



PathFinder: A Multi-Modal Multi-Agent System for Medical Diagnostic Decision-Making Applied to Histopathology

Supplementary Material

7. Triage agent

Figure 5 illustrates the architecture of the Triage Agent. To evaluate its effectiveness, we compared the performance of the Triage Agent against three MIL-based benchmark methods [25, 27, 45] for detecting Class 1 vs. Non-Class 1 cases in the M-Path dataset (details in Section 3). As summarized in Table 2, PathFinder’s Triage Agent, designed to assess whether a WSI is risky, outperforms the baseline methods.

Method	Class 1 F1	Non-Class 1 F1	Overall Accuracy
AMIL [25]	0.16	0.83	0.71
DSMIL [27]	0.35	0.86	0.77
TransMIL [45]	0.40	0.90	0.83
Triage Agent	0.57	0.95	0.91

Table 2. Comparison of Triage Agent with benchmark methods on Class 1 vs. Non-Class 1 classification. We report class-specific F1 due to imbalanced nature of the M-Path dataset.

8. VLM-based Navigation Agent

Our initial approach to designing the Navigator Agent explored a multi-modal architecture based on the LLaVA framework [31]. This design aimed to enable direct reasoning over image latents through an LLM. The architecture consisted of two main components:

1. A U-Net encoder [42] pre-trained on pathologist viewing behavior data (M-Path, details in Section 3), which served as the image encoder
2. The LLaMA-7B language model [48], which acted as the reasoning component

8.1. Training Process and Architecture

We first trained a complete U-Net on the M-Path dataset to learn meaningful representations of WSIs. For the Navigator implementation, we removed the U-Net’s decoder and retained only the encoder portion. This encoder was then integrated with LLaMA-7B following the LLaVA framework. The combined model was trained using instruction tuning, where each training instance consisted of:

- Input: A WSI and a list of previous observations and their descriptions obtained from the Description agent
- Output: Grid coordinates (row and column) identifying regions of interest within the WSI

The underlying hypothesis was that the LLM could effectively process the U-Net-encoded latent representations

to identify diagnostically relevant grid coordinates directly.

8.2. Limitations and Challenges with a LLaVA-based Navigator

This approach encountered several significant limitations:

1. **Data Scarcity:** The available navigation training dataset proved insufficient for the model to learn robust region selection strategies.
2. **Overfitting Patterns:** The model exhibited clear signs of overfitting:
 - Consistently selecting patches from the central regions of WSIs, regardless of input
 - Generating repetitive patch selections
 - Failing to generalize to novel slide patterns

8.3. Architectural Pivot

These limitations led us to revise our approach fundamentally. Instead of requiring the LLM to reason directly from latent representations, we returned to utilizing the complete U-Net architecture (including the decoder), and leverage the decoded attention maps for direct region sampling. This proved to be more robust with limited training data, and we simply conditioned our U-Net with the descriptions from the Description Agent to have the feedback loop between the agents. This experience highlighted the challenges of applying LLMs to specialized medical tasks with constrained training data, even when pre-training sub-modules (like our U-Net encoder in this case).

9. Description agent

We generated fine-tuning data for the Description Agent by prompting GPT-4 to extract short and concise histopathology findings from provided text. Figure 6 illustrates the prompt used and a sample of the data generated for fine-tuning the Description Agent.

10. Training Details of the Diagnosis Agent

We expand the training set to enhance diversity and robustness by resampling to create 20,000 cases, resulting in 100,000 trajectories for training. Each trajectory consists of a randomly selected number of descriptions (between five and ten), and we shuffle the sequence of descriptions within each trajectory to prevent over-fitting to any specific order. Each trajectory is formatted as a prompt to the LLM:

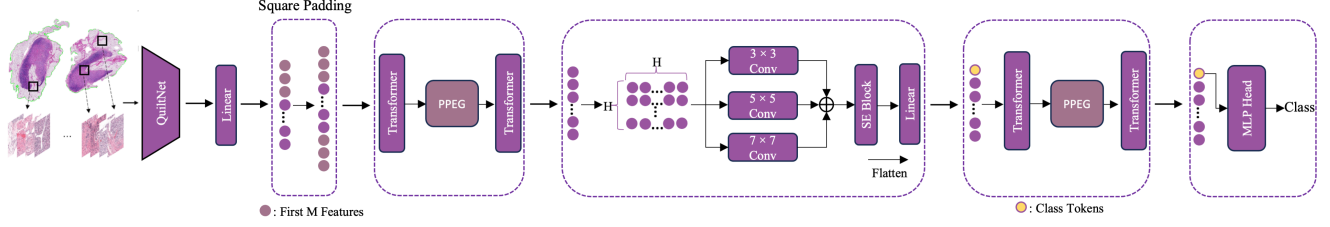


Figure 5. Overview of the Triage Agent architecture. Definitions of M and H can be found in Section 4.

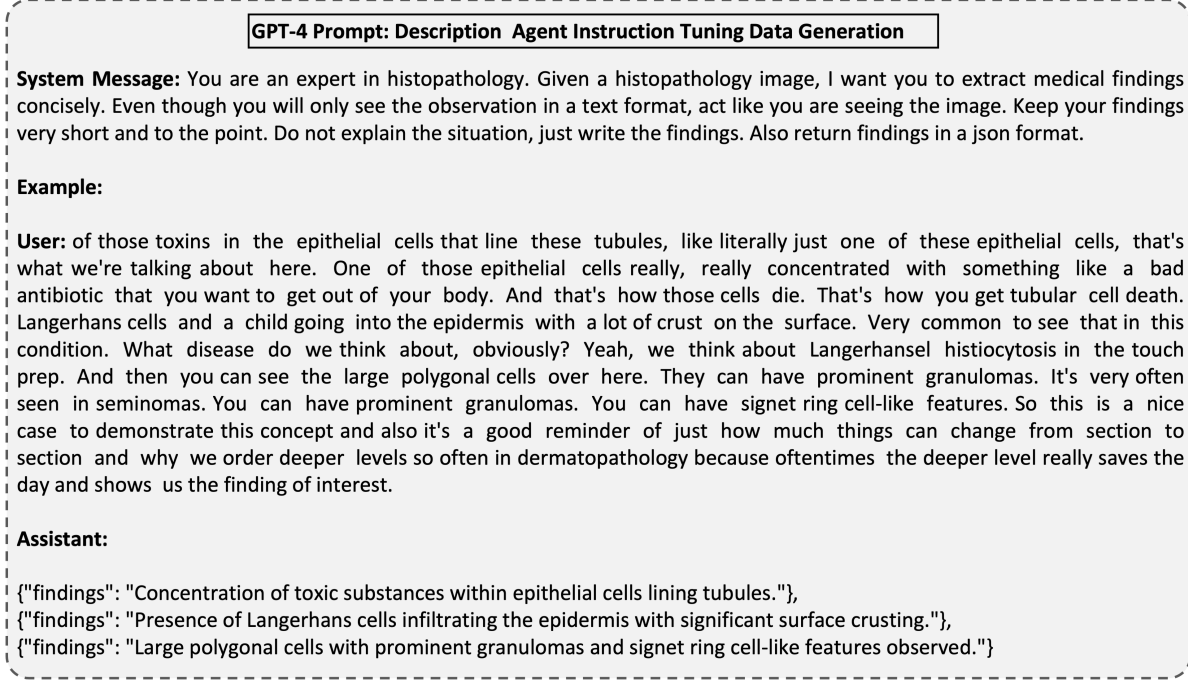


Figure 6. GPT-4 prompt to generate instruction-tuning dataset for the Description Agent.

“The image descriptions below are extracted from different patches from the same WSI; please tell me which class the image belongs to: descriptions”, where *descriptions* is the list of selected descriptions.

We fine-tune the LLM using LoRA (Low-Rank Adaptation) [17] with the scaling factor $\alpha = 8$, dropout rate 0.1, and rank parameter $r = 8$ in the LoRA layers. The model is trained using cross-entropy loss, with a learning rate of 5×10^{-5} , weight decay 0.001, and batch size 16.

11. Evaluation and experiments

This section provides details on the qualitative analysis conducted by pathologists and the prompt for our LLM-prompting experiments.

11.1. Qualitative Analysis of Descriptions Assessed by Pathologists

To evaluate the quality of the descriptions generated by the Description Agent, we cropped the region of interest from 25 WSIs from M-Path dataset and generated descriptions for these regions using three models: PathFinder’s Description Agent, GPT-4o, and LLaVa-Med. Figure 7 presents a few sample cases. Since our Description Agent is fine-tuned to produce short and concise descriptions, we ensured a fair comparison by prompting LLaVa-Med and GPT-4o with the instruction: *Describe the histology image concisely in less than 20 words*. We conducted a survey involving two pathologists who were asked to answer the following two questions regarding descriptions produced by the three models. The study was conducted in a double-blind, ran-

domized manner to ensure unbiased results:

1. **Selection:** Please select the description that you believe best matches the content of the image. (Options: Model A, Model B, Model C)
2. **Reason for Preference:** Please choose the primary reason for your preference. You may select more than one option if applicable. If Other, please specify.
 - **Correctness:** The description accurately reflects the features of the image.
 - **Detail:** The description provides a comprehensive analysis of the image.
 - **Relevance:** The description emphasizes the most pertinent aspects of the image.
 - **Other:** Please specify.

Figure 8 illustrates the distribution of reasons selected by pathologists for preferring each model. As shown, none of the models were preferred for their level of detail, which aligns with expectations since the models were specifically prompted to generate short and concise descriptions, inherently limiting detailed information. The majority of preferences were based on the correctness of the descriptions.

11.2. Prompt used for pre-trained LLM experiments

The following prompt was used in our experiments with pre-trained LLMs serving as the Diagnosis Agent to make a diagnosis based on the provided *descriptions*:

Prompt: Answer the following question related to skin cancer. Only use one of the four options given at the end.

The image descriptions below are extracted from different patches from the same whole slide image (WSI), please tell me which class the image belongs to:

{*descriptions*}

The options are:

”diagnosis: (I) mildly dysplastic nevi, moderately dysplastic nevi”

”diagnosis: (II) melanoma in situ and severely dysplastic nevi”

”diagnosis: (III) invasive melanoma stage pT1a”

”diagnosis: (IV) advanced invasive melanoma stage \geq pT1b”

Only output the complete text of the option you choose. Don’t add any more words.

12. Imitated Sampling Implementation

To simulate the navigation behavior of expert pathologists, we developed an *Imitated Sampling* algorithm that generates patch sequences reflecting human visual exploration of whole slide images (WSIs). This method mimics expert viewing patterns by modeling statistical distributions across multiple behavioral dimensions.

12.1. Data Collection and Distribution Modeling

We analyzed the viewing behavior of 12 board-certified pathologists as they reviewed 45 histopathology WSIs. From this, we extracted statistical distributions for:

- Zoom level frequencies
- Patch size distributions per zoom level
- Spatial transitions between consecutive patches
- Attention duration on specific regions

For each zoom level $z \in \{1, 5, 10, 20, 40, 50\}$, patch width and height were modeled as normal distributions:

$$w \sim \mathcal{N}(\mu_w^z, \sigma_w^z), \quad h \sim \mathcal{N}(\mu_h^z, \sigma_h^z)$$

with values clipped to the empirically observed ranges $[w_{\min}^z, w_{\max}^z]$ and $[h_{\min}^z, h_{\max}^z]$.

12.2. Algorithmic Implementation

The Imitated Sampling algorithm proceeds in the following steps:

Patch Count Determination: For each zoom level z , we sample the number of patches n_z using:

$$n_z = \mathcal{N}(\mu_n, \sigma_n) \times p_z$$

where $\mu_n = 278.48$, $\sigma_n = 131.64$, and p_z denotes the empirical frequency of patches at zoom level z .

Foreground Segmentation: Foreground tissue is segmented using a multi-stage thresholding and morphological pipeline:

$$M_{\text{bg}} = \Phi_{\text{thresh}}(I, \tau_{\text{lower}}, \tau_{\text{upper}})$$

$$M_{\text{fg}} = \neg M_{\text{bg}}$$

$$M_{\text{refined}} = \Phi_{\text{close}}(M_{\text{fg}}, K)$$

where Φ_{thresh} is a pixel intensity thresholding function, \neg denotes logical negation, and Φ_{close} is a morphological closing operation with kernel K .

Patch Generation: For each zoom level:

1. Sample patch dimensions:

$$w \sim \mathcal{N}(\mu_w^z, \sigma_w^z), \quad h \sim \mathcal{N}(\mu_h^z, \sigma_h^z)$$

2. Randomly select a center coordinate (x, y) from the foreground region Ω_{fg} :

$$(x, y) \sim \text{Uniform}(\Omega_{\text{fg}})$$

3. Accept the patch if it satisfies a tissue content threshold:

$$\frac{|\{(i, j) \in R_{x, y, w, h} : M_{\text{refined}}(i, j) \neq 0\}|}{w \times h} \geq \frac{t_z}{100}$$

where t_z is a zoom-dependent tissue coverage threshold (ranging from 15% at zoom 1 to 80% at zoom 50).

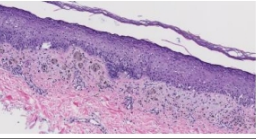
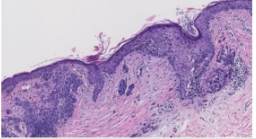
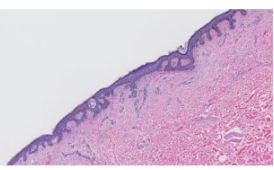
ROI Patch	PathFinder's Description Agent	LLaVA-Med	GPT-4o
	Melanin pigment incontinence in the epidermis, indicative of a fixed drug eruption.	The histology image shows a well-differentiated squamous cell carcinoma with a keratin pearl.	Hyperkeratotic epidermis with acanthosis, dermal fibrosis, and inflammatory infiltrate.
	Melanoma in situ with pagetoid spread and melanocytes in the epidermis.	The histology image shows a well-differentiated squamous cell carcinoma with a keratin pearl.	Epidermal hyperplasia with acanthosis, hyperkeratosis, and dermal inflammatory infiltrate.
	Normal skin with intact basal layer and no significant pathology.	The histology image shows a well-differentiated squamous cell carcinoma with a keratin pearl.	Epidermis with hyperkeratosis, acanthosis, and dermal papillae; underlying dermis with collagen fibers and blood vessels.

Figure 7. A comparison of descriptions generated by PathFinder’s Description Agent, GPT-4o, and LLaVa-Med for sample ROI patches selected by M-Path’s consensus panel of pathologists.

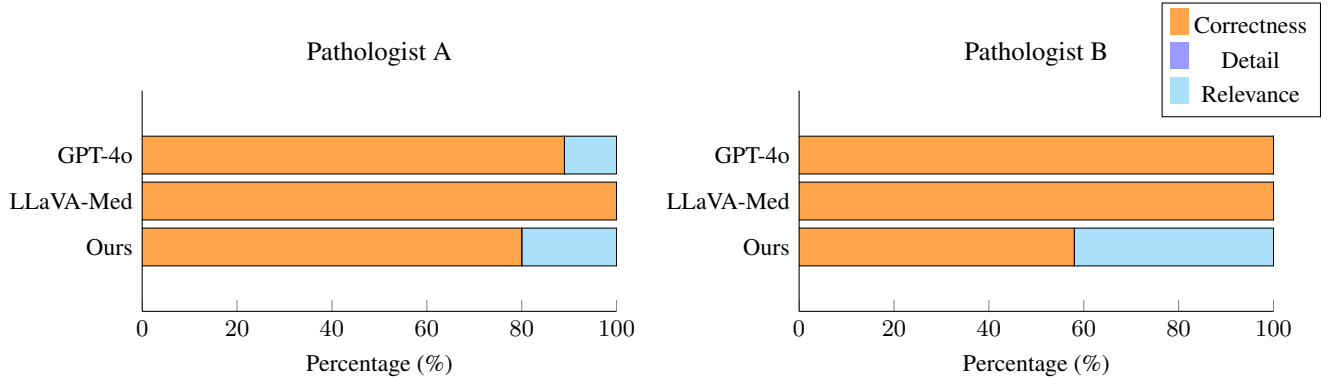


Figure 8. Expert human pathologist preferences for each model, segmented by the reasons for their choices. Each subplot corresponds to one pathologist and shows their ratings for PathFinder (Ours), LLaVA-Med, and GPT-4o.

Overlap Removal: To model realistic attention shifts and avoid redundant sampling, we enforce a maximum overlap criterion:

$$\frac{|R_i \cap R_j|}{\min(|R_i|, |R_j|)} \leq 0.3$$

where R_i and R_j are patch regions.

Coordinate Transformation: Patch coordinates are mapped back to the original image space:

$$X = x \cdot z, \quad Y = y \cdot z, \quad W = w \cdot z, \quad H = h \cdot z$$

12.3. Implementation Details

The implementation adapts thresholding parameters based on zoom level, reflecting the increasing specificity required

at higher zoom levels. A 15×15 kernel is used for morphological operations, selected to balance noise suppression and tissue continuity.

To improve sampling reliability, the system includes retry counters for edge-related failures ($n_{\text{retry}}^{\text{edge}}$) and content-based rejections ($n_{\text{retry}}^{\text{content}}$). This ensures robustness in difficult-to-segment regions.

To further emulate expert pathologists’ adaptive exploration, we implemented a residual patch propagation mechanism:

$$n_{z+1} = n_{z+1} + (n_z^{\text{target}} - n_z^{\text{actual}})$$

This simulates the human tendency to increase magnification when low-resolution views are inconclusive.

The final output is a sequence of $N = \sum_z n_z$ patches denoted as $(X_i, Y_i, W_i, H_i, z_i)$ for $i = 1, 2, \dots, N$, capturing the key characteristics of expert navigation: tissue prioritization, adaptive zooming, and spatial progression. Compared to uniform or random sampling, Imitated Sampling yields more representative and explainable trajectories for downstream WSI analysis.

13. Advantages and Limitations

Efficiency: While our approach prioritizes diagnostic accuracy and interpretability, it is also designed with computational efficiency in mind. Rather than exhaustively processing all possible regions in a WSI, which is often computationally prohibitive, PathFinder strategically selects and processes only 50 patches per WSI. This targeted selection significantly reduces the overall computational load and inference time, enabling faster analysis while maintaining strong diagnostic performance. By avoiding exhaustive search, PathFinder strikes a practical balance between accuracy, interpretability, and efficiency.

System Complexity: Although PathFinder introduces more architectural complexity compared to simpler, single-model baselines such as CONCH and MUSK, this added complexity comes with clear benefits. PathFinder’s modular design offers enhanced flexibility and interpretability. Each agent operates independently and can be evaluated, fine-tuned, or replaced without retraining the entire pipeline, making the system more adaptable to different datasets and tasks. Despite its multi-agent setup, PathFinder achieves strong performance, outperforming the best baseline by 9% in diagnostic accuracy while also generating human-interpretable explanations through descriptive patch-level outputs. This combination of accuracy and transparency makes PathFinder a compelling solution for high-stakes medical AI applications.

Generalizability: Identifying high-quality public datasets suitable for WSI-level classification remains a challenge, particularly due to the scarcity of datasets with reliable, slide-level diagnostic labels. We evaluated our pipeline on the TCGA-SKCM dataset. However, it is important to note that its diagnostic labels are derived from broader clinical context and metadata, rather than solely from histological features present in the WSI. As a result, classification based on WSI content alone is inherently difficult. The best-performing baseline on this dataset achieved an accuracy of 0.52. In comparison, our pipeline, when training only the Diagnosis Agent, achieved an accuracy of 0.54 using exhaustive patch search, and 0.50 using the T5-Navigator for patch selection. Notably, in addition to competitive accuracy, our pipeline provides explainability through textual descriptions of selected patches. Furthermore, our Description Agent is trained on Quilt-1M, a diverse dataset covering 12 cancer types, and the modular ar-

chitecture of our agents supports easy transfer and adaptation to new datasets, highlighting the potential for generalizability across tasks and domains.

Hallucination Risk in MLLMs: To mitigate this issue, we avoid general-purpose MLLMs and instead use a domain-specific model. Our Description Agent is based on Quilt-LLaVA, fine-tuned on histopathology data to produce concise and factual descriptions. These descriptions are not generated with zero-shot prompting but through supervised instruction tuning. Our Diagnosis Agent is trained (not prompted) on these domain-specific descriptions for classification, ensuring that the diagnosis is grounded in learned mappings rather than open-ended language generation.