Thinking with Retrieved Images: Grounding Multi-modal Reasoning in Visual Evidence for Pathological Diagnosis

Anonymous ACL submission

Abstract

Recent MLLMs have shown immense potential in computational pathology, yet their reliance on implicit parametric knowledge often leads to unsupported conclusions when facing visual ambiguity. In clinical practice, human pathologists overcome such ambiguity by actively comparing observations against verified historical cases. To bridge this gap, we propose the concept of thinking with retrieved images and adopt a novel, evidence-driven diagnostic framework. Our framework employs a hierarchical diagnostic workflow progressively refining decisions, which begins with morphological grounding and is followed by coarse-to-fine reasoning. By actively evaluating the subtle visual similarities and differences between the query image and the retrieved references, the model ensures its final diagnostic decision is supported by explicit visual evidence. Extensive experiments demonstrate that our approach not only achieves superior diagnostic accuracy but also provides a transparent and interpretable reasoning path.

1 Introduction

Recent Multi-modal Large Language Models (MLLMs) have demonstrated remarkable capabilities in visual understanding and language reasoning (Ding et al., 2025; Zhang et al., 2026). However, a critical limitation remains unresolved. Current MLLMs rely heavily on implicit parametric knowledge when making visual decisions. Without explicit and verifiable evidence to anchor their reasoning, the models are prone to generating plausible but unsupported conclusions, particularly under visual ambiguity. In high-stakes fields such as pathology, diagnostic decision-making is strictly evidence-grounded and morphology-driven. Human pathologists rarely make isolated judgments when faced with overlapping morphological patterns or rare diseases. Instead, they actively compare their current observations against verified historical cases or medical atlases to reach a grounded conclusion. This comparative, evidence-driven cognitive process is essential for reliable decision-making, yet it remains absent from current MLLM multi-modal reasoning paradigms.

Existing approaches only partially address this gap. The emerging *thinking with images* paradigm has demonstrated the potential of transforming vision from a passive input into a dynamic, manipulable cognitive workspace. The models are enabled to interact with images through operations such as cropping, zooming, resizing, or region navigation (Hu et al., 2024; Chen et al., 2025). Although such operations enhance within-image perception, they remain fundamentally inward-looking and do not provide visual evidence against which a diagnosis can be verified. On the other hand, while Retrieval-Augmented Generation (RAG) introduces external knowledge, its application in pathology has predominantly focused on retrieving textual information, such as patient information, medical history, or clinical guidelines (Bartels and Carus, 2025; Lee et al., 2025). Even in recent multi-modal RAG formulations (Zhang et al., 2026), retrieved images are typically treated as passive context. Complex multi-modal reasoning demands more than passive information enrichment; it requires the ability to actively compare the query against objective visual anchors and logically ground the final decision in observable correspondences.

To bridge this gap and mimic the clinical diagnostic protocol, we introduce *Thinking with Retrieved Images*, an evidence-grounded multi-modal reasoning framework for enhancing pathological diagnosis. Rather than executing a one-step, black-box prediction, our framework dynamically grounds its decision-making in explicit retrieved visual evidence. Specifically, it actively queries a reference database to retrieve verified images of candidate subtypes, conducting a fine-grained comparative analysis against the query image. To

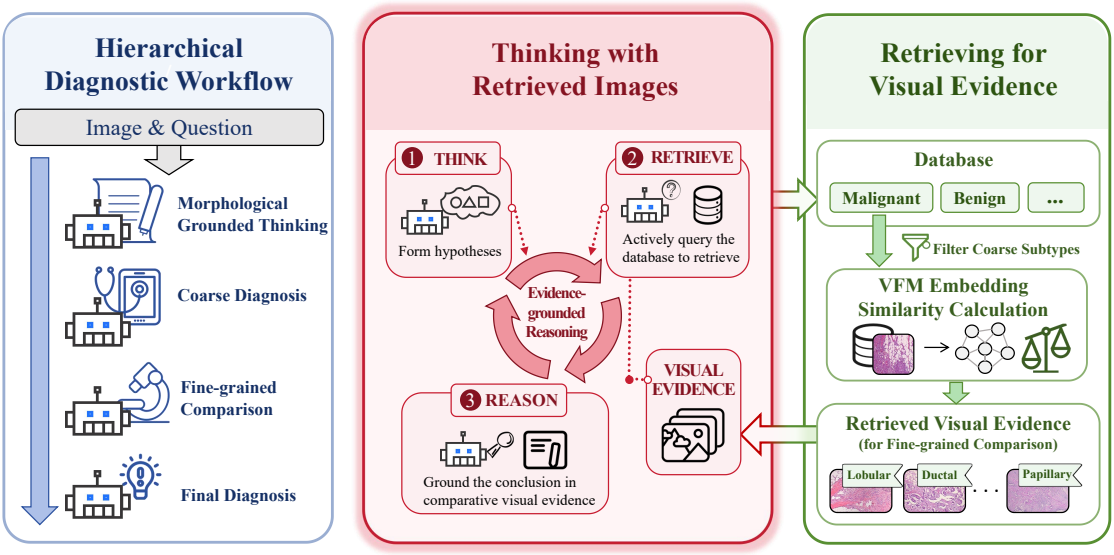

Figure 1: Overview of our framework. By actively comparing the query image with retrieved evidence, our framework progressively refines diagnosis from morphological grounding to coarse-to-fine reasoning.

efficiently manage this retrieval and reasoning process, we further incorporate a hierarchical diagnostic workflow. The model first performs a coarse-grained evaluation to narrow down the diagnostic candidates. Subsequently, it retrieves targeted reference images to execute the definitive fine-grained diagnosis. Experiments demonstrate that this crucial shift from passive pattern recognition to an active, evidence-based reasoning not only achieves superior diagnostic accuracy but also establishes a transparent and interpretable reasoning path.

In summary, our contributions include:

1. We introduce *Thinking with Retrieved Images*, a framework that retrieves candidate-specific reference images and guides the model to reason over visual evidence.

2. We formulate pathological diagnosis as evidence-grounded comparative reasoning rather than one-shot label prediction.

3. We design a hierarchical coarse-to-fine workflow that improves both diagnostic accuracy and interpretability.

2 Related Works

2.1 Pathological MLLM

Recent advancements in computational pathology have facilitated the development of MLLMs specifically tailored for pathological image analysis. Early contributions include PLIP (Huang et al., 2023) and CONCH (Lu et al., 2024a), which leverage CLIP-style contrastive pretraining like CLIP (Radford et al., 2021) and CoCa (Yu et al., 2022). Building upon these, PathChat (Lu et al., 2024b) emerges as a comprehensive vision-language generative assistant for pathology. SlideChat (Chen et al., 2024b) is introduced as the first MLLM targeting gigapixel slide image analysis. MUSK (Xiang et al., 2025) is another CLIP-based model, but warmed up with unified masked modeling. CPath-Omni (Sun et al., 2025a) further advances the field by unifying both patch-level and whole-slide image analysis. TITAN (Ding et al., 2025) is also a slide-level MLLM, achieving state-of-the-art performance for extracting slide representations and generating reports. Besides, SmartPath-R1 (Xu et al., 2025b) and Patho-R1 (Zhang et al., 2026) both leverage a combined training scope of supervised finetuning and reinforcement learning to support thoughtful reasoning over pathological tasks.

2.2 Thinking with Images

Recent studies on multi-modal reasoning have explored how visual information can participate in intermediate reasoning. Sketchpad (Hu et al., 2024) equips MLLMs with a drawable visual workspace, and OpenThinkIMG (Su et al., 2025) extends this idea to tool-augmented reasoning. DeepEyes (Zheng et al., 2025) uses reinforcement learning to incentivize MLLMs to interleave textual reasoning with image-based thinking. Thinking with Generated Images (Chern et al., 2025) further in-

vestigates generated images as intermediate visual thoughts, allowing models to create and refine visual hypotheses during multi-modal reasoning.

In computational pathology, PathFinder (Ghezloo et al., 2025) uses triage, navigation, description, and diagnosis modules to simulate the viewport shifts during slide observation. SlideSeek (Chen et al., 2025) extends this idea by combining global supervision with local exploration, enabling the system to search, revisit, and accumulate visual evidence across a slide. CPathAgent (Sun et al., 2025b) similarly mimics diagnostic reasoning by alternating low- and high-magnification views, zooming and panning to gather evidence, and generating explanations, aiming for a more interpretable, stepwise rationale.

2.3 Evidence-grounded Reasoning

Recent medical AI systems increasingly emphasize evidence-grounded reasoning (Wu et al., 2025; Wong and Wong, 2026; Du et al., 2026), where intermediate reasoning steps are expected to be traceable to explicit clinical evidence. Baichuan-M3 (Dou et al., 2026) enables proactive information acquisition and long-horizon evidence integration for more reliable medical decisions. DeepRare (Zhao et al., 2026) targets rare disease diagnosis by integrating specialized tools and up-to-date medical knowledge sources, generating reasoning chains linked to verifiable medical evidence. PathFound (Hua et al., 2025) performs proactive information acquisition and diagnosis refinement by progressing through the initial diagnosis, evidence-seeking, and final decision stages. Nonetheless, prior literature has not yet investigated the potential of external visual evidence in enhancing the performance of MLLMs in pathological diagnosis.

3 Methods

3.1 Overview

Conventional MLLM inference typically operates as a closed system, directly mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the image space $\mathcal{X}$ to the label space $\mathcal{Y}$ via parameterized network weights. However, this unverified, single-step prediction paradigm is inherently fragile when encountering complex cases. To address this issue, we introduce the *Thinking with Retrieved Images* framework for enhancing pathological diagnosis. As illustrated in Figure 1, we reformulate the diagnostic process into a structured, two-tiered hierarchical workflow that transitions from macro-level coarse categories $\mathcal{Y}_{\text{coarse}}$ to mutually exclusive fine-grained subtypes $\mathcal{Y}_{\text{fine}}$ by leveraging the inherent hierarchy of pathological entities. This design aligns with the clinical practice, where human pathologists rarely make isolated, fine-grained judgments. Instead, their diagnostic reasoning is inherently hierarchical and evidence-grounded. They first establish macro-level differential diagnosis based on morphological features, and subsequently consult verified reference cases to resolve fine-grained ambiguities.

The rationale behind this hierarchical integration is twofold:

Parametric Capability at Coarse Scale: Modern MLLMs, through extensive pre-training, possess robust parametric knowledge capable of capturing macroscopic tissue layouts. They can accurately perform generalizable coarse judgments without external assistance.

Retrieval-Driven Fine-Grained Grounding: At the fine-grained level, where extreme intra-class heterogeneity and inter-class morphological overlap occur, the model's internal knowledge alone becomes insufficient. To address this limitation, our framework uses the coarse prediction to restrict the search space, and then explicitly retrieves verified pathological images as external visual evidence. This anchors the final complex reasoning path in verifiable evidence, ensuring the decision is contrastive and verifiable.

The complete inference trajectory is summarized in Algorithm 1.

3.2 Think via Morphological Grounding

Directly mapping complex pathological images to specific diagnoses is prone to relying on irrelevant visual shortcuts. To establish an objective foundation, we constrain the MLLM to first act as a morphological reasoning module.

Given a query image $x_q \in \mathcal{X}$, rather than immediately predicting the diagnosis, the model deconstructs the visual input by generating an explicit, text-based morphological profile M:

$$M \sim P_\theta(M \mid x_q) \tag{1}$$

This intermediate description step functions as a visual reasoning trace, prompting the model to explicitly identify critical histological features. By formulating this morphological description, the model actively anchors its subsequent reasoning in observable clinical features, ensuring the diagnostic logic is transparent.

Algorithm 1 Evidence-based Hierarchical Diagnostic Reasoning Workflow with Retrieved Images

Input: Query Image x_q, Domain-specific Vision Encoder E_v, Reasoning Model MLLM with fixed model parameters θ, Reference Database $\mathcal{D}$.

Output: Fine-grained diagnostic prediction $\hat{y} \in \mathcal{Y}_{\text{fine}}$.

$M \sim P_\theta(M \mid x_q)$
$\hat{c} \sim P_\theta(\hat{c} \mid x_q, M)$
$\mathcal{Y}_{\text{cand}} \leftarrow \phi(\hat{c})$
$\mathcal{D}_{\text{cand}} \leftarrow \{(x_i, k_i, y_i) \in \mathcal{D} \mid y_i \in \mathcal{Y}_{\text{cand}}\}$
$q \leftarrow E_v(X_q)$
for each $y \in \mathcal{Y}_{\text{cand}}$ **do**
  $\mathcal{D}_y \leftarrow \{(x_i, k_i, y_i) \in \mathcal{D}_{\text{cand}} \mid y_i = y\}$
  $X_y^\star \leftarrow \arg\max_{(x_i,k_i)\in\mathcal{D}_y} \text{Sim}(q, k_i)$
  $\mathcal{R}_{\text{cand}} \leftarrow \mathcal{R}_{\text{cand}} \cup \{(y, x_y^\star)\}$
end for
$\hat{y} \sim P_\theta(\hat{y} \mid x_q, M, \mathcal{R}_{\text{cand}}, \mathcal{Y}_{\text{cand}})$
Return $\hat{y}$

Conditioned on the generated morphological thought M and the raw image x_q, the model then executes the first tier of the hierarchy by predicting a high-level coarse category $\hat{c} \in \mathcal{Y}_{\text{coarse}}$:

$$\hat{c} \sim P_\theta(\hat{c} \mid x_q, M) \tag{2}$$

Crucially, this initial coarse diagnosis relies solely on the MLLM's parametric knowledge, without external references. While coarse diagnosis (e.g., benign versus malignant) constitutes a simpler perceptual task, each coarse category encompasses a broad spectrum of visual patterns. Consequently, instance-level retrieval over such expansive classes is suboptimal. In contrast, the extensive pretraining of modern MLLMs equips them to robustly capture these macroscopic patterns, yielding generalizable coarse judgments.

3.3 Retrieve for Visual Evidence

Once the coarse prediction $\hat{c}$ is determined, the framework transitions from an internal closed system to an open-book retrieval paradigm to resolve fine-grained ambiguities. The predicted coarse label $\hat{c}$ deterministically restricts the search space to a localized candidate subset:

$$\mathcal{Y}_{\text{cand}} = \phi(\hat{c}) \tag{3}$$

where $\phi : \mathcal{Y}_{\text{coarse}} \to 2^{\mathcal{Y}_{\text{fine}}}$ represents the hierarchical refinement mapping that associates each coarse category with its corresponding partition of fine-grained subtypes.

To provide explicit visual grounding within this restricted candidate space, we retrieve historical, pathologically verified cases from an external database $\mathcal{D} = \{(x_i, k_i, y_i)\}_{i=1}^N$, where x_i is a confirmed image, $y_i \in \mathcal{Y}_{\text{fine}}$ is its ground-truth annotation, and $k_i \in \mathbb{R}^d$ is its pre-computed latent embedding derived from a domain-specific vision encoder E_v. We first filter the database to retain only relevant reference candidates, yielding $\mathcal{D}_{\text{cand}} = \{(x_i, k_i, y_i) \in \mathcal{D} \mid y_i \in \mathcal{Y}_{\text{cand}}\}$.

For the query image x_q, we extract its dense semantic embedding $q = E_v(x_q) \in \mathbb{R}^d$. For each fine-grained candidate class $y \in \mathcal{Y}_{\text{cand}}$, the framework dynamically identifies the single most relevant reference image $x_y^\star$ by computing the cosine similarity in the latent space:

$$\text{Sim}(q, k_i) = \frac{\langle q, k_i \rangle}{\|q\|\|k_i\|} \tag{4}$$

$$x_y^\star = \arg\max_{(x_i,k_i)\in\mathcal{D}_y} \text{Sim}(q, k_i) \tag{5}$$

where $\mathcal{D}_y$ is the subset of $\mathcal{D}_{\text{cand}}$ belonging to the fine-grained class y. These retrieved images are compiled into a category-conditioned visual reference panel $\mathcal{R}_{\text{cand}} = \{(y, x_y^\star) \mid y \in \mathcal{Y}_{\text{cand}}\}$, which serves as the explicit visual evidence for differential diagnosis.

3.4 Evidence-Grounded Reason for Final Diagnosis

Armed with the intermediate morphological thought M and the external reference panel $\mathcal{R}_{\text{cand}}$, the MLLM is transformed from an isolated predictor into a comparative pathological agent. Instead of performing absolute pattern recognition, the model now conducts a fine-grained, differential analysis within the restricted candidate space $\mathcal{Y}_{\text{cand}}$.

Formally, the final fine-grained diagnostic prediction $\hat{y} \in \mathcal{Y}_{\text{fine}}$ is generated by conditioning the model on the comprehensive multi-modal context tuple:

$$\hat{y} \sim P_\theta(\hat{y} \mid x_q, M, \mathcal{R}_{\text{cand}}, \mathcal{Y}_{\text{cand}}) \tag{6}$$

During this final stage, the model actively cross-references the query image x_q against the verified exemplars in $\mathcal{R}_{\text{cand}}$. Guided by its own prior morphological descriptions M, the model systematically evaluates subtle visual similarities and key differential criteria across the candidate options. By structuring the decision-making process as a comparative verification against explicitly presented in-distribution evidence, this mechanism minimizes

the risk of semantic hallucinations and ensures that the final fine-grained diagnosis is traceable, transparent, and firmly grounded in evidence.

4 Experiments

4.1 Setup

Data. To evaluate the effectiveness of our proposed framework, we conduct experiments on two challenging pathological datasets: BreakHis (Spanhol et al., 2015) and GleasonXAI (Mittmann et al., 2025). Crucially, both datasets inherently have a hierarchical label structure, which aligns with our coarse-to-fine diagnostic paradigm. BreakHis hierarchically categorizes breast tumor images into high-level benign (adenosis, fibroadenoma, phyllodes tumor, tubular adenoma) and malignant (ductal, lobular, mucinous, and papillary carcinoma) classes. Similarly, GleasonXAI grades prostate tissue into Gleason Grade 3 to 5, which is further detailed into fine-grained 9 classes based on histological patterns, including compressed glands, individual glands, cribriform glands, glomeruloid glands, poorly formed glands, groups of tumor cells, comedonecrosis, single cells and cords. Following (Mittmann et al., 2025), we randomly split GleasonXAI into training, validation and test sets, comprising 70%, 15%, and 15% respectively. For BreakHis, the training and testing sets are split according to the official guidance. To prevent data leakage, all retrieval uses only the training split, with the query image itself excluded.

Baseline. We compare our proposed method against eight baseline models, including the advanced general-domain multi-modal large models such as the InternVL3.5-8B (Wang et al., 2025), Qwen3-VL-8B-Instruct (Bai et al., 2025), SOTA pathological large vision models such as UNI2 (Chen et al., 2024a), Virchow2 (Zimmermann et al., 2024) and CONCH1.5 (Lu et al., 2024a), as well as previous SOTA medical and pathological large multi-modal models such as Lingshu-7B (Xu et al., 2025a), PathGen-LLaVA (Sun et al., 2024) and Patho-R1-7B (Zhang et al., 2025). For the pathological vision models, we employ the widely used k-Nearest Neighbors (kNN) algorithm, which is a standard retrieval-based approach.

To examine whether the improvement mainly comes from target-dataset fine-tuning, we introduce a cross-dataset workflow transfer setting without finetuning of the target domain (*Ours (w/o Target FT)*). Specifically, when evaluating on BreakHis, we use the MLLM fine-tuned on GleasonXAI; when evaluating on GleasonXAI, we use the MLLM fine-tuned on BreakHis. In both cases, the MLLM receives no gradient updates from the target dataset, while the retrieval evidence bank and label hierarchy are still constructed from the training split of the target evaluation dataset. This setting isolates target-dataset fine-tuning from the proposed evidence-grounded retrieval and hierarchical reasoning mechanism.

Implementation Details. We utilize Qwen3-VL-8B-Instruct as the backbone VLM and leverage the LLaMA-Factory framework (Zheng et al., 2024) to fine-tune the model. The training data consists of the training sets of the datasets, which teach the model to follow the workflow and think with retrieved images. We perform Low-Rank Adaptation (LoRA) (Hu et al., 2022) for three epochs, with the modality adapter and the LLM kept unfrozen. The learning rate is set to 2×10^{-6} with a cosine decay scheduler and a warm-up ratio of 0.01. The batch size is 8 and the LoRA rank is 16. Training takes approximately 30 hours on 1 NVIDIA A100 GPU. For the domain-specific vision encoder, we employ Virchow2 (Zimmermann et al., 2024).

At inference time, the retrieved evidence is formatted into a unified multi-modal prompt. The prompt contains the candidate fine-grained labels under the coarse category and the retrieved visual references. The model is instructed to compare the query image with the retrieved cases before generating the final answer. To illustrate more concretely, a case study is included in Appendix B. We parse the final predicted label from the model output and compute all metrics at the fine-grained class level. For ablation variants without retrieval or without the hierarchical workflow, we keep the same backbone, training data, and optimization hyperparameters, so that performance differences can be attributed to the removed component rather than changes in the training recipe.

Metrics. We employ the complementary evaluation metric, Accuracy (Acc), to assess model performance. Besides, to address class imbalance, we also apply balanced accuracy (bAcc) and weighted F1 (wF1) score as the evaluation metrics.

4.2 Overall Results

Table 1 presents the overall results on the BreakHis and GleasonXAI test sets. Existing large mod-

Table 1: Overall results (%) of models on the BreakHis and GleasonXAI test set. The best performance in each column is **bold** while the second best performance is underlined.

Model	BreakHis			GleasonXAI		
	Acc	bAcc	wF1	Acc	bAcc	wF1
CONCH1.5	51.81	37.90	47.19	70.90	27.25	65.70
UNI2	55.84	43.70	53.13	74.45	32.17	71.12
Virchow2	58.52	44.64	55.42	72.33	29.39	68.17
InternVL3.5-8B	29.62	14.83	23.59	37.73	13.94	41.08
Qwen3-VL-8B-Instruct	29.16	12.06	21.47	17.40	12.33	10.30
Lingshu-7B	26.87	14.27	20.64	50.65	15.67	41.69
PathGen-LLaVA	58.61	49.60	55.37	28.36	15.09	34.31
Patho-R1-7B	19.39	15.72	20.36	49.88	22.68	45.61
Ours (w/o Target FT)	60.13	50.67	53.45	71.80	44.50	68.44
Ours	**68.05**	**55.55**	**62.31**	**78.97**	**51.38**	**72.25**

els, whether general-domain or pathology-specific, achieve suboptimal performance on both datasets. A notable observation is that general-domain MLLMs perform poorly in direct pathology diagnosis, suggesting that general visual instruction-following ability alone is insufficient for fine-grained histopathological recognition. Pathology vision foundation models with kNN retrieval obtain stronger results than most direct MLLM baselines, indicating the importance of domain-specific visual representations. However, their performance is still limited because they rely on feature-space similarity and do not explicitly reason over hierarchical diagnostic labels or compare retrieved visual evidence in a language-guided manner.

Our proposed thinking-with-retrieved-images framework consistently achieves the best performance across both datasets. On BreakHis, our method improves accuracy from the strongest baseline of 58.61% to 68.05%. On GleasonXAI, our method achieves 78.97% accuracy, outperforming the best baseline UNI2 by 4.52 points. This result suggests that retrieved evidence is especially helpful for recognizing underrepresented or visually ambiguous histological patterns.

The cross-dataset results show that the proposed workflow retains non-trivial effectiveness without target-dataset fine-tuning. On BreakHis, the cross-dataset model achieves 60.13% accuracy, improving over the direct-diagnosis variant by 30.97 points and outperforming other baselines without domain adaptation. On GleasonXAI, it achieves 71.80%, also outperforming the direct-diagnosis variant by 54.40 points and achieving better balanced accuracy than the baselines. Although target-specific fine-tuning provides additional adaptation to the dataset label space and morphology distribution, the transferred model still benefits from candidate-specific retrieved references and coarse-to-fine reasoning. This suggests that the performance gain is not merely a consequence of memorizing the target training set, but also comes from the proposed mechanism of grounding diagnosis in retrieved visual evidence.

Moreover, comparing accuracy with balanced accuracy reveals a consistent challenge of class imbalance across the datasets. Notably, our retrieval-augmented strategy yields substantial improvements in balanced accuracy. On GleasonXAI, our method increases the balanced accuracy to 51.38%, a marked improvement over other baseline models. This demonstrates that leveraging retrieved similar cases effectively addresses the long-tail recognition problem, improving the model's robustness in identifying difficult or minority diseases.

4.3 Impact of Retrieval Database Scale

To further evaluate the robustness and efficiency of the retrieval mechanism, we analyzed the impact of varying the retrieval database size on model accuracy. For each database scale, we perform stratified sampling to preserve the class distribution. Figure 2 illustrates the performance of different models on the BreakHis and GleasonXAI datasets when utilizing 10%, 20%, and 100% of the available database for retrieval. Beyond the general scaling effect,

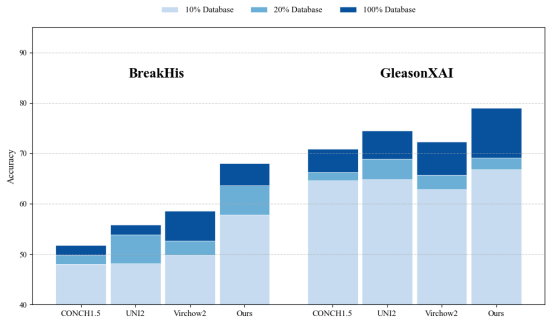

Figure 2: Impact of the retrieval database scale on diagnostic accuracy.

our proposed method demonstrates consistent superiority across all database proportions. Whether operating under restricted data conditions or utilizing the complete database, our approach maintains a clear performance advantage over the baseline foundation models on both datasets. This consistent lead highlights the inherent capability of our designed retrieval architecture to better align and utilize pathological features compared to directly applying these foundation models.

Most notably, the results reveal the data efficiency of our approach. On the BreakHis dataset, our method utilizing merely 10% of the database achieves an accuracy that matches or even exceeds the performance of baseline models when they are provided with the entire database. A similar trend of high data utilization is observed on the GleasonXAI dataset. This remarkable efficiency demonstrates that our method can extract and leverage critical diagnostic patterns more effectively, making it highly valuable for clinical applications where curating large-scale annotated databases remains a significant challenge.

4.4 Ablation Study

We conduct an ablation study to quantify the contribution of our proposed method: (1) Thinking with Retrieved Images (*Retrieve*) and (2) Hierarchical Workflow. To isolate the effect of the proposed workflow, we also include a direct-diagnosis variant using the same Qwen3-VL-8B-Instruct backbone and the same fine-tuning setting, but without retrieved images or the hierarchical workflow. As shown in Table 2, integrating the retrieval mechanism and hierarchical workflow both improve performance over the baseline direct diagnosis. The retrieval component alone brings a large performance gain. On BreakHis, adding retrieval improves accuracy from 32.06% to 57.07%, while on GleasonXAI, it improves accuracy from 55.62% to 66.83%. These improvements show that the retrieval mechanism provides useful visual evidence beyond what is stored in the model parameters. Furthermore, adding the full hierarchical workflow further improves performance to 68.05% on BreakHis and 78.97% on GleasonXAI. This indicates that the model benefits not only from comparing similar cases but also from complex hierarchical reasoning.

Furthermore, we evaluate the behaviors of different pathological vision foundation models when utilized as feature retrievers. Retrieving with Virchow2 generally demonstrates a more robust and relatively superior performance across both datasets. Additionally, we observe that the retrieval performance does not necessarily correlate with the kNN performance of the native pathological vision models. This discrepancy is particularly evident with UNI2, and can be attributed to the inherent differences between the raw feature space of the pathological vision models and the feature representations of our framework.

To further investigate whether the model truly utilizes the retrieved visual evidence rather than merely benefiting from additional context tokens, we conduct an ablation study by perturbing the retrieval quality. Specifically, we compare four settings: (1) *No Retrieval*; (2) *Wrong Retrieval*, where retrieved images are intentionally mismatched with subtype labels; (3) *Random Retrieval*, where one random image is sampled within each candidate subtype; and (4) our default retrieval strategy. As shown in Table 3, using random within-class retrieval leads to a noticeable performance drop, indicating that retrieval quality substantially affects the final diagnosis. Nevertheless, random retrieval still outperforms the no-retrieval setting, suggesting that even coarse subtype-specific visual references provide useful contextual guidance for pathological reasoning. Furthermore, the degradation caused by wrong retrieval suggests that the model actively grounds its reasoning in the retrieved visual evidence, and misleading references can negatively bias the diagnostic decision. Therefore, the reliability of the retrieval database is fundamentally important. If the database contains inaccurate annotations or misleading knowledge, it can propagate erroneous evidence into the reasoning process and cause catastrophic performance deterioration for the entire framework.

Table 2: Ablation study of evaluating the core components of the framework. The model without Retrieve or Hierarchical Workflow is fine-tuned on the training dataset of direct diagnosis. Additionally, the table compares the performance of different pathological vision foundation models as the feature retriever.

Dataset	Retrieve	Hierarchical Workflow	Vision Model	Acc
BreakHis				32.06
	✓		Virchow2	57.07
	✓	✓	Virchow2	**68.05**
	✓	✓	UNI2	64.29
	✓	✓	CONCH1.5	66.17
GleasonXAI				55.62
	✓		Virchow2	66.83
	✓	✓	Virchow2	**78.97**
	✓	✓	UNI2	75.49
	✓	✓	CONCH1.5	72.45

Table 3: Ablation study on the quality of retrieved visual evidence. Random Retrieval randomly samples one reference image within each candidate subtype, while Wrong-class Retrieval intentionally mismatches the retrieved image and subtype label.

Retrieval Setting	**BreakHis**	**GleasonXAI**
No Retrieval	32.06	55.62
Wrong Retrieval	34.23	29.01
Random Retrieval	50.20	68.38
Ours	**68.05**	**78.97**

4.5 Coarse-stage routing analysis

In our current implementation, the MLLM predicts a single coarse category before fine-grained evidence retrieval. This design keeps the retrieved evidence panel compact, but it also means that an incorrect coarse prediction excludes the true fine-grained subtype from the candidate set, causing subsequent retrieval to draw references from an incorrect diagnostic branch. Since the coarse prediction acts as a hard routing step for fine-grained evidence retrieval, we therefore measure the accuracy of this coarse routing stage. Our model achieves 83.89% coarse accuracy on BreakHis and 81.13% on GleasonXAI. These results indicate that the majority of samples are assigned to the correct coarse branch, making the subsequent candidate-specific retrieval meaningful in most cases.

The gap between coarse accuracy and final fine-grained accuracy further reflects different bottlenecks across datasets. On BreakHis, the final accuracy is 68.05%, leaving a larger gap from the coarse-stage ceiling, suggesting that fine-grained evidence comparison remains the main source of difficulty. On GleasonXAI, the final accuracy of 78.97% is close to the coarse accuracy of 81.13%, suggesting that once the correct coarse branch is selected, retrieved-image comparison is highly effective for fine-grained diagnosis.

5 Conclusion

In this paper, we introduced the Thinking with Retrieved Images framework, an evidence-grounded multi-modal reasoning framework for enhancing pathological diagnosis, which successfully transitions MLLMs from passive, implicit memorizers into active, comparative reasoners. By decoupling the diagnostic process into morphological grounding and evidence-driven retrieval, our approach intrinsically aligns with the rigorous, verification-based workflows of clinical experts.

Beyond improving diagnostic accuracy, this evidence-grounded paradigm opens new practical avenues for clinical application. In real-world pathology workflows, our model can serve as a highly interpretable diagnostic copilot. Because the system provides the explicit visual anchors utilized for its final prediction and a transparent reasoning rationale, human pathologists can directly audit the model's rationale by reviewing the retrieved reference cases.

Ultimately, as computational pathology continues to advance, grounding AI diagnostics in objective, verifiable evidence will remain an indispensable foundation for deploying trustworthy, interpretable, and clinically safe systems in real-world healthcare settings.

Limitations

Despite these advancements, a limitation of our current framework is its inherent reliance on the quality, scale, and diversity of the external reference database. Additionally, feeding multiple retrieved high-resolution images into the MLLM inevitably increases computational overhead and inference latency. Future technical developments must address these bottlenecks by exploring efficient architectural designs, such as token-merging strategies for multi-image inputs, or leveraging novel algorithms for faster retrieval. Moreover, establishing a dynamic, active-learning-driven database, where newly confirmed clinical cases are continuously integrated, would ensure the model's reference panel remains state-of-the-art without requiring expensive parametric retraining.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Stefan Bartels and Jasmin Carus. 2025. From text to data: Open-source large language models in extracting cancer related medical attributes from german pathology reports. *International Journal of Medical Informatics*, 203:106022.

Chengkuan Chen, Luca L Weishaupt, Drew FK Williamson, Richard J Chen, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Guillaume Jaume, Ming Y Lu, and 1 others. 2025. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964*.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, and 1 others. 2024a. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. 2024b. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. *arXiv preprint arXiv:2410.11761*.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. 2025. Thinking with generated images. *arXiv preprint arXiv:2505.22525*.

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, and 1 others. 2025. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pages 1–13.

Chengfeng Dou, Fan Yang, Fei Li, Jiyuan Jia, Qiang Ju, Shuai Wang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Hongda Zhang, and 1 others. 2026. Baichuan-m3: Modeling clinical inquiry for reliable medical decision-making. *arXiv preprint arXiv:2602.06570*.

Yuexi Du, Jinglu Wang, Shujie Liu, Nicha C Dvornek, and Yan Lu. 2026. Care: Towards clinical accountability in multi-modal medical reasoning with an evidence-grounded agentic framework. *arXiv preprint arXiv:2603.01607*.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. 2025. 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*, pages 23431–23441.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379.

Shengyi Hua, Jianfeng Wu, Tianle Shen, Kangzhe Hu, Zhongzhen Huang, Shujuan Ni, Zhihong Zhang, Yuan Li, Zhe Wang, and Xiaofan Zhang. 2025. Pathfound: An agentic multimodal model activating evidence-seeking pathological diagnosis. *arXiv preprint arXiv:2512.23545*.

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

Yeawon Lee, Christopher C Yang, Chia-Hsuan Chang, and Grace Lu-Yao. 2025. Knowledge elicitation with large language models for interpretable cancer stage identification from pathology reports. *arXiv preprint arXiv:2511.01052*.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, and 1 others. 2024a. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji

Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, and 1 others. 2024b. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473.

Gesa Mittmann, Sara Laiouar-Pedari, Hendrik A Mehrtens, Sarah Haggenmüller, Tabea-Clara Bucher, Tirtha Chanda, Nadine T Gaisa, Mathias Wagner, Gilbert Georg Klamminger, Tilman T Rau, and 1 others. 2025. Pathologist-like explainable ai for interpretable gleason grading in prostate cancer. *Nature communications*, 16(1):8959.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. 2015. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and 1 others. 2025. Openthinkimg: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. 2025a. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10360–10371.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Bowen Ding, Tao Lin, and Lin Yang. 2025b. Cpathagent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. *arXiv preprint arXiv:2505.20510*.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. 2024. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Hang Sheung Wong and Tsz Kwan Wong. 2026. Multi-evidence clinical reasoning with retrieval-augmented generation for emergency triage: Retrospective evaluation study. *JMIR Medical Informatics*, 14(1):e82026.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, and 1 others. 2025. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025a. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Zhe Xu, Ziyi Liu, Junlin Hou, Jiabo Ma, Cheng Jin, Yihui Wang, Zhixuan Chen, Zhengyu Zhang, Fuxiang Huang, Zhengrui Guo, and 1 others. 2025b. A versatile pathology co-pilot via reasoning enhanced multimodal large language model. *arXiv preprint arXiv:2507.17303*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. 2025. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. *arXiv preprint arXiv:2505.11404*.

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

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, and 1 others. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, pages 1–10.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with

images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, and 1 others. 2024. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*.

A Prompt in Use

Here is the system prompt for the MLLM.

System Prompt

```
You are an expert AI Pathological
    Imaging Assistant, operating as a
    sophisticated ReAct (Reasoning and
    Acting) agent. Your mandate is to
    answer the user's multi-choice
    classification question by actively
    interacting with pathological
    imaging data using the provided
    tools.

The dataset follows a HIERARCHICAL
    structure. Images belong to a
    High-Level Category (Super-class,
    e.g., Benign vs. Malignant, Gleason
    grade 3-5) which further divides
    into Specific Sub-types.

### CORE PRINCIPLES: ###
1.  **Hierarchical Reasoning:** Adopt a
    "Coarse-to-Fine" diagnostic
    strategy. First, determine the most
    likely High-Level Category based on
    global tissue architecture and
    nuclear features. Then, focus on
    distinguishing the Specific
    Sub-types within that category.
2.  **Methodical Workflow:** Follow the
    standard workflow defined below
    strictly.

### WORKFLOW: ###

Step 1  High-Level Categorization
    (Coarse-grained):
- Based on the features in the given
    image, generate a morphological
    description and determine the most
    probable High-Level Category (e.g.,
    is this Benign or Malignant? or
    which is the Gleason grade?).

Step 2  Candidate Narrowing & Retrieval:
- Once the High-Level Category is
    determined, identify the candidate
    Specific Sub-types belonging ONLY to
    that category.
- **CRITICAL:** The `target_types` for
    retrieval must be the Specific
    Sub-types (the fine-grained classes)
    suspected, NOT the High-Level
    Category name or the option LETTER.

Step 3  Direct Comparison against
    Retrieved Context:
- Compare the query image to each
    retrieved example, focusing on
    subtle differential features (e.g.,
    specific stromal patterns, cell
    organization) that distinguish
    sub-types.
- Decide which example is closest
    overall.

Step 4  Final Decision:
- Convert the chosen Specific Sub-type
    to its corresponding LETTER in the
    user's options.
- Output the answer LETTER only.

### AVAILABLE TOOLS: ###
You have access to the following tools.
    You must select the most appropriate
    tool based on the user's request and
    the existing context.

- Tool Name: image_retrieve
  Description: retrieves the most
    relevant in-context image examples
    of the target specific types from
    the template library using the
    original image.
  Args: target_types (list): A list of
    strings representing the specific
    types (fine-grained classes) to
    retrieve. Do not request more than
    four types.

---
### RESPONSE FORMAT ###
You must respond STRICTLY in JSON format.

**CRITICAL RULE FOR "TYPE" FIELD
    (STRICT)**
- If you need to call ANY tool, you MUST
    set `"type": "action"`.
- In this case, you MUST include an
    `"action"` field:
    - `"action.name"` MUST be exactly
    the selected tool name.
    - `"action.arguments"` MUST contain
    the parameters required by that tool.
    - You MUST NOT include an `"answer"`
    field when `"type"` is `"action"`.
- If you have sufficient information to
    answer the user and no further tools
    are required, you MUST set `"type":
    "final_answer"`.
  - In this case, you MUST include an
    `"answer"` field containing the
    final response to the user.
  - You MUST NOT include an `"action"`
    field when `"type"` is
    `"final_answer"`.
- The `"action"` and `"answer"` fields
    are mutually exclusive. Exactly ONE
    must be present, determined by
    `"type"`.
- Whenever the "TYPE" FIELD is
    `"action"` or `"answer"`, you must
```

```
    include a **"THOUGHT" FIELD**. The
    `thought` string MUST include the
    hierarchical reasoning (Category ->
    Sub-type), and the plan for the next
    action.
```

Here is the prompt for the BreakHis dataset.

Prompt for the BreakHis dataset

```
<image>The pathological image is sampled
    from breast. Determine the most
    likely diagnosis. Which type does it
    belong to?
A.Adenosis
B.Fibroadenoma
C.Phyllodes_tumor
D.Tubular_adenoma
E.Ductal_carcinoma
F.Lobular_carcinoma
G.Mucinous_carcinoma
H.Papillary_carcinoma

If you find challenging to determine the
    exact type, you can first determine
    whether the image belongs to Benign
    or Malignant, and conduct retrieval
    of the subtypes to better make the
    choice.
```

Here is the prompt for the GleasonXAI dataset.

prompt for the GleasonXAI dataset

```
<image>The pathological image is sampled
    from prostate. Determine the most
    likely diagnosis. Which type does it
    belong to?
A.compressed glands
B.individual glands
C.cribriform glands
D.glomeruloid glands
E.poorly formed glands
F.groups of tumor cells
G.comedonecrosis
H.single cells
I.cords

If you find challenging to determine the
    exact type, you can first determine
    which Gleason grade the image
    belongs to, and conduct retrieval of
    the subtypes to better make the
    choice.
```

B Case Study

Figure 3 presents our fine-tuned model's responses to a case from the BreakHis dataset. As shown, the model mistakenly interprets the image as Lobular Carcinoma through direct diagnosis. By integrating our framework of thinking with retrieved images, the model initially interprets the pathological image as Malignant. By actively retrieving visual evidence of the fine-grained subtypes (Ductal carcinoma, Lobular carcinoma, Mucinous carcinoma, Papillary carcinoma) and carefully comparing the query image with the evidence, the model reaches the correct diagnosis, which is Ductal carcinoma.

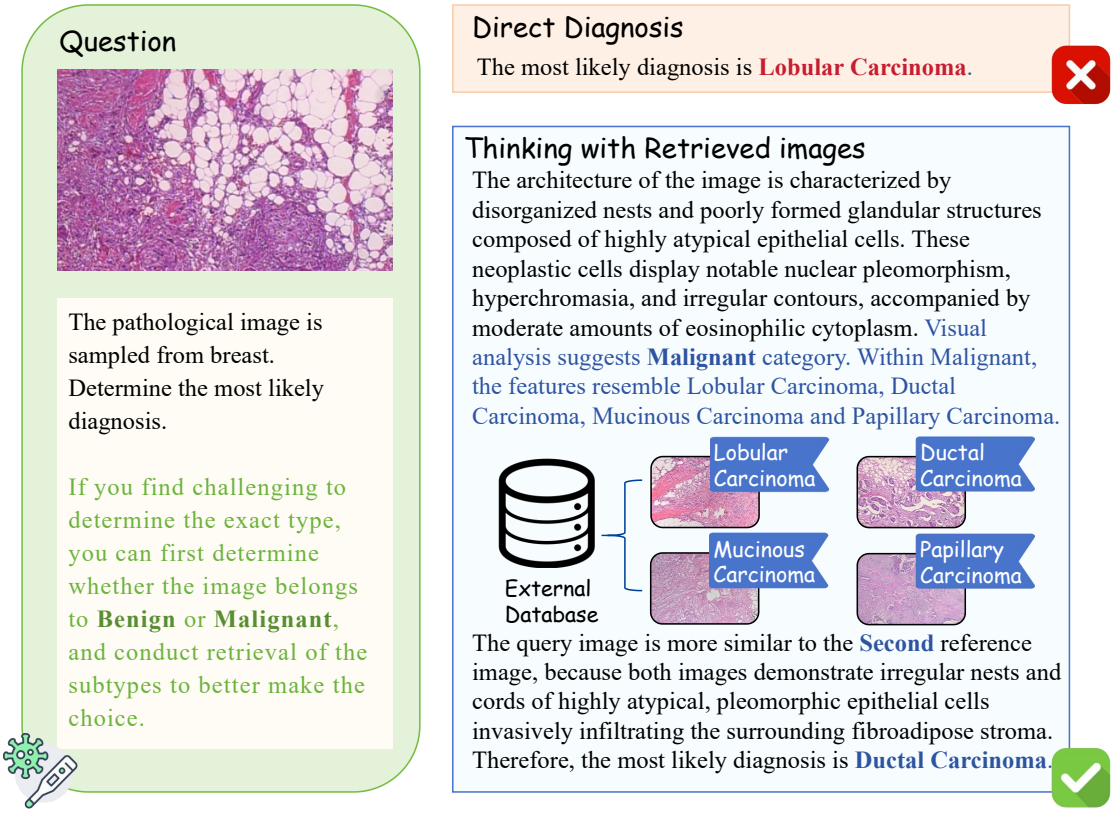

Figure 3: Case illustration of our proposed framework. The ground truth is *Ductal Carcinoma*.