, 2025.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Ashish Vaswani et al. Attention is all you need, 2017.

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

Sheng Wang, Ruiming Wu, Charles Herndon, Yihang Liu, Shunsuke Koga, Jeanne Shen, and Zhi Huang. Pathology-cot: Learning visual chain-of-thought agent from expert whole slide image diagnosis behavior. *arXiv preprint arXiv:2510.04587*, 2025.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Luca L Weishaupt, Chengkuan Chen, Drew FK Williamson, Richard J Chen, Guillaume Jaume, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Ming Y Lu, et al. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964*, 2025.

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.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18802–18812, 2022.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28418–28426, 2026.

Xiao Zhou, Luoyi Sun, Dexuan He, Wenbin Guan, Ge Wang, Ruifen Wang, Lifeng Wang, Xiaojun Yuan, Xin Sun, Ya Zhang, et al. Knowledge-enhanced pretraining for vision-language pathology foundation model on cancer diagnosis. *Cancer Cell*, 44(4):777–791, 2026.

A Appendix

A.1 More Related Work

Multiple Instance Learning for WSI Classification Early neural MIL methods introduced trainable permutation-invariant aggregation, with attention-based MIL Ilse et al. (2018) further providing patch-level importance scores that make slide-level prediction more interpretable. DSMIL Li et al. (2021) subsequently proposes a dual-stream architecture that jointly models instance-level and bag-level representations to improve weakly supervised WSI classification. In computational pathology, CLAM Lu et al. (2021) extends attention-based MIL with instance-level clustering constraints and demonstrated data-efficient WSI classification using only slide-level supervision. Recent methods have improved MIL from different perspectives. DTFD-MIL Zhang et al. (2022) constructs pseudo-bags and performs double-tier feature distillation to mitigate the limited number of slide-level training samples. TransMIL Shao et al. (2021) models correlations among instances with a Transformer-based architecture, relaxing the independent-instance assumption commonly made in conventional MIL.

Field	Description
key	Unique feature identifier (e.g., `glandular_acinar_pattern`).
description	Human-readable morphology description, injected into VLM prompts.
scale	Native observation scale: `region`, `detail`, or `both`.
strength_at_native_scale	KB-asserted diagnostic strength at native scale (`strong`/`moderate`/`weak`).
strength_at_other_scale	KB-asserted strength when observed at non-native scale.
requires_cross_scale_confirmation	If `true`, the feature must be confirmed at the opposite scale (drives M1).
confirms_with	List of partner feature keys that confirm this feature at the opposite scale.
requires_same_scale_combination	If `true`, the feature requires a co-occurring partner on the same patch.
combines_with	Partner feature keys for same-scale combination.
polarity	Owning class name or `AMBIG` for ambiguous features.

Table 6: FeatureSpec fields in PathKB.

More recently, Flow-MIL Ma et al. (2025) proposes a normalizing-flow-based latent semantic embedding space to enhance the expressiveness and separability of pathology features. While these methods significantly improve slide-level classification accuracy, they primarily optimize predictive performance rather than evidence reliability. Consequently, the resulting diagnosis is often difficult to verify and offers limited support for resolving prediction-evidence inconsistencies.

Pathology Foundation Models Advances in pathology foundation models have shifted the field from task-specific feature extractors to large-scale reusable representations. Phikon Filiot et al. (2023) and the subsequently released Phikon-v2 Filiot et al. (2024) demonstrate the effectiveness of large-scale self-supervised pretraining and provide highly transferable pathology representations across diverse downstream tasks. UNI Chen et al. (2024) further scales pretraining to over 100 million histology image patches, providing highly transferable representations across diverse computational pathology tasks. CONCH Lu et al. (2024a) introduces vision-language pretraining for histopathology by learning from large-scale image-caption pairs, enabling language-aligned pathology applications. Concurrently, Virchow Vorontsov et al. (2024) investigates large-scale pathology pretraining and delivers competitive performance across a wide range of clinical and research benchmarks. More recently, TITAN Ding et al. (2025) extends pathology foundation modeling to slide-level through multimodal learning and shows SOTA transfer performance across diverse downstream tasks. These advances have substantially improved pathology representation learning and enabled stronger downstream models. However, foundation models primarily enhance feature quality and do not address how diagnostic evidence should be validated, reconciled, or revised during decision making.

A.2 Pathology Knowledge Base Construction

PathKB supplies the diagnostic criteria against which Defender and Critic claims are adjudicated. Each task is paired with a compact set of YAML entries, grounded in the WHO diagnostic criteria for the corresponding organ system, stating which morphologies are exclusive to which class and at which scale each is reliably observable.

A.2.1 KB Curation Principles and Construction protocol.

The KB is curated following four principles:

1. **Presence-based, not threshold-based.** All entries are qualitative—no percentages, no "more than / greater than" comparisons. This reflects the VLM's capability bound: it can reliably report *presence* of a morphology but not *degree*.
2. **Scale-aware.** Every pathognomonic feature is annotated with its native observation scale. Features visible only at detail scale (e.g., keratin pearls, intercellular bridges) are assigned `weak` strength at region scale, preventing the VLM from making a strong claim based on a scale where the feature is not reliably visible.
3. **Mimic exclusion.** For each pathognomonic feature, common mimics are documented and must be excluded before the feature can be cited as Type A. For example, before reporting `metastatic_deposit`, the agent must rule out sinus histiocytosis, reactive follicular hyperplasia, and other mimics listed in the KB.
4. **Asymmetric strength.** Features that are pathognomonic (exclusive to one class) are assigned `strong` strength; features that are suggestive but not exclusive are assigned `moderate` or `weak`. This directly controls the decision threshold: only `strong` features can contribute to STRONG agent strength, which is required for ACCEPT or REJECT.

KB entries were produced by a draft–audit–refine loop rather than written from scratch. A drafting model (*Gemini-3.1-Flash* Team et al. (2023)) generated an initial entry for each class and borderline distinction, grounded in the diagnostic criteria of the WHO Classification of Tumours Kleihues et al. (2000). An auditing model (*Claude Opus 4.8*) then reviewed each draft against the four principles above over two rounds, returning targeted revision requests—quantitative thresholds to be requalified as presence statements (P1), strength assertions missing a scale qualifier (P2), pathognomonic features lacking a documented mimic list (P3), and strength labels inconsistent with class exclusivity (P4)—which the drafting model then applied. Entries whose morphologic claims could not be traced to the WHO criteria were revised or discarded. Both models are used only at KB authoring time; neither is involved at inference.

A.2.2 PATHOGNOMONIC FEATURE BLOCK SCHEMA

Alongside the natural-language KB, each task YAML defines a structured `features_per_class` schema that encodes the same clinical knowledge as machine-checkable rules. This is the "single source of truth" for feature strength and cross-scale confirmation.

```
features_per_class:
  LUAD:
    pathognomonic:
      - key: glandular_acinar_pattern
        description: "Glandular acinar pattern with luminal
            differentiation"
        scale: detail
        strength_at_native_scale: weak
        strength_at_other_scale:  weak

      - key: intracellular_mucin
        description: "Visible intracellular mucin droplets signet-ring
             forms"
        scale: detail
        strength_at_native_scale: strong   # LUSC does not
        strength_at_other_scale:  weak     # produce mucin

  LUSC:
    pathognomonic:
      - key: keratin_pearls
        description: "Rounded eosinophilic keratinised centres within
            squamous nests"
        scale: detail
        strength_at_native_scale: strong   # LUAD never
        strength_at_other_scale:  weak     # keratinises
```

KB-asserted strength lookup. The diagnostic strength of a cited feature is *never* self-asserted by the LLM. It is looked up from this table as a function of the feature f and the scale s at which it was

Scale	Patch Size	Magnification	Diagnostic Purpose
Region	4096×4096 px	$20\times$	Tissue architecture, growth patterns
Detail	224×224 px	$20\times$	Nuclear atypia, cytological differentiation

Table 7: Dual-scale patch pool.

observed:

$$\text{Strength}(f, s) = \begin{cases} \texttt{strength[native]} & \text{if } s = \texttt{f.scale} \\ \texttt{strength[other]} & \text{otherwise} \end{cases} \tag{10}$$

This ensures that the decision boundary depends on curated clinical knowledge, not on the LLM's self-assessed confidence.

A.2.3 Feature Consistency Pairs

The task YAML also encodes `feature_consistency_pairs`: morphological features that are mutually exclusive in the same field of view. When both are cited on the same patch, both are invalidated (Mechanism M2).

```
feature_consistency_pairs:
  - feature_a: glandular_acinar_pattern
    feature_b: solid_growth_pattern
    reason: "True gland formation and solid sheets are architecturally
         exclusive in the same field;  if both appear, one is a
        misread."
```

Co-citation on *different* patches is permitted, as this reflects tumour heterogeneity (e.g., a tumour may have glandular and solid areas in different fields).

A.2.4 Rule Block File Structure

The rule block file uses a simple list-of-entries schema:

```
- id: NSCLC_LUAD_ANCHOR
  title: LUAD anchor -- glands and/or mucin
  text: Call LUAD when the tumour forms true glands/acini that enclose
       lumens, or when tumour cells contain intracytoplasmic mucin ...
```

Each entry has three fields:

- `id`: a unique identifier cited by agents in `kb_citation`;
- `title`: a human-readable summary, also used to build the KB index;
- `text`: the full clinical guidance, written in natural language.

A.3 Details in Verifier

A.3.1 Multi-Scale Patch Pool

The Verifier operates on two spatial scales, motivated by the clinical distinction between *architectural* and *cytological* diagnosis. As shown in Table 7, each scale is further partitioned by MIL-assigned role:

- **Support** ($\mathcal{S}$): examined by the Defender agent;
- **Counter** ($\mathcal{C}$) + **Context** ($\mathcal{X}$): examined by the Critic agent.

A spatial mapping links each region patch to its constituent detail patches (those whose coordinates fall within the region's bounding box), enabling cross-scale confirmation (§A.3.3).

A.3.2 Two-Stage Feature Extraction (Stage A → Stage B)

Patch-level feature extraction uses a two-stage text pipeline.

Stage A — VLM Narrative. A vision-language model (VLM) examines each H&E patch and produces: (i) an MCQ classification vote (one letter per class plus "non-diagnostic"), and (ii) a free-text narrative describing the visible morphology. The narrative is deliberately rich and observational (3–6 sentences) because Stage B parses it—a terse answer-only reply starves Stage B of extractable evidence.

Stage B — Structured Extraction. A text-only LLM reads the Stage A narrative and outputs strict JSON:

```
{
  "predominant_pattern": "LUAD",
  "features": {
    "glandular_acinar_pattern": "yes",
    "intracellular_mucin": "no",
    "keratin_pearls": "unclear"
  },
  "evidence_quotes": {
    "glandular_acinar_pattern": "gland-forming structures with lumina"
  },
  "extraction_confidence": "high"
}
```

Three hard rules govern Stage B extraction:

1. A feature not explicitly mentioned in the narrative is labelled `unclear` (not `no`); only explicit absence yields `no`.
2. Every `yes` must be backed by a verbatim `evidence_quote` from the narrative; a `yes` without a quote is automatically downgraded to `unclear`.
3. No feature absent from the pathologist's description may be added (precision over recall).

A.3.3 Zoom-In: Cross-Scale Confirmation and Detail Cascade

The "zoom-in" mechanism is the core multi-scale reasoning component. It comprises two mechanisms that share the same spatial mapping but serve different purposes, plus a fallback that re-runs the agents at detail scale when the region-scale assessment is unresolved.

Spatial Mapping Construction For each region patch r at coordinate (x_r, y_r) with side length L_{region}, the set of detail patches within its field of view is:

$$\begin{aligned}\text{Detail}(r) = \big\{ d \mid\ & x_d \in [x_r, x_r + L_{\text{region}}], \\ & y_d \in [y_r, y_r + L_{\text{region}}] \big\}\end{aligned} \tag{11}$$

An 8-neighbourhood expansion extends this to detail patches within one tile step ($\pm L_{\text{region}}$) of the region boundary, allowing nearby confirmation.

Mechanism M1: Cross-Scale Feature Confirmation Each pathognomonic feature in the KB carries an optional flag `requires_cross_scale_confirmation` with a list of `confirms_with` partner features. When a Defender or Critic agent cites a region-scale feature f flagged for confirmation, the Verifier checks whether any of f's `confirms_with` partners was extracted as `yes` by Stage B on a detail patch within (or adjacent to) the same region patch.

Unconfirmed features are **invalidated** (removed from the decision computation). When a region feature survives M1 confirmation, its detail-scale confirmer is upgraded from MODERATE to STRONG strength, reflecting that multi-scale agreement is itself strong evidence.

Confirmation source switch. The source of confirmation evidence is configurable per task:

- `stage_b` (default): Stage B's structured `yes` signals on in-region detail patches provide confirmation.
- `agent_only`: confirmation must come from a detail-scale Defender/Critic agent citation (more conservative; if the detail cascade never ran, all cross-scale-requiring features are invalidated).

Verifier Configuration	BRCA	NSCLC	RCC	ESCA	BLCA	THCA	Avg.
Patho-R1 + Qwen3-4B	89.42	94.13	94.39	100.00	56.20	71.09	84.21
Qwen3.5-9B	91.78	93.13	94.52	97.62	52.23	69.96	83.21

Table 8: Comparison of different VLM–LLM configurations within the Verifier on six TCGA subtype classification benchmarks.

Method	BRCA	NSCLC	Avg.
Reviewer with Qwen3-4B planner	92.54	94.63	93.59
Reviewer w/o planner (KEEP only)	89.38	93.85	91.62

Table 9: Ablation study of planner guidance in the Reviewer module.

Mechanism M2: Same-Patch Consistency Pairs The KB encodes `feature_consistency_pairs`: morphological features that are physically impossible to co-occur in the same field of view (e.g., `glandular_acinar_pattern` and `solid_growth_pattern` in NSCLC). If both features of a pair are cited on the *same* patch, both are invalidated. Co-citation on *different* patches is permitted, as this reflects tumour heterogeneity (e.g., DCIS with microinvasion).

A.3.4 Detail Cascade Fallback

When the region-scale cascade returns UNCERTAIN and the task has `enable_detail_cascade_fallback = True`, the Defender and Critic are re-run at the detail scale on the same slide. The detail assessment applies the same deterministic decision policy as the region pass. If it produces ACCEPT or REJECT, it overrides the region UNCERTAIN; otherwise the verdict remains UNCERTAIN.

A.3.5 Defender Agent

The Defender examines *support* patches and reports whether the pathognomonic features of the MIL-predicted class are visible. Its output schema:

```
{
  "narrative": "1-2 sentence summary",
  "pathognomonic_present": "yes | no | unclear",
  "pathognomonic_features_cited": [
    {"feature": "<key>", "patch": "(x,y)", "kb_citation": "<KB id>"}
  ],
  "supporting_features_cited": [...],
  "contradicting_observations": "specific opposing-class features in
     the SAME patches",
  "image_quality": "good | limited | poor",
}
```

A.3.6 Critic Agent

The Critic examines *counter* and *context* patches, reporting pathognomonic features for *alternative* classes. Its central distinction is between two evidence types:

- **Type A** (challenger evidence): a feature pathognomonic for an alternative class is *present* (positive evidence). Only Type A can trigger a REJECT.
- **Type B** (ambiguous evidence): features that are compatible with an alternative class but insufficient to challenge. These are recorded but never trigger a flip.

A critical rule: the absence of the predicted class's features alone is never Type A in subtyping tasks. Only positive evidence for an alternative class counts.

A.3.7 CRITIC OUTPUT SCHEMA

```
{
  "narrative": "1-2 sentence summary",
  "type_A_findings": [
    {"feature": "<key>", "patch": "(x,y)",
     "favours_class": "<alt class>", "kb_citation": "<KB id>"}
  ],
  "type_B_findings": [
    {"feature": "<key>", "patch": "(x,y)", "why_only_type_B": "..."}
  ],
  "best_challenger": "<alt class> | null",
  "per_class_evidence": {"<alt class>": <int>},
  "image_quality": "good | limited | poor",
}
```

Algorithm 1: Planner-Guided Sequential Reviewer

Input: MIL prediction (y_0, q_0), Verifier report $\mathcal{R}$, feature set $\mathcal{F}$
Output: Final prediction $\hat{y}$

$\mathcal{P} \leftarrow \text{BuildEvidencePools}(\mathcal{R})$
$\mathcal{M}_0 \leftarrow \text{InitializeMemory}(y_0, q_0, \mathcal{R})$
$(\mathcal{B}, \mathcal{O}) \leftarrow (\emptyset, \emptyset)$
for $r = 1, \ldots, R_{\max}$ **do**
  $\mathcal{C}_r \leftarrow \text{BuildReviewState}(\mathcal{M}_0, \mathcal{B}, \mathcal{O}, r)$
  $\pi_r \leftarrow \text{Planner}(\mathcal{C}_r)$
  if $\pi_r = \text{STOP}$ **then**
    break
  end if
  if $\text{IsExpansion}(\pi_r)$ **then**
    $\mathcal{P} \leftarrow \text{ExpandEvidence}(\mathcal{P}, \pi_r)$
  end if
  $\mathcal{P}_r \leftarrow \text{PrepareReviewBatch}(\mathcal{P}, \pi_r, r)$
  $\mathcal{O} \leftarrow \text{MultiScaleReview}(\mathcal{P}_r, \mathcal{F})$
  $\mathcal{B} \leftarrow \text{UpdateBelief}(\mathcal{M}_0, \mathcal{O})$
  $z \leftarrow \text{Arbitration}(\mathcal{B}, \mathcal{R})$
  if $z \neq \text{UNCERTAIN}$ **then**
    break
  end if
end for
$z \leftarrow \text{Arbitration}(\mathcal{B}, \mathcal{R})$
if $z = \text{REVISE}$ **then**
  $\hat{y} \leftarrow \arg\max_{c \in \mathcal{Y} \setminus \{y_0\}} S(c)$
else
  $\hat{y} \leftarrow y_0$
end if
return $\hat{y}$

A.4 DETAILS OF THE REVIEWER

Algorithm 1 summarizes the complete Reviewer procedure for verifier-uncertain cases.

A.5 MORE EXPERIMETAL DETAILS

A.5.1 DETAILS FOR DATASETS

Specifically, the evaluated TCGA and CAMELYON datasets Weinstein et al. (2013); Ehteshami Bejnordi et al. (2017) include:

- **TCGA-RCC** (Renal Cell Carcinoma): a three-class classification task consisting of kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), and kidney renal papillary cell carcinoma (KIRP), containing 422 training slides and 238 testing slides.

Method	BRCA	NSCLC	RCC	ESCA	BLCA	THCA	Avg.
Flow-MIL Ma et al. (2025)							
Predictor	86.9	93.6	91.8	95.2	65.4	74.7	84.6
w/o Rev.	88.4	94.1	91.8	97.6	64.7	**76.3**	85.5
w/o Ver.	85.5	93.2	91.0	97.6	46.8	71.7	81.0
Full	**88.4**	**94.7**	**94.2**	**97.6**	**71.8**	**76.3**	**87.2**

Table 10: Ablation of the Verifier and Reviewer with Flow-MIL on six TCGA subtyping benchmarks (BACC, %). "Predictor" denotes the original MIL predictor; "w/o Rev." and "w/o Ver." remove the Reviewer and Verifier, respectively; and "Full" denotes the complete ReMIL-Agent framework. Best results are shown in bold.

- **TCGA-NSCLC** (Non-Small Cell Lung Cancer): a binary classification task distinguishing lung adenocarcinoma (LUAD) from lung squamous cell carcinoma (LUSC), containing 642 training slides and 195 testing slides.
- **TCGA-ESCA** (Esophageal Carcinoma): a binary classification task distinguishing adenocarcinoma (EAC) from esophageal squamous cell carcinoma (ESCC), containing 83 training slides and 31 testing slides.
- **TCGA-BRCA** (Breast Invasive Carcinoma): a binary classification task distinguishing invasive ductal carcinoma (IDC) from invasive lobular carcinoma (ILC), containing 637 training slides and 226 testing slides.
- **TCGA-THCA** (Thyroid Carcinoma): a three-class classification task consisting of papillary adenocarcinoma (PA), papillary carcinoma columnar cell (PCCC), and papillary carcinoma follicular variant (PCFV), containing 323 training slides and 119 testing slides.
- **TCGA-BLCA** (Bladder Urothelial Carcinoma): a binary classification task distinguishing transitional cell carcinoma (TCC) from papillary transitional cell carcinoma (PTCC), containing 227 training slides and 79 testing slides.
- **CAMELYON16** (Cancer Metastases in Lymph Nodes 2016): a binary classification task distinguishing lymph-node slides containing breast cancer metastases (*Tumor*) from slides without metastases (*Normal*), containing 270 training slides and 129 testing slides.

A.6 More Ablation Studies

Table 10 complements the ABMIL-based component ablation reported in the main text by evaluating the same configurations with Flow-MIL Ma et al. (2025) as the Predictor. Removing either the Verifier or the Reviewer reduces the average BACC relative to the complete framework, while applying the Reviewer without prior verification causes the largest degradation. These results are consistent with the ABMIL findings and further support that verification and review provide complementary benefits across different MIL backbones.

A.6.1 Verifier Model Configuration Analysis

We further investigate the impact of different VLM–LLM configurations within the Verifier. All experiments in this analysis use ABMIL Ilse et al. (2018) as the Predictor and evaluate only the Verifier stage without invoking the Reviewer. As shown in Table 8, we compare a pathology-specialized configuration, where Patho-R1 Zhang et al. (2026) is used for visual evidence extraction and Qwen3-4B Yang et al. (2025) performs evidence structuring and verification, against a unified general-purpose configuration where Qwen3.5-9B Yang et al. (2025) handles both visual description and reasoning. The Patho-R1-based verifier achieves an average BACC of 84.21%, outperforming the unified Qwen3.5-9B Yang et al. (2025) configuration by 1.00%. This indicates that pathology-oriented visual evidence extraction is more beneficial for verification than directly optimizing diagnostic prediction ability. Although Qwen3.5-9B Yang et al. (2025) achieves competitive standalone classification performance, the Verifier does not require direct image-level diagnosis but reliable morphological evidence extraction, where pathology-specific visual reasoning provides more effective support for evidence adjudication.

Dataset	Stage	Processed WSIs	Avg. Time / WSI	Avg. Time / Processed WSI	Total Time
TCGA-BLCA $N = 79$	Predictor	79	0.50	0.50	39.22
	Verifier	33	18.34	43.91	1449.18
	Reviewer	26	1.83	5.57	144.86
	Full ReMIL-Agent	–	**20.67**	**49.98**	**1633.27**
TCGA-THCA $N = 119$	Predictor	119	0.41	0.41	49.30
	Verifier	47	42.16	106.75	5017.23
	Reviewer	41	2.48	7.21	295.54
	Full ReMIL-Agent	–	**45.06**	**114.37**	**5362.08**

Table 11: Stage-wise time breakdown of ReMIL-Agent. "Processed WSIs" denotes the number of WSIs processed by each stage. "Avg. Time / WSI" is computed over all WSIs in the dataset, whereas "Avg. Time / Processed WSI" is computed only over the WSIs processed by the corresponding stage. "Total Time" denotes the accumulated time of each stage over all processed WSIs. For the Full ReMIL-Agent, "Avg. Time / Processed WSI" is the sum of the stage-wise average time of the Predictor, Verifier, and Reviewer.

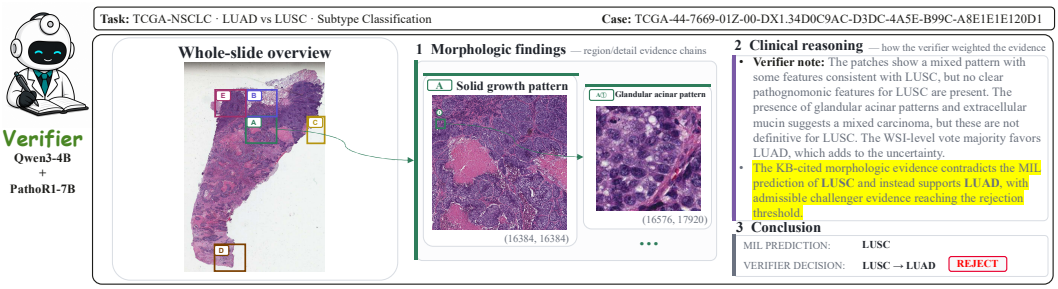

Figure 4: Case study on the TCGA-NSCLC subtype classification task. The Verifier identifies morphologic evidence contradicting the MIL prediction and directly rejects LUSC in favor of LUAD.

A.6.2 Planner Effect in Reviewer

We further investigate the contribution of the planner in the Reviewer. As shown in Table 9, we compare the full Reviewer with Qwen3-4B Yang et al. (2025) as the planner against a simplified variant that directly applies KEEP-based Zhou et al. (2026) evidence assessment without planner guidance. The planner-guided Reviewer achieves higher BACC on both TCGA-BRCA and TCGA-NSCLC, improving the average performance from 91.62% to 93.59%. This improvement demonstrates that the planner effectively schedules evidence examination and enables more adaptive multi-round review, rather than relying on fixed evidence aggregation.

A.7 More Efficiency Analysis

Beyond the overall efficiency comparison with PathAgent Chen et al. (2026), we further analyze the computational cost of ReMIL-Agent by reporting the stage-wise runtime breakdown of different components on the TCGA-BLCA and TCGA-THCA datasets in Table 11.

A.8 More Case Studies

We provide additional case studies to demonstrate different verification and correction patterns of ReMIL-Agent.

Figure 4 presents a TCGA-NSCLC subtype classification case, where the MIL predictor initially assigns LUSC. The Verifier identifies contradictory morphologic evidence, including glandular acinar patterns and extracellular mucin, while no definitive LUSC-specific features are observed. The KB-

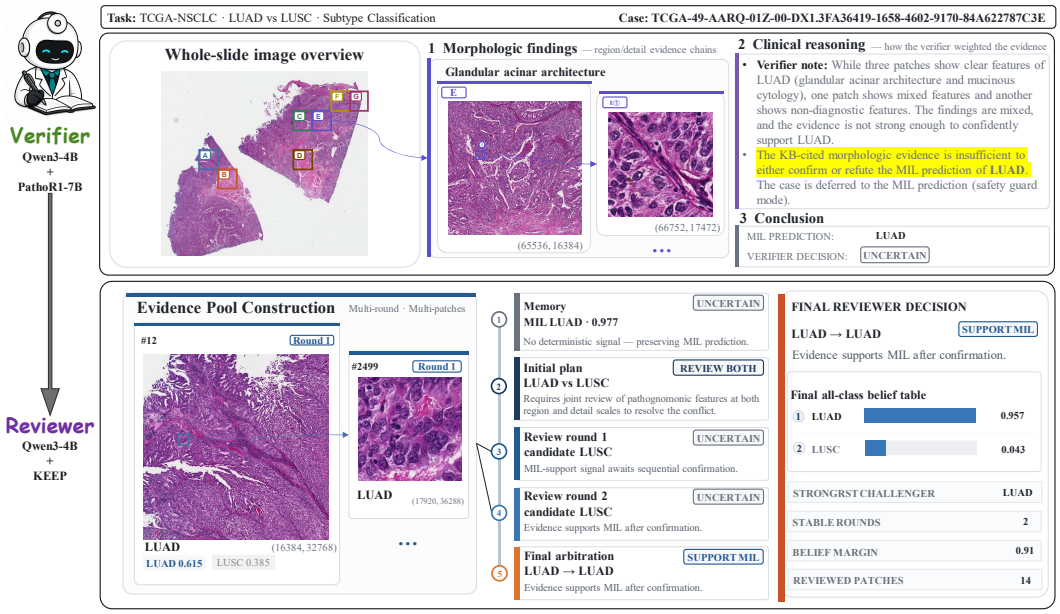

Figure 5: Case study on the TCGA-NSCLC LUAD–LUSC subtype classification task. The Verifier identifies supportive but incomplete morphologic evidence and returns UNCERTAIN. The Reviewer performs planner-guided sequential evidence review and confirms the original LUAD prediction, resulting in a SUPPORT_MIL decision.

cited evidence supports the challenger class LUAD and reaches the rejection threshold, leading the Verifier to directly reject the MIL prediction and correct the result to LUAD.

Figure 5 presents a TCGA-NSCLC subtype classification case, where the MIL predictor initially assigns LUAD with high confidence. The Verifier performs multi-scale morphologic verification and identifies supportive evidence for LUAD, including glandular acinar architecture and mucinous cytology. However, due to mixed and non-diagnostic findings from several patches, the evidence is insufficient to confidently confirm or refute the MIL prediction, leading to an UNCERTAIN decision under the safety guard mechanism. The Reviewer subsequently performs planner-guided sequential evidence review over region- and detail-scale patches. After accumulating consistent evidence across multiple review rounds and confirming the absence of decisive contradictory signals, the Reviewer outputs SUPPORT_MIL and preserves the original LUAD prediction.