PathoTool: A Tool-Using Agent for Reliability-Aware Pathology Diagnosis

Anonymous ACL submission

Abstract

Recent advances in VLMs have moved toward agentic systems capable of invoking external tools to refine reasoning process. Pathology diagnosis naturally fits this paradigm, as morphological patterns in HE images may require additional immunophenotypic evidence for reliable decision-making. Recent virtual staining methods enable the generation of virtual IHC images from HE, creating a new opportunity for tool-using agents in diagnosis. However, such generated evidence may contain unreliable or conflicting signals, which can interfere with reasoning when directly incorporated. To address this, we propose PathoTool, a tool-using agent for reliability-aware pathology diagnosis. PathoTool performs morphology-based diagnosis from HE images while estimating the confidence, and invokes a virtual staining tool to generate IHC evidence. The HE confidence is further updated through a Confidence Re-evaluation tool conditioned on the generated IHC, which determines whether virtual staining influences the final decision. In addition, an Immunophenotype Conflict Filter is introduced to suppress inconsistent or contradictory IHC signals, ensuring coherent reasoning. Experiments on NSCLC and BRCA classification demonstrate that PathoTool achieves a more balanced and reliable performance improvement.

1 Introduction

Recent advances in vision-language models (VLMs) are driving a shift from passive visual recognition toward agentic systems that can actively invoke external tools to support decision-making under uncertainty. Rather than relying solely on single-step inference from fixed observations, such agents can iteratively acquire, transform, and refine intermediate evidence through tool use, enabling dynamic interaction between reasoning and external operations. This paradigm is particularly valuable in uncertain visual scenarios, where relevant cues may be implicit, weakly expressed, or insufficiently utilized during the initial reasoning process. By revisiting the same input through tool-augmented transformations and auxiliary evidence generation, tool-using agents can improve their ability to surface latent patterns and revise unreliable initial predictions.

Pathology diagnosis provides a representative example of uncertainty-driven visual reasoning. As the gold standard for cancer diagnosis (Bera et al., 2019), in pathology workflow, hematoxylin and eosin (HE) images serve as the primary source of visual evidence for forming an initial diagnostic hypothesis based on tissue morphology (Belsare and Mushrif, 2012). However, morphological cues alone are sometimes insufficient for reliable decision-making due to inter-class similarity, tissue heterogeneity, and ambiguous structural patterns, which introduce substantial uncertainty (Ronen et al., 2021). To address this limitation, immunohistochemistry (IHC) is commonly used as an additional source of complementary signals to help resolve these ambiguous scenarios. Accordingly, pathology diagnosis can be viewed as an iterative refinement process, where an initial morphology-based hypothesis is progressively updated through the acquisition and interpretation of additional evidence under uncertainty.

With the development of virtual staining methods (Li et al., 2023; Klöckner et al., 2025), recent studies (Aggarwal et al., 2025; Zhou et al., 2025; Fischer et al., 2026) have explored incorporating virtual IHC as auxiliary evidence to support pathological diagnosis in vision-based frameworks. This introduces a natural setting for agentic multimodal systems, where an agent can dynamically invoke a generative tool to produce virtual IHC that amplifies morphological cues in HE associated with specific immunophenotypic states, allowing the model to revisit and emphasize diagnostically relevant patterns that may be overlooked or underweighted

during the initial morphology-based inference.

However, the reliability of such tool-generated signals is not always guaranteed (Kataria et al., 2025), as the generation process may introduce artifacts or inconsistent expression patterns (Bai et al., 2023; Huang et al., 2024), and erroneous outputs may in turn mislead the reasoning process when directly incorporated. Moreover, conflicts may arise among multiple generated signals with mutually incompatible semantic interpretations, further complicating decision-making under uncertainty. Therefore, a key challenge in agentic visual reasoning lies not only in invoking external tools to generate auxiliary evidence, but also in coordinating additional tool-based operations to assess the reliability of generated signals, resolve semantic conflicts, and determine how such evidence should influence the final reasoning process.

Motivated by these observations, we propose PathoTool, a tool-using agentic reasoning framework for pathology diagnosis. In PathoTool, a VLM first performs morphology-based diagnosis from HE images while estimating the confidence of the current prediction. To further refine uncertain predictions, the agent can invoke a virtual staining tool to generate virtual IHC images, which are subsequently interpreted by the VLM to extract immunophenotypic evidence.

Inspired by the diagnostic behavior of pathologists, who may reconsider their initial morphological assessment when confronted with strong and internally consistent IHC signals (De Matos et al., 2010; Kohale et al., 2023), the agent will further invoke a Confidence Re-evaluation (Re-Conf) tool to revise the HE-based confidence according to the generated IHC evidence. If the updated confidence remains high, the final decision follows the original HE prediction; otherwise, the prediction is revised according to IHC-guided diagnostic rules.

In addition, different virtual IHC markers may produce mutually incompatible immunophenotypic interpretations, leading to unreliable reasoning signals. To address this issue, the agent can additionally invoke an Immunophenotype Conflict Filter (IC-Filter), which identifies and suppresses conflicting virtual IHC evidence. We conduct systematic experiments on both NSCLC and BRCA classification under patch-level and slide-level settings. Experimental results demonstrate that PathoTool effectively balances morphological reasoning and generated IHC evidence, reducing the adverse impact of virtual staining errors while significantly improving diagnostic reliability and decision consistency in morphologically ambiguous cases.

Our contributions are summarized as follows: (1) We propose PathoTool, an agentic reasoning framework that invokes external tools for reliability-aware pathology diagnosis. (2) We introduce Re-Conf, an explicit confidence re-evaluation mechanism that leverages immunophenotype consistency to refine HE diagnostic confidence. (3) We propose IC-Filter, a principled strategy for handling semantic conflicts among virtual IHC predictions. (4) Experiments on NSCLC and BRCA classification demonstrate the advantages of PathoTool in terms of reliability and interpretability.

2 Related Work

Virtual Staining Virtual staining techniques (Li et al., 2023) enable the synthesis of virtual IHC iamges from HE images, providing an intermediate view that augments morphology-based analysis. Recent approaches (Chen et al., 2024; Liu et al., 2025; Klöckner et al., 2025) are primarily built upon generative models for cross-modality image translation, which have demonstrated promising results in generating visually plausible IHC-like signals. Recent studies (Aggarwal et al., 2025; Zhou et al., 2025; Fischer et al., 2026) have further explored leveraging such virtual IHC signals in purely visual models to support downstream diagnostic modeling, including tumor subtype classification and cellular composition analysis. These approaches demonstrate that virtual staining can serve as a complementary representation that highlights immunophenotype-related cues that are not explicitly encoded in raw HE morphology.

Pathology Agents Recent advances in large multimodal models have led to the emergence of pathology-oriented agents capable of reasoning over whole-slide images (WSIs). These systems typically formulate pathology diagnosis as a sequential decision-making or navigation problem, where the model iteratively selects regions of interest, aggregates information across patches, and produces slide-level predictions. PathChat+ (Weishaupt et al., 2025) introduces a multi-agent copilot design, where multiple specialized components collaborate to refine pathology reasoning through structured evidence integration. CPathAgent (Sun et al., 2026) further develops agentic workflows for WSI analysis, enabling iterative region selection and hierarchical aggrega-

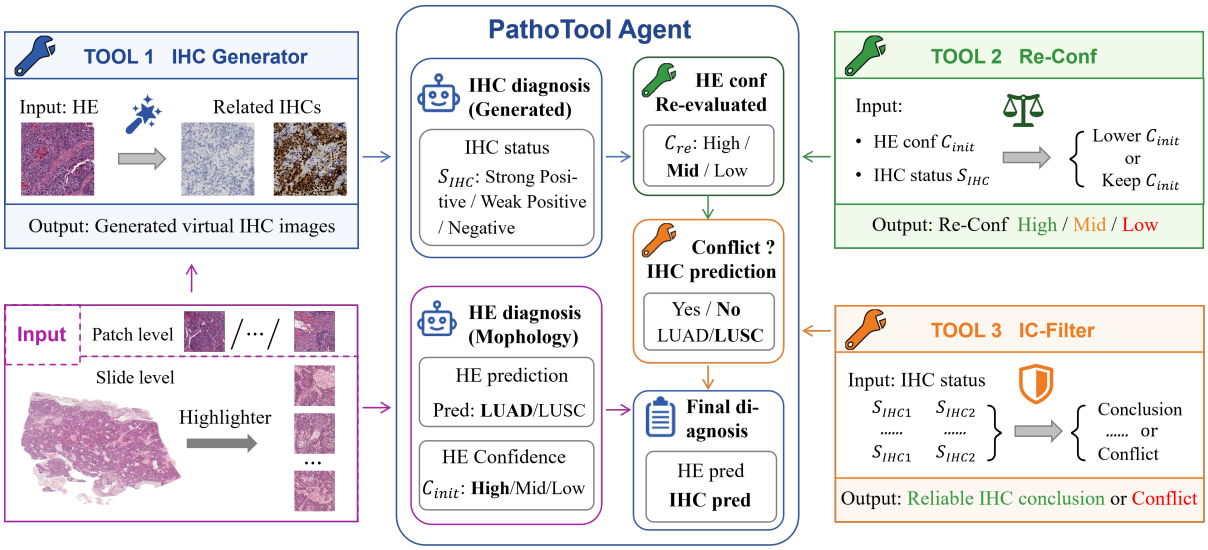

Figure 1: Pipeline for our PathoTool. The PathoTool supports both patch and slide level inputs, where representative diagnostic regions are first sampled by a Highlighter for slide-level inputs. The agent initially performs diagnosis solely based on HE images and provides a confidence C_{init}. Then, our PathoTool analyzes the immunophenotypes of virtual IHC generated by the **IHC Generator**, while **Re-Conf** explicitly re-evaluates the initial HE confidence, and the **IC-Filter** handles the contradictory IHC semantics. Final decisions are made based on the re-evaluated HE confidence C_{re} and the virtual IHC evidence after IC-Filter.

tion of patch-level information. PathFinder (Ghezloo et al., 2025) explores navigation-based reasoning over gigapixel histopathology images, emphasizing sequential exploration of discriminative regions for slide-level diagnosis. However, existing approaches remain constrained to HE images and primarily focus on spatial exploration rather than invoking external tools to generate additional evidence modalities. As a result, their reasoning process is largely bounded by the information explicitly present in the input slide.

Tool-Using Agents Early work on tool-augmented large language models showed that LLMs can move beyond purely language-only reasoning by invoking external tools during problem solving. ReAct (Yao et al., 2022) enables LLMs to interact with knowledge sources or task environments while maintaining interpretable intermediate reasoning. Toolformer (Schick et al., 2023) trains language models in a self-supervised manner to decide when and how to call tools such as calculators, search engines, QA systems, translation systems, and calendars. ToolLLM (Qin et al., 2024) further extendes this capability to open-domain ecosystems encompassing thousands of real-world APIs. More recently, the tool-using agent paradigm has expanded from text-only reasoning to multi-modal reasoning, where agents can jointly perceive visual inputs, plan tool calls, and use external modules for perception, search, code execution, or domain-specific analysis (Wang et al., 2025; Gao et al., 2025; Ashraf et al., 2025; Li et al., 2026). These studies indicate a broader shift from using tools merely as auxiliary text-based APIs toward treating tool usage as a general interface for agentic multimodal reasoning.

3 Methods

PathoTool is a tool-using agentic reasoning framework that reliably integrates virtual IHC into HE-based diagnosis, enhancing diagnostic stability while reducing the impact of staining artifacts. PathoTool first establishes a HE-only diagnostic baseline and estimates the initial confidence from the HE morphology (Sec. 3.1). The agent then invokes an IHC Generator tool to synthesize diagnostically relevant virtual IHC images, which are subsequently analyzed. (Sec. 3.2). To ensure evidence reliability, instead of treating virtual stains as absolutely trustworthy, the agent further invokes the Re-Conf (Sec. 3.3) and IC-Filter tools (Sec. 3.4). These tools reassess the initial confidence in light of the pre-analyzed IHC status and eliminate potential conflicts within the virtual IHC information. Finally, PathoTool fuses HE predictions with IHC analysis, guided by the re-evaluated confidence, to produce the final diagnosis (Sec. 3.5).

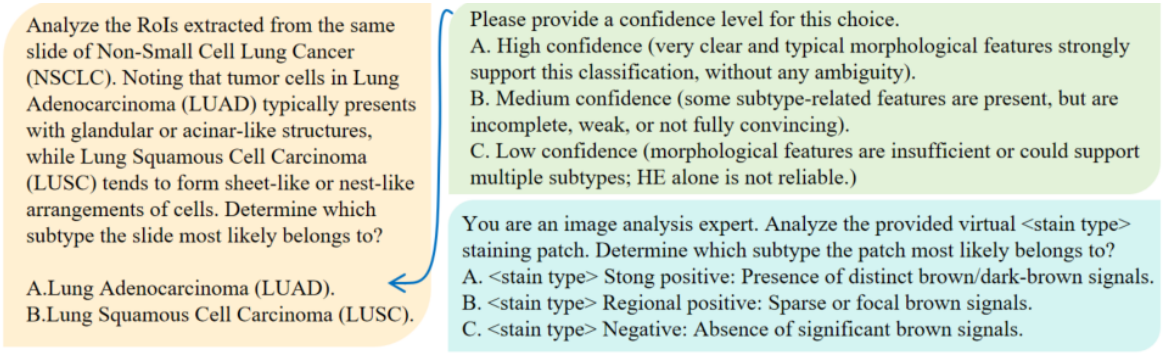

Figure 2: Prompts on NSCLC classification for the HE-based diagnosis (yellow), the initial HE confidence C_{init} (green) and the IHC status S_{IHC} (blue), where <stain type> refers to the IHC related to diagnosis.

3.1 HE-only Diagnosis and Confidence

As HE images provide the fundamental morphological baseline for histopathological assessment, our PathoTool begins with an HE-based diagnosis. The agent is applicable to both patch-level and slide-level settings. For slide-level inputs, a Highlighter module (Hua et al., 2025) is first applied to localize representative diagnostic regions and sample multiple patches. Given the input patch or sampled slide regions, the agent generates an initial morphological diagnostic conclusion alongside a confidence assessment $C_{init} \in \{\text{High}, \text{Mid}, \text{Low}\}$, as prompted in Fig. 2. Considering the potential uncertainty and error introduced by virtual staining, this confidence assessment determines whether the final decision relies on the initial HE prediction or the analysis of generated IHC images.

3.2 IHC Generating and Analysis

IHC Generator. To augment morphological findings, PathoTool invokes the IHC Generator tool to virtual stain IHC images that correspond to specific diagnostic immunophenotypes. For example, in NSCLC classification, we virtual stain TTF-1 and P40, where **TTF-1 positivity** typically supports **LUAD**, and **P40 positivity** supports **LUSC**. Similarly, in BRCA IDC/ILC subtyping, E-cadherin is virtually stained, where **E-cadherin negativity** typically supports **ILC**, whereas both normal glandular tissues and IDC exhibit E-cadherin positivity. These generated virtual IHC images serve to emphasize and amplify critical diagnostic signals that might otherwise be overlooked in the primary HE images. Simultaneously, the analysis of these generated IHC images also provides inputs for the confidence re-evaluation (**Re-Conf**) and immunophenotype conflict filter (**IC-Filter**) module.

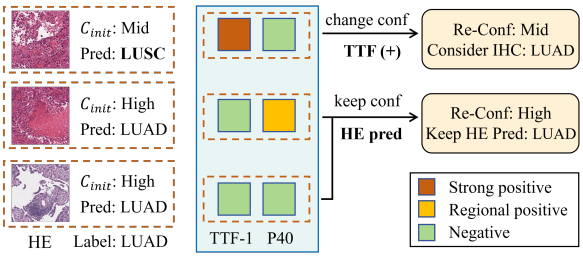

Figure 3: Confidence Re-evaluation. To ensure reliability, Re-Conf re-evaluates the HE-based diagnostic confidence C_{init} according to the virtual IHC evidence.

Generated IHC Analysis. PathoTool employs the same VLM for HE analysis to interpret the generated IHC. As illustrated in Fig. 1, the VLM evaluates each virtual IHC patch to assign an expression status $S_{\text{IHC}} \in \{\text{Strong Positive}, \text{Regional or Weak Positive}, \text{Negative}\}$. Taking NSCLC classification as an example, as prompted in Fig. 2, Strong Positive indicates the presence of distinct brown signals, Regional Positive means sparse or focal brown signals, while Negative indicates the absence of significant brown signals.

3.3 Confidence Re-evaluation

In routine diagnostic practice, strong and internally consistent IHC evidence may occasionally alter a pathologist's confidence in the initial morphology-based diagnosis. Motivated by this observation, PathoTool introduces a Re-Conf tool that reassesses the HE-based diagnostic confidence C_{init} using evidence derived from generated IHC images.

Specifically, as illustrated in Fig. 3, when two semantically complementary IHC markers are available (e.g., TTF-1 and P40), virtual IHC evidence for patch-level inputs is considered strong and internally consistent only when one of the two markers is predicted as Strong Positive while the other is

Negative. In such cases, if the diagnostic implication of the virtual IHC contradicts the HE-based prediction and the C_{init} is High, Re-Conf downgrades the re-evaluated confidence C_{re} to Mid. Otherwise, C_{init} will be preserved. For slide-level inputs comprising a set of patches $\mathcal{P}$, Re-Conf is triggered by PathoTool only if the markers IHC1 and IHC2 satisfy both Eq. (1) and Eq. (2):

$$\exists p \in \mathcal{P}, S^{\mathrm{p}}_{\mathrm{IHC1}} = \text{Strong Positive} \quad (1)$$

$$\forall p \in \mathcal{P}, S^{\mathrm{p}}_{\mathrm{IHC2}} = \text{Negative}. \quad (2)$$

As for the single-marker scenario, the activation criterion of Re-Conf depends on the specific discriminative property of the marker. If the positive status is discriminative, Re-Conf is triggered when Eq. (1) is satisfied. Conversely, for markers where negative expression is discriminative, Re-Conf is triggered when Eq. (2) is satisfied.

By confidence re-evaluation based on IHC status, Re-Conf balances morphological reasoning and virtual IHC evidence. It prevents unreliable virtual staining from unnecessarily perturbing high-confidence HE-based decisions, while also mitigating excessive confidence in erroneous HE predictions by highlighting diagnostically salient IHC cues amplified through virtual IHC generation.

3.4 Immunophenotype Conflict Filter

After the invocation of the IHC Generator by PathoTool, when multiple IHC markers are involved, the pathology VLM in our agent predicts the expression status for each IHC marker. However, different IHC markers may encode mutually exclusive diagnostic semantics (e.g., TTF-1 positive supports LUAD, P40 positive supports LUSC). Due to noise or instability introduced by the IHC Generator tool, the model may simultaneously predict multiple semantically incompatible markers as positive on the same patch. Such cases should be regarded as immunophenotype conflicts and cannot be directly treated as reliable diagnostic evidence.

To address this issue, our PathoTool invokes the Immunophenotype Conflict Filter (IC-Filter) tool to handle conflicting IHC status during reasoning over generated IHC. Within IC-Filter, both Strong Positive and Regional Positive are treated as positive. For each patch, if the predicted IHC result conflicts (e.g., TTF-1 regional positive and P40 strong positive, as illustrated in Fig. 4) , the virtual IHC evidence of that patch is considered to contain intra-patch conflicts and is therefore excluded from subsequent decision making.

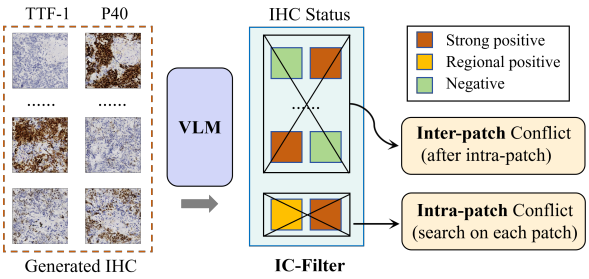

Figure 4: Immunophenotype Conflict Filter. To ensure reliability, IC-Filter searches intra-patch conflicts for each patch. For slide-level input (multiple patches after the Highlighter), IC-Filter further searches for inter-patch conflicts after removing intra-patch conflicts.

For slide-level inputs, conflicts may additionally arise across different patches sampled from the same slide. IC-Filter is first applied independently at the patch level to remove patches with internal IHC conflicts. Among the remaining patches, if one patch exhibits TTF-1 positivity while another exhibits P40 positivity, the slide-level IHC evidence is regarded as contradictory and unreliable, which is called inter-patch conflicts. In this case, the final diagnosis directly follows the HE-based prediction. Notably, inter-patch conflict detection is performed after excluding patches with intra-patch conflicts. And for the single-marker scenario, IC-Filter directly returns the IHC predictions without conflict checking. The IC-Filter ensures that the IHC conclusion for each input is either identified as unreliable or is both self-consistent and unique.

3.5 Final Decision Making

After invoking the IHC generator, Re-Conf, and IC-Filter, PathoTool integrates HE conclusion and virtual IHC evidence to produce the final diagnosis. If the re-evaluated HE-based confidence C_{re} remains High, the model considers the morphological evidence to be sufficiently stable and directly adopts the initial HE prediction, avoiding the potential staining artifacts.

When the re-evaluated confidence C_{re} is Mid or Low, virtual IHC is introduced as auxiliary evidence. After applying IC-Filter to remove all conflicts, the remaining virtual IHC outcomes fall into only two cases: (1) virtual IHC supports a single, non-conflicting diagnostic conclusion. In such case, the final decision follows the IHC-driven conclusion; (2) all relevant markers are negative, providing no discriminative conclusion. The model then falls back to the HE-based prediction.

Table 1: Slide-level quantitative performance on NSCLC classification. "HE only" denotes using only HE input for diagnosis, while "+ stain" indicates that PathoTool invokes the IHC Generator tool on that stain.

Setting	bACC	$\text{Recall}_{\text{LUAD}}$	$\text{Recall}_{\text{LUSC}}$
HE only	80.3	79.1	81.4
+ TTF-1	80.6 (↑0.3)	**84.2** (↑5.1)	77.0 (↓4.4)
+ P40	80.4 (↑0.1)	75.5 (↓3.6)	**85.4** (↑4.0)
+ both	**82.7** (↑2.4)	81.3 (↑2.2)	84.2 (↑2.8)

Table 2: Slide-level quantitative performance on BRCA classification. "HE only" denotes using only HE input for diagnosis, while "+ E-cadherin" indicates that PathoTool invokes the E-cadherin IHC Generator tool.

Setting	bACC	$\text{Recall}_{\text{IDC}}$	$\text{Recall}_{\text{ILC}}$
HE only	58.8	**96.9**	20.7
+ E-cadherin	**70.8** (↑12.0)	91.1 (↓5.8)	**50.5** (↑29.8)

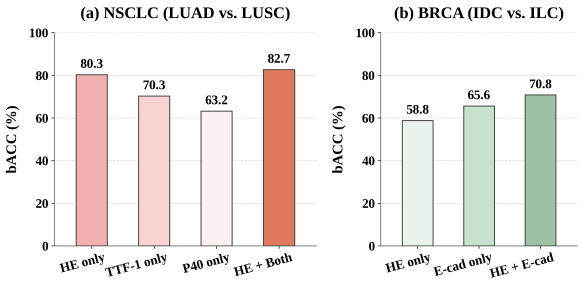

Figure 5: Slide-level comparison with "IHC only" on both NSCLC and BRCA classification. "IHC only" means that the VLM will only process with the generated virtual IHC, without initial HE information.

4 Experiments and Results

Datasets. For slide-level tasks, we conduct experiments on two cohorts from the public TCGA database: TCGA-NSCLC (507 LUAD, 512 LUSC slides) and TCGA-BRCA (802 IDC, 198 ILC slides). For patch-level tasks, we construct datasets from pathologist-annotated tumor regions. For NSCLC, a senior pathologist delineate coarse tumor regions on 25 LUAD and 25 LUSC slides from TCGA. We extract raw patches from these regions and employ the same VLM used in experiments to distinguish tumor from non-tumor patches, thus filtering out non-tumor components. This process yield a balanced test set comprising 10472 tumor patches, with 5237 LUAD and 5235 LUSC. Following the same protocol, we construct an IDC/ILC patch-level test set from 13 IDC and 5 ILC slides in TCGA-BRCA, obtaining 5039 patches, including 3589 IDC and 1450 ILC patches.

Implementation Detail We utilize the VLM backbone from (Hua et al., 2025), which is fine-tuned on Qwen2.5-VL-7B-Instruct (Bai et al., 2025). For slide-level analysis, we sample $N = 5$ representative diagnostic patches for each slide at $10\times$ magnification with a size of 512×512, balancing tissue architecture with cytological detail. We train PyramidPix2pix models (Liu et al., 2022) to generate virtual IHC images. For NSCLC classification, the training datasets are constructed from 18 NSCLC slides with paired TTF-1 and P40 stains. For BRCA subtyping, the training cohort included 10 IDC and 4 ILC slides with the E-cadherin staining. All training datasets are collected from the Hospital X.

4.1 Slide-level Classification

NSCLC classification The slide-level quantitative results on NSCLC classification are summarized in Table 1. When introducing a single virtual IHC marker, the model exhibits a clear class-biased effect that is consistent with the diagnostic semantics of that marker. Specifically, incorporating TTF-1, where TTF-1 positivity indicates LUAD, leads to a notable increase in $\text{Recall}_{\text{LUAD}}$, accompanied by a decrease in $\text{Recall}_{\text{LUSC}}$. Conversely, using P40 alone, where P40 positivity indicates LUSC, slightly improves $\text{Recall}_{\text{LUSC}}$ while marginally reducing $\text{Recall}_{\text{LUAD}}$. These observations suggest that a single virtual IHC marker tends to reinforce the decision toward its corresponding subtype, shifting the recall balance between classes. Nevertheless, both settings yield improvements in bACC over the HE-only baseline, indicating that invoking tools to incorporate virtual IHC evidence can still enhance overall discriminability despite the class-specific trade-offs.

When both TTF-1 and P40 are jointly incorporated under the PathoTool framework, the model achieves the best overall performance, with a 2.4% improvement in bACC over the HE only baseline. Importantly, although the "HE + both" configuration does not yield the maximum single-class Recall growth due to the mechanistic adaptation of tools like Re-Conf and IC-Filter when handling multiple virtual IHC stains, consistent gains are observed across both $\text{Recall}_{\text{LUAD}}$ and $\text{Recall}_{\text{LUSC}}$. This result indicates that invoking tools on multi IHC markers provides complementary IHC evidence, while the Re-Conf and IC-Filter tools effectively mitigate marker-specific biases, enabling more balanced and reliable decisions in morpho-

Table 3: Patch-level quantitative performance on NSCLC classification. "HE only" denotes using only HE input for diagnosis, while "+ stain" indicates that PathoTool invokes the IHC Generator tool on that stain.

Setting	bACC	$\text{Recall}_{\text{LUAD}}$	$\text{Recall}_{\text{LUSC}}$
HE only	76.3	75.2	77.4
+ TTF-1	78.3 (↑2.0)	**82.7** (↑7.5)	73.9 (↓3.5)
+ P40	78.4 (↑2.1)	72.9 (↓2.3)	**84.0** (↑6.6)
+ both	**80.8** (↑4.5)	78.1 (↑2.9)	83.6 (↑6.2)

Table 4: Patch-level quantitative performance on BRCA classification. "HE only" denotes using only HE input for diagnosis, while "+ E-cadherin" indicates that PathoTool invokes the E-cadherin IHC Generator tool.

Setting	bACC	$\text{Recall}_{\text{IDC}}$	$\text{Recall}_{\text{ILC}}$
HE only	54.7	**95.4**	14.1
+ E-cadherin	**76.1** (↑21.4)	76.5 (↓18.9)	**75.7** (↑61.6)

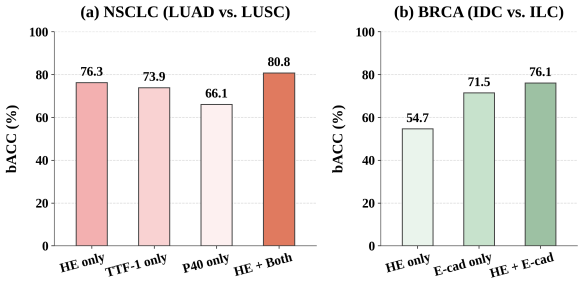

Figure 6: Patch-level comparison with "IHC only" on both NSCLC and BRCA classification. "IHC only" means that the VLM will only process with the generated virtual IHC images, without initial HE information.

logically ambiguous cases.

Concurrently, as illustrated in Fig. 5, we evaluate the diagnostic performance using only the virtual IHC without HE ("IHC only"), and compare it against the "HE only" baseline as well as the full PathoTool ("HE + both") configuration. This comparison serves to reflect the synthesis quality of the IHC Generator and the individual utility of virtual IHC. In the NSCLC classification, "HE + both" achieves the highest bACC, whereas the "IHC only" underperforms compared to "HE only". This suggests that while the IHC Generator broadly captures and synthesizes correct IHC expressions for those HE images where prominent diagnostic signals are clearly presented, it inevitably introduces generative artifacts and staining noise. Crucially, by jointly deploying the IHC Generator, Re-Conf, and IC-Filter, PathoTool effectively rectifies the error introduced by virtual IHC, ultimately yielding more robust and reliable diagnostic outcomes.

BRCA Subtyping To further validate the generalizability of PathoTool, we evaluate it on the challenging BRCA IDC/ILC subtyping task. Clinically, the vast majority of invasive breast cancers are classified as IDC, and pathologists frequently misdiagnose ILC as IDC due to deceptive morphological overlaps. To address this, PathoTool invokes virtual staining for E-cadherin. Distinct from the NSCLC scenario, E-cadherin serves as a negatively discriminative marker, where both normal glands and IDC exhibit strong E-cadherin positivity, while E-cadherin negativity uniquely points to ILC.

In Table 2, incorporating E-cadherin leads to a minor decrease in $\text{Recall}_{\text{IDC}}$, but drives a monumental surge in $\text{Recall}_{\text{ILC}}$ from 20.7% to 50.5%. Concurrently, this targeted enhancement yields a substantial 12.0% substantial improvement in overall bACC over the HE-only baseline.

Furthermore, as illustrated in Fig. 5, the bACC scales progressively across the "HE only", "E-cadherin only", and "HE + both" configurations. This stepwise improvement further demonstrates that in challenging scenarios like BRCA subtyping, the IHC Generator can successfully unearth critical diagnostic cues that are easily overlooked during routine HE interpretation; moreover, PathoTool achieves a subsequent leap in reliability through its complementary tool-invocation mechanism.

4.2 Patch-level Classification

NSCLC classification Patch-level results on NSCLC classification are summarized in Table 3. Similar to the slide-level classification, When virtual TTF-1 and P40 are jointly integrated into the PathoTool framework, all evaluation metrics show consistent improvements, yielding the best bACC. This improvement further benefits from the application of the Re-Conf and IC-Filter mechanisms when multiple IHC markers are incorporated. These results further indicate that, at the patch level, multi-marker virtual IHC also provides complementary constraints and thereby improving overall diagnostic reliability.

BRCA Subtyping At the patch level, invoking tools on E-cadherin within PathoTool yields an even more pronounced performance leap than at the slide level, driving a 61.6% surge in $\text{Recall}_{\text{ILC}}$ and a 21.4% increase in bACC. On one hand, this sharp contrast confirms the overall efficacy of PathoTool; on the other hand, it highlights that compared to

Table 5: Ablation study on the magnification, the presence of C_{init}, Re-Conf and IC-Filter for NSCLC classification, where the VLM will process with both virtual TTF-1 and P40. Values in parentheses denote the performance gain or loss compared to the HE-only baseline. Both experiments on patch and slide-level are conducted.

Task	Mag	C_{init}	Re-Conf	IC-Filter	bACC	$\text{Recall}_{\text{LUAD}}$	$\text{Recall}_{\text{LUSC}}$	$\text{Prec}_{\text{LUAD}}$	$\text{Prec}_{\text{LUSC}}$
Slide	10×	✗	✗	✓	78.9 (↓1.4)	75.5 (↓3.6)	82.2 (↑0.8)	80.8 (0.0)	77.3 (↓2.4)
		✓	✗	✓	81.1 (↑0.8)	**81.9** (↑2.8)	80.3 (↓1.1)	80.4 (↓0.4)	81.7 (↑2.0)
		✓	✓	✗	82.1 (↑1.8)	79.9 (↑0.8)	**84.4** (↑3.0)	83.5 (↑2.7)	80.9 (↑1.2)
		✓	✓	✓	**82.7** (↑2.4)	81.3 (↑2.2)	84.2 (↑2.8)	**83.6** (↑2.8)	**81.9** (↑2.2)
	20×	✗	✗	✓	78.3 (↓1.1)	74.6 (↓1.3)	82.0 (↓0.8)	80.4 (↓1.0)	76.5 (↓1.2)
		✓	✗	✓	80.6 (↑1.2)	**80.9** (↑5.0)	80.3 (↓2.5)	80.2 (↓1.2)	**80.9** (↑3.2)
		✓	✓	✗	80.7 (↑1.3)	76.9 (↑1.0)	**84.4** (↑1.6)	**83.0** (↑1.6)	78.7 (↑1.0)
		✓	✓	✓	**81.4** (↑2.0)	79.5 (↑3.6)	83.4 (↑0.6)	82.6 (↑1.2)	80.4 (↑2.7)
Patch	10×	✗	✗	✓	80.7 (↑4.4)	74.2 (↓1.0)	**87.2** (↑9.8)	**85.3** (↑8.4)	77.2 (↑1.5)
		✓	✗	✓	79.0 (↑2.7)	**80.2** (↑5.0)	77.8 (↑0.4)	78.3 (↑1.4)	**79.7** (↑4.0)
		✓	✓	✗	80.7 (↑4.4)	78.1 (↑2.9)	83.3 (↑5.9)	82.4 (↑5.5)	79.1 (↑3.4)
		✓	✓	✓	**80.8** (↑4.5)	78.1 (↑2.9)	83.6 (↑6.2)	82.7 (↑5.8)	79.2 (↑3.5)
	20×	✗	✗	✓	80.2 (↑4.2)	72.0 (↑0.2)	**88.3** (↑8.0)	**86.9** (↑7.3)	74.6 (↑2.0)
		✓	✗	✓	78.8 (↑2.8)	**77.4** (↑5.6)	80.2 (↓0.1)	80.8 (↑1.2)	**76.8** (↑4.2)
		✓	✓	✗	80.2 (↑4.2)	75.4 (↑3.6)	85.0 (↑4.7)	84.3 (↑4.7)	76.3 (↑3.7)
		✓	✓	✓	**80.3** (↑4.3)	75.4 (↑3.6)	85.2 (↑4.9)	84.6 (↑5.0)	76.4 (↑3.8)

the information-rich slide-level setting, the baseline model heavily struggles to rely solely on HE morphology for patch-level diagnosis, which is further confirmed by Fig. 6.

4.3 Ablation Study

Table 5 presents ablation studies on the magnification, initial confidence C_{init}, Re-Conf and IC-Filter for NSCLC classification. In all configurations, the VLM thinks with both virtual TTF-1 and P40.

For both 10× and 20× magnification, introducing virtual IHC leads to consistent performance gains over the HE-only baseline, indicating that PathoTool is insensitive to input scales and exhibits good robustness and generalization.

At the slide level, omitting C_{init} (all samples are treated as low-confidence and virtual IHC is always incorporated, with Re-Conf naturally disabled) causes performance to drop below the HE-only baseline. Conversely, at the patch level, removing C_{init} still yields substantial bACC gains. This divergence occurs because HE-based predictions at the patch level are far less stable than their slide-level counterparts, allowing the virtual IHC introduced by PathoTool to provide higher marginal returns. In contrast, slide-level HE decisions are relatively reliable, and the indiscriminate incorporation of virtual IHC may inadvertently override correct HE-based judgments due to generative staining noise or artifacts.

When we remove IC-Filter, where the IHC conclusion is determined by the count of positive patches for each marker, the bACC is lower than the PathoTool. This drop highlights the pivotal role of IC-Filter in eliminating IHC contradictions and enhancing evidence reliability. Concurrently, the bACC under this setting still outperforms the HE-only baseline, demonstrating the efficacy of C_{init} and Re-Conf working as a whole.

As for the usage of Re-Conf tool, by integrating C_{init} and IC-Filter. it yields the optimal bACC by dynamically adjusting confidence in a reliability-aware manner. With Re-Conf primarily rectifying over-confident HE predictions and IC-Filter eliminating contradictions among virtual IHC stains, our PathoTool achieves significant and more stable improvements on both LUAD and LUSC.

5 Conclusion

In this work, we present PathoTool, a tool-using agent for reliability-aware pathology diagnosis. PathoTool enables a VLM to invoke external tools during diagnostic reasoning, including an IHC Generator for virtual staining, a Confidence Re-evaluation for adaptive decision adjustment, and an Immunophenotype Conflict Filter for resolving inconsistent IHC signals. Through explicit coordination between morphology-based reasoning and tool-derived evidence, PathoTool allows generated virtual IHC signals to be incorporated in a controlled and reliable manner. Experiments on both NSCLC and BRCA classification demonstrate that PathoTool consistently improves diagnostic reliability under both patch and slide level settings.

Limitations

Although PathoTool improves the reliability of VLM-based pathology diagnosis through tool invocation, several limitations still remain.

First, virtual IHC generation inevitably introduces noise, artifacts, and inaccurate immunophenotypic signals, which may still mislead downstream reasoning. To mitigate this issue, PathoTool selectively incorporates generated evidence through the interaction between the initial confidence C_{init} and the Re-Conf mechanism, allowing the system to reassess and regulate the influence of virtual IHC on the final decision. In addition, IC-Filter further improves reliability by identifying and suppressing explicit conflicts among generated IHC signals. Nevertheless, the overall framework still depends on the quality and stability of the virtual staining module.

Second, the current IC-Filter tool is primarily designed for scenarios involving two or more diagnostically different IHC markers, which is a prevalent clinical reality. Although our experiments demonstrate that PathoTool remains effective in single-marker settings, particularly for challenging BRCA IDC/ILC subtyping where morphology-only diagnosis is difficult, the conflict-filtering capability is sometimes limited.

Finally, the current framework focuses on pathology classification tasks. Extending the tool-using reasoning paradigm to broader pathology scenarios with additional external tools, and more complex clinical workflows remains an important direction for future work.

Ethical Consideration

This work is intended solely for research purposes and is not designed for direct clinical deployment or autonomous medical decision-making. All experiments in this study are retrospective and conducted offline without involvement in real-time patient diagnosis or treatment workflows. For training the virtual staining module, private pathology data were collected under institutional regulations and fully de-identified prior to use, with all personally identifiable information removed to protect patient privacy. No human subjects were contacted or intervened in this study, and all evaluation experiments were conducted exclusively on publicly available datasets. The proposed system is intended to assist diagnostic reasoning rather than replace professional pathologists, and any clinical interpretation or decision should remain under the supervision of qualified medical experts.

References

Arpit Aggarwal, Mayukhmala Jana, Amritpal Singh, Tanmoy Dam, Himanshu Maurya, Tilak Pathak, Sandra Orsulic, Kailin Yang, Deborah Chute, Justin A Bishop, and 1 others. 2025. Artificial intelligence-based virtual staining platform for identifying tumor-associated macrophages from hematoxylin and eosin-stained images. *European Journal of Cancer*, 220:115390.

Tajamul Ashraf, Amal Saqib, Hanan Ghani, Muhra AlMahri, Yuhao Li, Noor Ahsan, Umair Nawaz, Jean Lahoud, Hisham Cholakkal, Mubarak Shah, and 1 others. 2025. Agent-x: Evaluating deep multimodal reasoning in vision-centric agentic tasks. *arXiv preprint arXiv:2505.24876*.

Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. 2023. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

AD Belsare and MM Mushrif. 2012. Histopathological image analysis using image processing techniques: An overview. *Signal & Image Processing*, 3(4):23.

Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715.

Fuqiang Chen, Ranran Zhang, Boyun Zheng, Yiwen Sun, Jiahui He, and Wenjian Qin. 2024. Pathological semantics-preserving learning for h&e-to-ihc virtual staining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–394. Springer.

Leandro Luongo De Matos, Damila Cristina Trufelli, Maria Graciela Luongo De Matos, and Maria Aparecida da Silva Pinhal. 2010. Immunohistochemistry as an important tool in biomarkers detection and clinical practice. *Biomarker insights*, 5:BMI–S2185.

Maximilian Fischer, Alexander Muckenhuber, Robin Peretzke, Luay Farah, Constantin Ulrich, Sebastian Ziegler, Philipp Schader, Lorenz Feineis, Hanno Gao, Shuhan Xiao, and 1 others. 2026. Contrastive virtual staining enhances deep learning-based pdac subtyping from h&e-stained tissue cores. *The Journal of Pathology*, 268(1):89–98.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. 2025. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. In *International Conference on Learning Representations*, volume 2025, pages 13354–13385.

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.

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*.

Luzhe Huang, Yuzhu Li, Nir Pillar, Tal Keidar Haran, William Dean Wallace, and Aydogan Ozcan. 2024. Autonomous quality and hallucination assessment for virtual tissue staining and digital pathology. *arXiv e-prints*, pages arXiv–2404.

Tushar Kataria, Shikha Dubey, Mary Bronner, Jolanta Jedrzkiewicz, Ben J Brintz, Shireen Y Elhabian, and Beatrice S Knudsen. 2025. Building trust in virtual immunohistochemistry: Automated assessment of image quality. *arXiv preprint arXiv:2511.04615*.

Pascal Klöckner, José Teixeira, Diana Montezuma, João Fraga, Hugo M Horlings, Jaime S Cardoso, and Sara P Oliveira. 2025. H&e to ihc virtual staining methods in breast cancer: an overview and benchmarking. *Npj digital medicine*, 8(1):384.

Mangesh G Kohale, Anupama V Dhobale, Nandkishor J Bankar, Obaid Noman, Kajal Hatgaonkar, and Vaishnavi Mishra. 2023. Immunohistochemistry in pathology: A review. *Journal of Cellular Biotechnology*, 9(2):131–138.

Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. 2023. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 632–641. Springer.

Pengxiang Li, Zhi Gao, Bofei Zhang, Yapeng Mi, Xiaojian Shawn Ma, Chenrui Shi, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and 1 others. 2026. Iterative tool usage exploration for multimodal agents via step-wise preference tuning. *Advances in Neural Information Processing Systems*, 38:59496–59528.

Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. 2022. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1815–1824.

Ziwen Liu, Eduardo Hirata-Miyasaki, Soorya Pradeep, Johanna V Rahm, Christian Foley, Talon Chandler, Ivan E Ivanov, Hunter O Woosley, See-Chi Lee, Sudip Khadka, and 1 others. 2025. Robust virtual staining of landmark organelles with cytoland. *Nature Machine Intelligence*, 7(6):901–915.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, and 1 others. 2024. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pages 9695–9717.

Shira Ronen, Rami N Al-Rohil, Elizabeth Keiser, George Jour, Priyadharsini Nagarajan, Michael T Tetzlaff, Jonathan L Curry, Doina Ivan, Lavinia P Middleton, Carlos A Torres-Cabala, and 1 others. 2021. Discordance in diagnosis of melanocytic lesions and its impact on clinical management. *Archives of Pathology & Laboratory Medicine*, 145(12):1505–1515.

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

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Bowen Ding, Tao Lin, and Lin Yang. 2026. 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.

Chenyu Wang, Weixin Luo, Sixun Dong, Xiaohua Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. 2025. Mllm-tool: A multimodal large language model for tool agent learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6678–6687. IEEE.

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, and 1 others. 2025. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964*.

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

Shun Zhou, Yanbo Jin, Jiaji Li, Jie Zhou, Linpeng Lu, Kun Gui, Yanling Jin, Yingying Sun, Wanyuan Chen, Qian Chen, and 1 others. 2025. Uncertainty-assisted virtual immunohistochemical detection on morphological staining via semi-supervised learning. *Optics and Lasers in Engineering*, 184:108657.

A Appendix

A.1 Prompts

A.1.1 NSCLC Classification

Here is the prompt for the slide-level HE diagnosis in NSCLC classification.

Slide-level HE Diagnosis (With confidence for HE diagnosis)

```
Analyze the Regions of Interest (RoIs)
    extracted from the same slide of
    Non-Small Cell Lung Cancer (NSCLC).
    Noting that tumor cells in Lung
    Adenocarcinoma (LUAD) typically
    presents with glandular or
    acinar-like structures, while Lung
    Squamous Cell Carcinoma (LUSC) tends
    to form sheet-like or nest-like
    arrangements of cells. Determine
    which subtype the slide most likely
    belongs to?

A.Lung Adenocarcinoma (LUAD).
B.Lung Squamous Cell Carcinoma (LUSC).

Answer with the option's letter from the
    given choices directly.


<Answer>


Provide a confidence level for this
    choice.
A. High confidence (very clear and
    typical morphological features
    strongly support this
    classification, without any
    ambiguity).
B. Medium confidence (some
    subtype-related features are
    present, but are incomplete, weak,
    or not fully convincing).
C. Low confidence (morphological
    features are insufficient or could
    support multiple subtypes; HE alone
    is not reliable).

Answer with the option's letter from the
    given choices directly.
```

Here is the prompt for the patch-level HE diagnosis in NSCLC classification.

Slide-level HE Diagnosis (With confidence for HE diagnosis)

```
Analyze the patch extracted from a slide
    of Non-Small Cell Lung Cancer
    (NSCLC). Noting that tumor cells in
    Lung Adenocarcinoma (LUAD) typically
    presents with glandular or
    acinar-like structures, while Lung
    Squamous Cell Carcinoma (LUSC) tends
    to form sheet-like or nest-like
    arrangements of cells. Determine
    which subtype the patch most likely
    belongs to?

A.Lung Adenocarcinoma (LUAD).
B.Lung Squamous Cell Carcinoma (LUSC).

Answer with the option's letter from the
    given choices directly.


<Answer>


Provide a confidence level for this
    choice.
A. High confidence (very clear and
    typical morphological features
    strongly support this
    classification, without any
    ambiguity).
B. Medium confidence (some
    subtype-related features are
    present, but are incomplete, weak,
    or not fully convincing).
C. Low confidence (morphological
    features are insufficient or could
    support multiple subtypes; HE alone
    is not reliable.)

Answer with the option's letter from the
    given choices directly.
```

Here is the prompt for the TTF-1 IHC analysis in NSCLC classification.

TTF-1 Analysis

```
You are an image analysis expert.
    Analyze the provided virtual TTF-1
    IHC staining patch. Determine which
    subtype the slide most likely
    belongs to?

A. TTF-1 positive: Presence of distinct
    brown/dark-brown signals in the
    tissue.
B. TTF-1 focal positive: Sparse or focal
    brown signals, including isolated
    small glands C. TTF-1 negative:
    Absence of significant brown signals.

Answer with the option's letter from the
    given choices directly.
```

Here is the prompt for the P40 IHC analysis in NSCLC classification.

P40 Analysis

```
You are an image analysis expert.
    Analyze the provided virtual P40 IHC
```

```
    staining patch. Determine which
    subtype the slide most likely
    belongs to?

A. P40 positive: Presence of distinct
    brown/dark-brown signals in the
    tissue.
B. P40 focal positive: Sparse or focal
    brown signals
C. P40 negative: Absence of significant
    brown signals.

Answer with the option's letter from the
    given choices directly.
```

A.1.2 BRCA IDC/ILC Subtyping

Here is the prompt for the slide-level HE diagnosis in BRCA subtyping.

Slide-level HE Diagnosis (With confidence for HE diagnosis)

```
Analyze the Regions of Interest (RoIs)
    extracted from the same slide of
    Breast Invasive Carcinoma. Noting
    that Invasive Ductal Carcinoma (IDC)
    typically shows cohesive, nest-like
    or cluster-forming tumor cells,
    while Invasive Lobular Carcinoma
    (ILC) typically shows single-file,
    linear or loosely discohesive tumor
    cells. Determine which subtype the
    slide most likely belongs to?

A.Invasive Ductal Carcinoma (IDC).
B.Invasive Lobular Carcinoma (ILC).

Answer with the option's letter from the
    given choices directly.


<Answer>


Provide a confidence level for this
    choice.

A. High confidence (very clear and
    typical morphological features
    strongly support this
    classification, without any
    ambiguity).
B. Medium confidence (some
    subtype-related features are
    present, but are incomplete, weak,
    or not fully convincing).
C. Low confidence (morphological
    features are insufficient or could
    support multiple subtypes; HE alone
    is not reliable).

Answer with the option's letter from the
    given choices directly.
```

Here is the prompt for the patch-level HE diagnosis in BRCA subtyping.

Patch-level HE Diagnosis (With confidence for HE diagnosis)

```
Analyze the patch extracted from a slide
    of Breast Invasive Carcinoma. Noting
    that Invasive Ductal Carcinoma (IDC)
    typically shows cohesive, nest-like
    or cluster-forming tumor cells,
    while Invasive Lobular Carcinoma
    (ILC) typically shows single-file,
    linear or loosely discohesive tumor
    cells. Determine which subtype the
    patch most likely belongs to?

A.Invasive Ductal Carcinoma (IDC).
B.Invasive Lobular Carcinoma (ILC).

Answer with the option's letter from the
    given choices directly


<Answer>


Provide a confidence level for this
    choice.

A. High confidence (very clear and
    typical morphological features
    strongly support this
    classification, without any
    ambiguity).
B. Medium confidence (some
    subtype-related features are
    present, but are incomplete, weak,
    or not fully convincing).
C. Low confidence (morphological
    features are insufficient or could
    support multiple subtypes; HE alone
    is not reliable).

Answer with the option's letter from the
    given choices directly.
```

Here is the prompt for the E-cadherin analysis

E-cadherin Analysis

```
You are an image analysis expert.
    Analyze the provided virtual
    E-cadherin IHC staining patch.
    Determine which subtype the slide
    most likely belongs to?

A. E-cadherin positive: Presence of
    distinct yellow, brown or dark-brown
    signals in the tissue.
B. E-cadherin focal positive: Sparse or
    focal brown/yellow signals.
C. E-cadherin negative: Absence of
    significant yellow or brown signals.

Answer with the option's letter from the
    given choices directly.
```

A.2 Case Study

Fig. 7 represents a case where the C_{init} changes after Re-Conf, the initial HE diagnosis is wrong, and the virtual IHC corrects the HE diagnosis.
Fig. 8 represents a case where the C_{init} is unchanged after Re-Conf, and the initial HE diagnosis is correct.

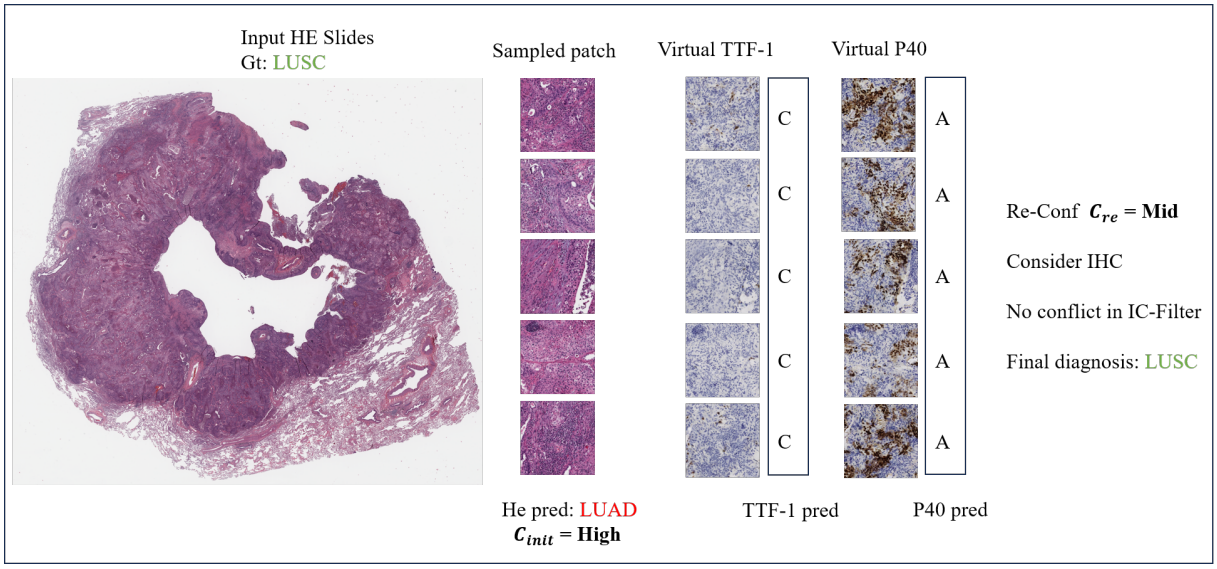

Figure 7: Case for C_{init} changed. In Slide-level NSCLC classification, the C_{init} is High confidence, and the C_{re} changes to Medium confidence after Re-Conf.

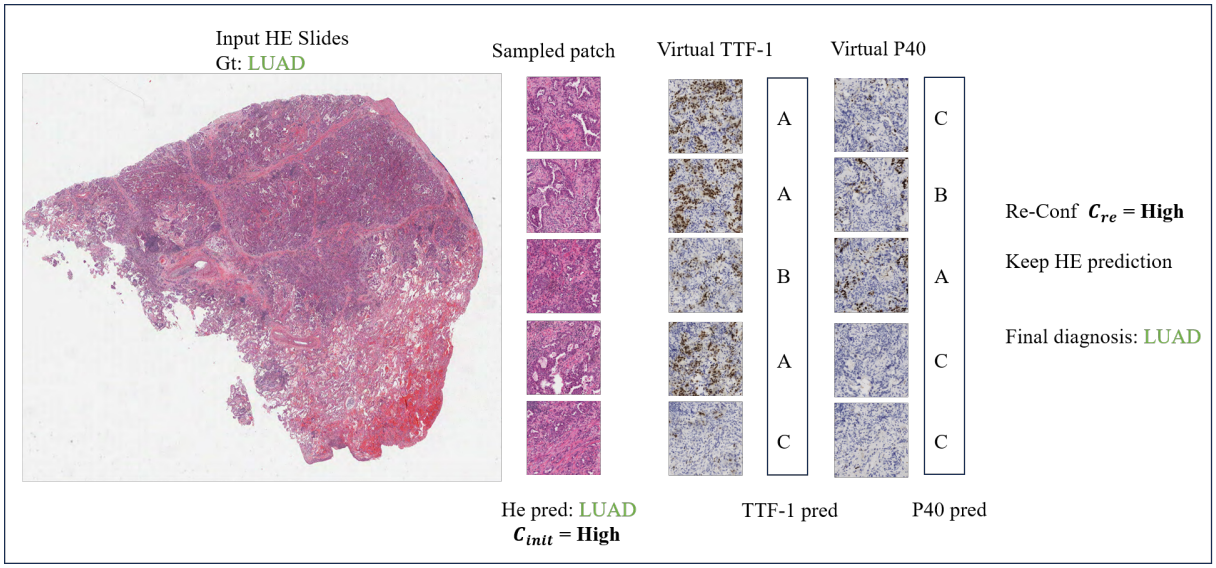

Figure 8: Case for C_{init} unchanged. In Slide-level NSCLC classification, the C_{init} is High confidence, and the C_{re} is still High confidence after Re-Conf.