



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

Fatemeh Ghezloo<sup>1,2\*</sup> Mehmet Saygin Seyfioglu<sup>1\*</sup> Rustin Soraki<sup>1\*</sup>  
 Wisdom O. Ikezogwo<sup>1\*</sup> Beibin Li<sup>2\*</sup> Tejoram Vivekanandan<sup>1</sup>  
 Joann G. Elmore<sup>3</sup> Ranjay Krishna<sup>1,4</sup> Linda Shapiro<sup>1</sup>  
<sup>1</sup> University of Washington <sup>2</sup> Microsoft Research  
<sup>3</sup> David Geffen School of Medicine, UCLA <sup>4</sup> Allen Institute for AI (AI2)

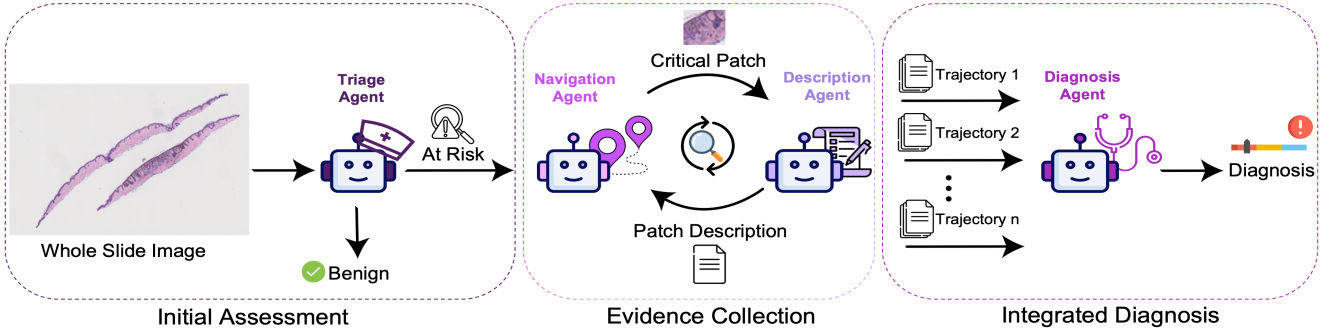


Figure 1. We propose **PathFinder**, a system capable of navigating image patches in a whole slide image, describe each patch to collect evidence, and produce a diagnosis. Pathfinder’s process is interpretable and reminiscent of pathologists. Our system consists of multiple steps carried out by multi-modal agents: 1) Initial Assessment by Triage Agent; 2) Evidence Collection by Navigation and Description Agents; and 3) Integrated Diagnosis by Diagnosis Agent.

## Abstract

Diagnosing diseases through histopathology whole slide images (WSIs) is fundamental in modern pathology but is challenged by the gigapixel scale and complexity of WSIs. Trained histopathologists overcome this challenge by navigating the WSI, looking for relevant patches, taking notes, and compiling them to produce a final holistic diagnostic. Traditional AI approaches, such as multiple instance learning and transformer-based models, fail short of such a holistic, iterative, multi-scale diagnostic procedure, limiting their adoption in the real-world. We introduce *PathFinder*, a multi-modal, multi-agent framework that emulates the decision-making process of expert pathologists. *PathFinder* integrates four AI agents—the Triage Agent, Navigation Agent, Description Agent, and Diagnosis Agent—that collaboratively navigate WSIs, gather evidence, and provide comprehensive diagnoses with natural language explanations. The Triage Agent classifies the WSI as benign or risky; if risky, the Navigation and Description Agents iteratively focus on significant regions, gen-

erating importance maps and descriptive insights of sampled patches. Finally, the Diagnosis Agent synthesizes the findings to determine the patient’s diagnostic classification. Our Experiments show that *PathFinder* outperforms state-of-the-art methods in skin melanoma diagnosis by 8% while offering inherent explainability through natural language descriptions of diagnostically relevant patches. Qualitative analysis by pathologists shows that the Description Agent’s outputs are of high quality and comparable to GPT-4o. *PathFinder* is also the first AI-based system to surpass the average performance of pathologists in this challenging melanoma classification task by 9%, setting a new record for efficient, accurate, and interpretable AI-assisted diagnostics in pathology. Data, demo, code and models are available at <https://pathfinder-dx.github.io/>.

## 1. Introduction

Medical diagnosis of histopathology through the examination of whole slide images (WSIs) is a cornerstone of modern pathology. WSIs are high-resolution, digitally scanned

histopathology cases, providing an extensive view of tissue architecture and cellular detail. Pathologists *navigate* these gigapixel-scale images to identify morphological features and spatial relationships critical for accurate diagnoses. They start with a low magnification to identify suspicious regions and then zoom into image patches for detailed examination [11, 32]. They gather evidence across patches and accumulate them together to make a final holistic diagnosis. This process is the gold standard. However, it is labor-intensive and requires significant expertise to interpret complex visual information effectively. It is becoming increasingly unsustainable due to the rising number of cancer cases globally.

The shift towards more efficient diagnostic methods in medical imaging is essential, yet must maintain accuracy. Recent advancements in deep learning report achieving expert-level performance, promising such a scalable approach [47]. However, current methods typically divide WSIs into smaller patches for independent analysis, making diagnoses without the holistic context [1, 13, 23, 26, 44, 46, 53, 54, 56]. Transformer-based models attempt to capture both local and global patterns but are not scalable with the high-resolution demands of WSI [3, 14, 45, 52, 55].

In contrast, we propose PathFinder, a multi-modal and multi-agent system designed to mimic the decision-making process of expert pathologists by integrating four AI agents: Triage, Navigation, Description, and Diagnosis. The system begins with the Triage Agent, classifying the WSI as benign or risky; if risky, the Navigation and Description Agents iteratively examine patches, generating natural language descriptions and refining their focus with each cycle. Finally, these detailed insights are integrated by the Diagnosis Agent to produce an accurate and holistic diagnostic classification. Figure 1 demonstrates an overview of PathFinder’s pipeline.

Our experiments demonstrate that the proposed agentic system significantly outperforms prior state-of-the-art (SOTA) methods on WSI skin melanoma grading, achieving an accuracy of 74% on the M-Path Skin Biopsy dataset [9]. This marks an 8% improvement over the best baseline with accuracy of 66% and a 9% improvement over the 65% average performance of pathologists [9]. Our proposed system is also fully explainable from the patches visited to the description of the patches and the final diagnoses, which takes into consideration all the patch-wise information. To the best of our knowledge, PathFinder is the first AI-based system capable of surpassing the average performance of pathologists on this challenging melanoma classification task.

## 2. Related Work

**Multi-modal Histopathology Models.** There have been a series of studies in histopathology that leverage WSI-level

and patch-level images to train unimodal classifiers based on multiple instance modeling leveraging pretrained feature extractors [35, 45, 51]. More recently unimodal foundational models trained on varying self-supervised objectives have achieved significant improvements on performance downstream [4, 22, 49, 53]. With the introduction of large-scale multi-modal datasets in histopathology, we have seen significant advancements, with the emergence of large language models and vision-language models for histopathology. For instance, studies like Quilt-1M [23] and PathGen-1.6M [46] curate large histopathology image-text paired dataset and train CLIP-based models to learn joint vision-language representations, significantly enhancing clinical histology downstream tasks on patch-level. On the WSI-level, PathAlign [1] aligns diagnostic texts from pathology reports with corresponding WSIs, facilitating applications such as automatic report generation and case/patient-level visual question answering, moving towards a more clinically integrated and holistic diagnostic process. While many other studies like Quilt-LLaVA [44], SlideChat [5], and PathChat [34] train histopathology Multi-modal Large Language models (MLLM) and improve on diagnostic reasoning tasks, none of these models effectively automatically navigate the giga-pixel scale WSIs towards a diagnosis.

### **The role of Navigation in Histopathology Diagnosis.**

Computational pathology studies have tried to capture and analyze the navigation patterns of pathologists when reviewing digital slide images [11, 36, 37, 41] specifically characterizing mouse patterns, zooming in/out, and panning the field of view (FOV) to piece out morphological clues towards a diagnosis. Often, these studies juxtapose the navigation patterns of junior and senior-level pathologists. NaviPath [13], presents a human-AI collaborative navigation system designed to seamlessly integrate into pathologists’ workflows, considering the specific domain knowledge and navigation strategies required for effective examination of pathology scans.

**Multi-agent Systems.** The concept of multi-agent systems has gained traction in AI research, particularly for tasks requiring dynamic behavior and contextual understanding. Recent research has demonstrated the potential of large foundation models in creating interactive agent-based AI systems including interactions between robots, environments, and humans in the field of robotics [8, 15, 50]. These systems can perform complex tasks by leveraging the strengths of individual agents utilizing collaboration and coordination. The potential of multi-agent systems in handling real-world scenarios has been demonstrated in recent studies including but not limited to role-playing [29], reasoning [6], gaming [19] and software engineering [16]. In the medical domain some studies have explored role-playing providers (clinicians) treating patients and accumulating proficiency with increasing interactions [10, 30].

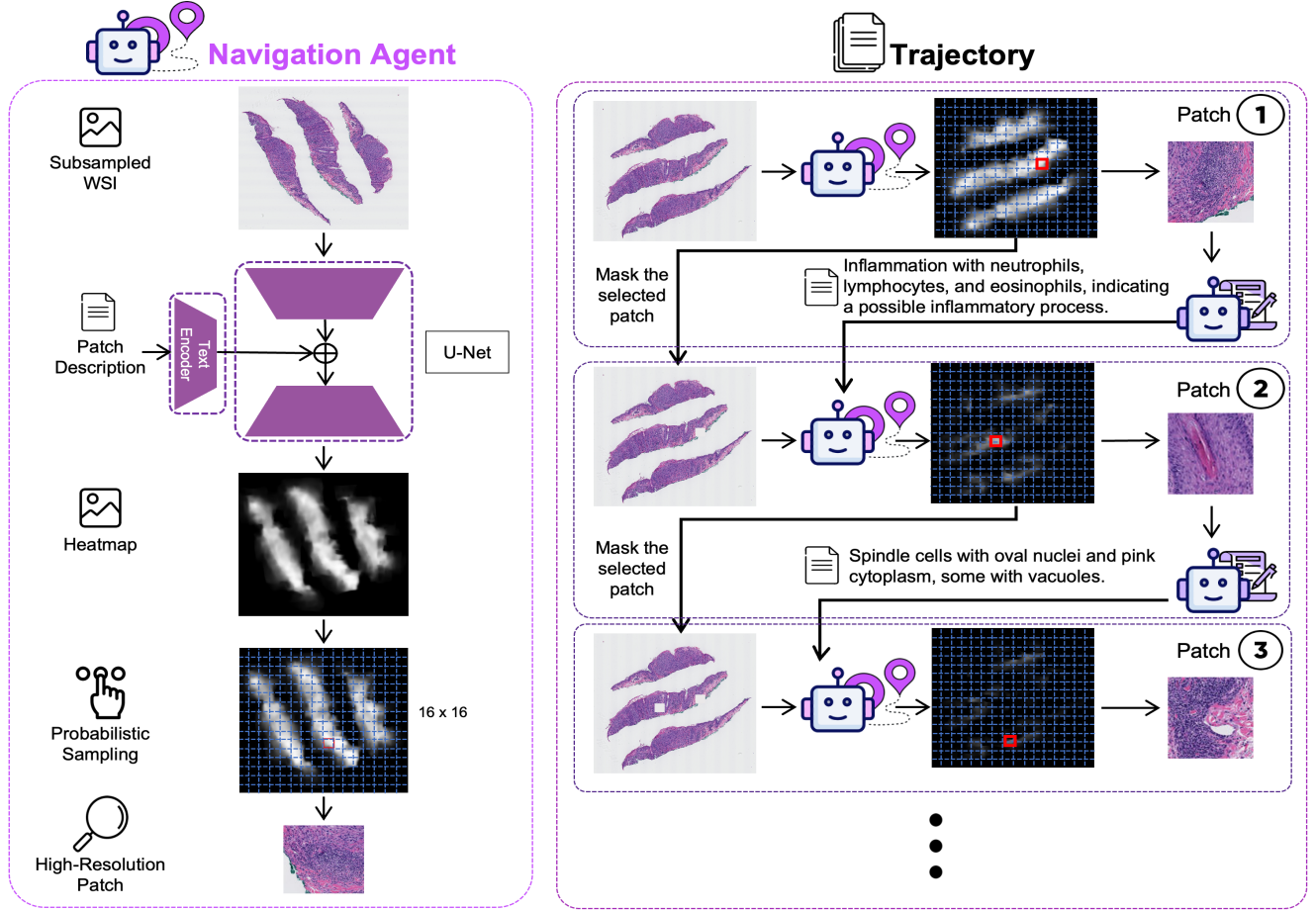


Figure 2. The left panel illustrates the Navigation Agent, as outlined in Section 4.2. The right panel presents the iterative trajectory generation process, which employs both the Navigation Agent and Description Agent, as described in Section 4.4.

These studies are centered around multiple providers; however, in medical image analysis, multi-agent systems can simulate the collaborative nature of different sub-tasks within the human-expert decision-making processes, including region navigation, understanding and holistic diagnosis.

### 3. Datasets

To lay the groundwork for describing our agents, we first start by introducing the different datasets used for training and evaluating our system.

**M-Path Skin Biopsy WSIs.** The skin biopsy WSIs in this dataset originate from M-Path study [2, 9, 38], consisting of 238 melanocytic lesion specimens stained with Hematoxylin and eosin (H&E). A consensus reference panel of three dermatopathologists, each with internationally recognized expertise, independently interpreted all 238 cases and established a consensus diagnosis for each case through a series of review meetings. There are 4 diagnostic classes

in this dataset: class 1 with 35 cases (mild and moderate dysplastic nevi); class 2 with 86 cases (severe dysplasia/melanoma in situ); class 3 with 70 cases (invasive melanoma stage pT1a); and class 4 with 47 cases (advanced invasive melanoma stage pT1b or more). For model development, the dataset is divided into training, validation, and test sets with a 168/35/35 case split, maintaining consistent class distribution across these sets.

**M-Path Pathologists' Viewport Data.** The M-Path study conducted viewport data collection, recruiting 87 pathologists from 10 U.S. states. Eligibility criteria included completion of residency and/or fellowship training and recent experience interpreting skin specimens in clinical practice. Pathologists' viewport data was gathered through an online digital slide viewer developed using Microsoft's open-source Silverlight-based HD View SL, a gigapixel image viewer. This viewer enabled pathologists to navigate each image by panning and zooming up to 60x magnification. During interpretation, the web-based viewer automatically

logged viewport tracking data, capturing a rectangular image area displayed on the pathologist’s screen at any given moment. For each interpretation (unique pathologist-case pair), the system recorded a list of viewport coordinates, magnification levels, and timestamps. Data from 32 pathologists who completed the M-Path study were included in the current study. Detailed methodology of the M-Path study is available in [38].

## 4. PathFinder

The multi-agent multi-modal framework proposed in this study includes four agents: 1) Triage Agent ; 2) Navigation Agent ; 3) Description Agent ; and 4) Diagnosis Agent. The details of training data and model architectures are described below. Figure 1 demonstrates how the four agents interact with each other towards the final goal which is diagnosing a WSI.

### 4.1. Triage Agent

The Triage Agent is an image-only transformer-based model tasked with separating class 1 (nevus/mild atypia and moderate atypia/dysplasia) from the rest in the M-Path dataset (Refer to section 3 for M-Path class definitions). We describe the data preparation, model architecture, and training details below.

**Data Generation.** Each whole slide image (WSI) is divided into non-overlapping  $512 \times 512$  patches at  $10\times$  magnification. Background patches (saturation less than 15) are discarded. If fewer than 150 patches remain, we randomly select additional patches from the WSI, apply the saturation filter again, and include the ones that pass. These additional patches may overlap with existing patches but ensure that each WSI contains sufficient information. All patches are then rearranged based on their spatial coordinates. The patches are embedded using the Quilt-Net image encoder [23], resulting in a feature vector of shape  $(N, 768)$  per WSI, where  $N$  is the total number of patches for the WSI.

**Model Architecture.** The Triage Agent includes several sequential stages (See Fig 1 in the Appendix): The feature vector is initially projected from  $(N, 768)$  to  $(N, dim)$  using a linear layer to align with the model’s embedding dimension  $dim$ . For compatibility with 2D processing, the vector is reshaped into a square grid through padding to dimensions  $H \times H$ , where  $H$  is the smallest integer satisfying  $H \times H \geq N$ , with padding achieved by repeating the first  $M = H^2 - N$  features. The padded vector is then processed through a transformer block, followed by positional encoding via the Pyramid Position Encoding Generator (PPEG) [45], and an additional transformer block, where each transformer block contains a single self-attention layer. Subsequently, multi-scale convolutional layers and a squeeze-and-excitation (SE) block [18] refine

the vector, capturing spatial patterns across scales and emphasizing key features. The output is then flattened and transformed back to the embedding dimension. A learnable class token is appended to capture global context, and the modified vector is passed through another transformer block, positional encoding, and a final transformer block. Finally, the class token is pooled and passed through an MLP head to produce the model’s output.

**Training Details.** We used binary cross-entropy loss for the classification task. The embedding dimension  $dim$  is set to 512. Training hyperparameters are as follows: a batch size of 1, learning rate of  $2 \times 10^{-4}$ , weight decay of  $1 \times 10^{-5}$ , and gradient accumulation over 32 steps. Training is conducted for up to 100 epochs, with early stopping after 30 epochs without improvement to prevent overfitting. Our approach achieved higher F1-score and accuracy compared to other methods that are directly comparable for this task (See Table 1 in the Appendix).

### 4.2. Navigation Agent

The Navigation Agent is designed to mimic a pathologist’s methodical approach to identifying regions of interest (ROIs) in whole slide images (WSIs). Unlike traditional systems that scan the entire WSI in a single, mechanistic sweep, our Navigation Agent adopts a more human-like, iterative process collaborating with the Description Agent. It begins by pinpointing an initial ROI, much as a pathologist would focus on one area at a time. This selected ROI is then relayed to the Description Agent, which provides a natural-language description of the area. Figure 2 illustrates the workflow of the Navigation Agent in the left panel.

In our initial attempt, we designed the Navigation Agent using a multi-modal architecture inspired by LLaVA [31], integrating an image encoder and a large language model (LLM). The image encoder extracted features from a low-resolution version of the WSI, and the LLM processed these features along with previous text descriptions to predict the next ROI. Specifically, the WSI was divided into a grid of patches, and the LLM would output the grid coordinates of the most relevant patch based on both visual and textual inputs. However, this approach faced significant challenges due to the limited size of our training dataset. The model tended to overfit, frequently selecting central patches regardless of the input (see Appendix 2 for details). This limitation prompted the exploration of more data-efficient methods that could better generalize from limited samples.

To overcome these challenges, we restructured the Navigation Agent to directly generate an importance map over the WSI, conditioned on textual descriptions from previous observations. This approach removes the dependency on the LLM for spatial selection and leverages a feedback mechanism between the image and text modalities. Let  $I^{(t)}$  be the input WSI at iteration  $t$ , with previ-



ously selected patches masked out to avoid re-sampling and  $D^{(1:t)} = \{D^{(1)}, D^{(2)}, \dots, D^{(t)}\}$  be the set of textual descriptions up to iteration  $t$ . At each iteration  $t$ , the Navigation Agent processes the masked WSI  $I^{(t)}$  to predict an importance map  $M^{(t)}$ , indicating the likelihood of each region being the next ROI. The importance map is conditioned on the aggregated textual information from previous descriptions. We define the importance map generation as  $M^{(t)} = f_{\text{Nav}}(I^{(t)}, E^{(t-1)})$  where,  $f_{\text{Nav}}$  is the Navigation Agent’s function (implemented as a lightweight U-Net [42]) that has four layers in both encoder and decoder and is conditioned with text embeddings of descriptions, as well as the masked version of the WSI that masks the earlier predicted ROIs, and  $E^{(t-1)}$  is the aggregated text embedding up to iteration  $t - 1$ .  $E^{(t-1)}$  is computed by encoding each description  $D^{(k)}$  using a pre-trained Text-to-Text-Transfer-Transformer (T5) text encoder [40] and averaging the embeddings:

$$E^{(t-1)} = \frac{1}{t-1} \sum_{k=1}^{t-1} \text{T5}_{\text{text}}(D^{(k)}) \quad (1)$$

At the first iteration ( $t = 1$ ), since there are no prior descriptions, the importance map is generated solely from the unmasked WSI  $M^{(1)} = f_{\text{Nav}}(I^{(1)})$ . From the importance map  $M^{(t)}$ , we then statistically sample the next patch to analyze. The probability  $p_{(i,j)}^{(t)}$  of selecting a location  $(i, j)$  is proportional to its importance score:

$$p_{(i,j)}^{(t)} = \frac{M_{(i,j)}^{(t)}}{\sum_{(i',j')} M_{(i',j')}^{(t)}} \quad (2)$$

We then sample the patch coordinates  $(i^*, j^*)$  based on this probability distribution:  $(i^*, j^*) \sim p_{(i,j)}^{(t)}$ . The selected high-resolution patch corresponding to  $(i^*, j^*)$  is sent to the Description Agent, which generates a new textual description  $D^{(t)}$ . The new description  $D^{(t)}$  is encoded and incorporated into the aggregated text embedding  $E^{(t)}$ :

$$E^{(t)} = \frac{1}{t} \sum_{k=1}^t \text{T5}_{\text{text}}(D^{(k)}) \quad (3)$$

This updated embedding  $E^{(t)}$  is then used to condition the Navigation Agent in the next iteration, enabling the model to refine its importance map  $M^{(t+1)}$  based on both the visual information from  $I^{(t+1)}$  and the accumulated textual insights. Therefore, we refer to it as the Text-conditioned Visual Navigator.

**Training details.** To train the Navigation Agent, we constructed a dataset from M-Path [38] consisting of WSIs and sequences of textual descriptions for the most important patches. Each training sample includes: The WSI and

the corresponding masked versions, the set of descriptions  $D^{(1:t)}$  for each iteration generated by Quilt-LLaVA [44] and the ground truth importance maps derived from pathologist annotations. We minimized the binary cross-entropy loss between the predicted importance maps  $M^{(t)}$  and the ground truth maps  $\hat{M}^{(t)}$ . Finally, to prevent overfitting, we paraphrased each description multiple times while preserving the semantic meaning, augmenting the textual data available for training.

### 4.3. Description Agent

We utilize Quilt-LLaVA [44], a multi-modal large language model capable of describing individual histopathology patches, as our Description Agent. While the original Quilt-LLaVA generates highly detailed findings, in this work, we instruction-tuned the model to produce more concise summaries, optimizing for computational efficiency. Using captions from the Quilt-1M dataset [21], we prompted GPT-4 to generate a list of findings as concise as possible. This process yielded 102,000 instruction-tuning samples. Figure 2 in the Appendix presents our prompt and a sample data entry generated for tuning the Description Agent. The Quilt-LLaVA 7B model was instruction-tuned for one epoch to obtain the Description Agent.

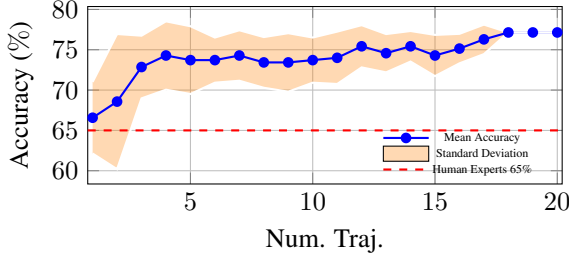
### 4.4. Diagnosis Agent

The Diagnosis Agent is a language-only model that analyzes all the gathered natural text descriptions produced by the Description Agent over all the patches identified by the Navigation Agent, to analyze natural text descriptions of histopathological findings and classify them into three categories (classes 2, 3, and 4).

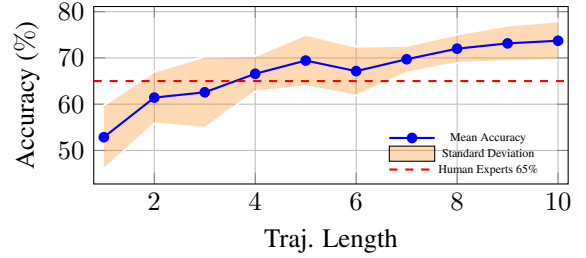
**Data Generation.** To train the Diagnosis Agent, we generated diagnostic trajectories—sequences of patch descriptions that simulated how a pathologist examined a whole slide image (WSI). Using our Navigation Agent, we proceeded as follows.

We first obtained a heatmap for a sub-sampled WSI ( $512 \times 512$  pixels) using a text-conditioned U-Net model, which highlighted regions of diagnostic significance. The WSI was divided into a  $16 \times 16$  grid, creating 256 patches of  $32 \times 32$  pixels each. Each patch received an importance score based on the mean intensity of the heatmap over the patch, indicating its diagnostic relevance. These scores were normalized across all patches.

To generate a single trajectory, we iterated the following steps ten times, yielding ten patches per case. At each iteration, a patch was selected using weighted probabilistic sampling based on the normalized importance scores, introducing variability and ensuring different patches were chosen across iterations. The selected patch was then cropped from the high-resolution  $10 \times$  WSI, and a description was generated by the Description Agent. Selected patches were



(a) Majority voting accuracy for 1-20 trajectories. Each trajectory contains 10 patches visited by the navigator.



(b) Majority voting accuracy for 5 trajectories with 1-10 patches per trajectory.

Figure 3. Ablation results. We ran 10 experiments, and plotted both the mean and standard deviation.

masked on the WSI to prevent reselection, and all descriptions generated thus far were combined into a single text input for the next iteration. The generation of a single trajectory is presented in the right panel of Figure 2.

For each WSI in the training and validation sets, we generated five ( $n = 5$ ) different trajectories, each containing ten patch descriptions, to capture various examination patterns. For the test set, we extracted additional trajectories ( $n = 20$ ) to assess the effect of trajectory number on diagnosis results. To introduce further variability, we used a LLaMA 3.1 Instruct model [7] at each iteration to rephrase the text descriptions. This approach effectively simulated the variability among pathologists, who might examine a single case using different patterns while seeking diagnostically relevant regions.

**Training Details.** The Diagnosis Agent consists of a large language model (LLM) with a classification head on top. The classification head maps the LLM’s output (vocabulary size) to the number of classes, producing the final classification probabilities using a single linear layer.

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:

*“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”,* 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  $5 \times 10^{-5}$ , weight decay 0.001, and batch size 16. Due to resource constraints and the limited size of the dataset, we selected GPT-2 [39] as the base pre-trained LLM.

## 5. Experiments and Results

This section outlines the experimental setup and evaluates the performance of the proposed PathFinder framework. First, we conduct a qualitative assessment of the descriptions generated by the Description Agent, comparing them to two vision-language models (VLMs). Next, we evaluate PathFinder on the M-Path dataset for melanoma diagnosis (see Section 3), benchmarking it against state-of-the-art transformer-based and MIL-based baselines, as well as public and private large language models (LLMs) using prompting without additional training. Finally, we analyze PathFinder’s performance under various configurations, altering the Triage, Navigation, and Description Agents. Detailed evaluations are provided in the following subsections.

### 5.1. Pathologist Evaluation of Description Quality

To assess the quality of descriptions generated by our Description Agent, we conducted a survey in which two expert pathologists rated descriptions produced by our Description agent in comparison to those generated by GPT4-o [20] and LLaVA-Med [28]. We selected 25 cases from the M-Path dataset, sampling across the four diagnostic classes. For each case, we cropped the consensus region of interest, manually labeled by a panel of expert dermatopathologists as the area most representative of the diagnosis. Using this region, we prompted our Description Agent, LLaVA-Med and GPT4-o to generate concise descriptions of each histopathology patch. These descriptions were then presented to two expert pathologists in a randomized, double-blind format. Each pathologist was asked to respond to two questions for each case to indicate their preferred description and the reason for their preference. Refer to Appendix 4.1 for a detailed explanation of this assessment. The results shown in Figure 4 indicate that, PathFinder’s Description agent achieves comparable performance to GPT-4o while being significantly more cost-effective, operating with just 7B parameters - a fraction of GPT-4o’s size.

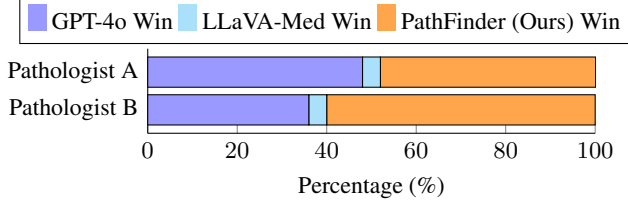


Figure 4. Expert human pathologist preferences for each model in assessing description quality, evaluated in a double-blind survey for unbiased comparison.

## 5.2. PathFinder Evaluation

For evaluating PathFinder, we utilize the M-Path dataset which contains histopathology WSIs of melanocytic skin tissue. As outlined in 4.4, multiple trajectories are generated per case to simulate the variability in diagnostic patterns observed among pathologists, who may assess a single case with diverse visual strategies to identify diagnostically significant regions. To mitigate randomness in our results, we evaluated PathFinder 10 times on the test set, each time using a different random subset of 5 trajectories selected from the total of 20. For each Whole Slide Image (WSI), majority voting is performed on the predictions from the 5 selected trajectories to produce the final result. The overall performance is then reported in Table 1 as the mean of the results across the 10 runs. We balanced the testing dataset to ensure that each diagnostic class is represented by an equal number of samples. Consequently, the micro-averaged F1 score, precision, and recall are equivalent to the accuracy reported in the table. We opted to use micro-averaged metrics in our clinical evaluation, because they appropriately balance the importance of different stages of skin cancer, which is crucial for assessing the overall reliability and effectiveness of the diagnostic tool.

We compared Pathfinder to four state-of-the-art baseline models: 1) three transformer-based models all utilizing the ScAtNet architecture [12, 32, 52] and 2) four MIL-based model, using ABMIL [24] with different backbones. ScAtNet utilizes a MobileNetV2 backbone [43] to extract multi-scale features from images at 7.5x, 10x, and 12.5x magnification. For the first baseline model, these feature vectors are subsequently fed into ScATNet which aggregates information of the three scales to perform the diagnostic task using Transformer blocks. The second approach [12] augments the WSI with ROI heatmaps generated by the U-Net model, appending these maps as a fourth input channel and using ScAtNet for classification. The third baseline model, SAG [32], converts diagnostically relevant entities into attention signals, integrating these with ScAtNet and employing an attention-guiding loss function that combines heuristic guidance (HG) and tissue guidance (TG) based on disease-specific prior knowledge such as tissue, structure,

and cellular information. In addition to the original ABMIL model [24], we extended our evaluation by incorporating three additional ABMIL variants, each using a different pathology-specific foundation model as a backbone: CONCH [33], UNI2-h [4], and QuiltNet [23]. ABMIL aggregates information across instances using an attention mechanism that assigns weights to each instance, allowing the model to capture its contribution to the final bag label in a permutation-invariant manner.

Then, we conducted comprehensive experiments to evaluate PathFinder by examining different architectures for each agent component, achieving 74% accuracy that surpasses both human experts (65%) and previous state of the art (66% best). Our evaluation focused on three main aspects:

**Navigator Architectures.** First, to quantify the importance of Description Agent feedback, we tested a Visual-Only Navigator that employs weighted probabilistic sampling for patch selection without iterative feedback in a single pass over the WSI. Additionally, we implemented Imitated Sampling, which leverages pathologists’ viewing pattern distributions (viewport width, height, and zoom level) from our M-Path dataset (Section 3) to statistically sample patches as important WSI regions. If pathologists spend more time focusing on a region, we gave a higher chance of sampling to that region. Both Imitated Sampling and Vision-Only Navigator performed similarly (64% and 63% respectively), indicating that both pure statistical and learned “sampling”, regardless of source, has limited effectiveness. To further assess the necessity of our iterative navigation approach, we added a non-iterative baseline, selecting the top 10 patches from ABMIL attention scores using three different pathology-specific foundation model backbones. The best model with CONCH[33] backbone performed substantially lower than our best navigation-based approach (74% vs 54%). This confirms that a purely image-based, non-iterative selection approach is insufficient, as it lacks the ability to iteratively refine patch selection based on evolving textual descriptions. Furthermore, we evaluated text-conditioned visual navigators using either CLIP-based or T5-based text encoders. The T5-based navigator significantly outperformed its CLIP-based counterpart (74% vs 62%), suggesting CLIP’s 77-token limit constrains its ability to effectively process multiple descriptions (we simply truncate descriptions exceeding 77 tokens, then average if a description is long). Finally, our navigation-based approach (74%) outperformed exhaustive search (68%), which utilizes all non-background patches of the WSI, suggesting that selective patch sampling helps avoid confusion from irrelevant regions.

**Description Agents.** We compared a fine-tuned version of Quilt-LLaVA (optimized for concise descriptions) against off-the-shelf LLaVA-Med. The fine-tuned version showed

Methods	Accuracy	F-1 score
<i>Baselines</i>		
Human Experts [9]	0.65	0.65
ScAtNet [52]	0.62	0.62
ScAtNet + ROI Heatmap [12]	0.63	0.63
ScAtNet + SAG [32]	0.60	0.60
ABMIL [24]*	0.46	0.47
ABMIL w/ CONCH [33]*	0.66	0.60
ABMIL w/ UNI2-h [4]*	0.66	0.66
ABMIL w/ QuiltNet [23]*	0.61	0.63
<i>LLM Prompting Baselines</i>		
BioMistral-7B	0.43	0.43
Mistral-Nemo-Instruct-2407	0.41	0.41
GPT-4o	0.49	0.49
Meta-Llama-3-8B-Instruct	0.31	0.31
LLaVA-Med-v1.5-Mistral-7b	0.43	0.43
Quilt-LLaVA-v1.5-7b	0.29	0.29
<i>Ours</i>		
PathFinder + ABMIL w/ UNI2-h Attention-Based Top Patches	0.46	0.46
PathFinder + ABMIL w/ QuiltNet Attention-Based Top Patches	0.46	0.46
PathFinder + ABMIL w/ CONCH Attention-Based Top Patches	0.54	0.54
PathFinder + T5-Based Text-Conditioned Visual Navigator + No Triage Agent	0.58	0.58
PathFinder + T5-Based Text-Conditioned Visual Navigator + LLaVA-Med Descriptions	0.56	0.56
PathFinder + CLIP-Based Text-Conditioned Visual Navigator + LLaVA-Med Descriptions	0.60	0.60
PathFinder + Imitated Sampling	0.63	0.63
PathFinder + Vision-Only Navigator	0.64	0.64
PathFinder + CLIP-Based Text-Conditioned Visual Navigator	0.62	0.62
PathFinder + Exhaustive search	0.68	0.68
<b>PathFinder + T5-Based Text-Conditioned Visual Navigator</b>	<b>0.74</b>	<b>0.74</b>

\* ABMIL result is based on a single run and does not use majority voting

Table 1. Majority voting performance for whole slide image (WSI) diagnosis on the M-Path dataset. Accuracy is reported, and the F-1 score is identical due to the balanced testing set. Finally, coverage here is the percent of patches used across all trajectories.

superior performance (74% vs 56%), demonstrating better guidance for the navigator. Notably, when paired with LLaVA-Med descriptions, the T5-based Navigator showed no advantage over the CLIP-based version (56% vs 60%). This suggests that a more powerful text encoder like T5 can actually be detrimental when processing lower-quality descriptions, potentially steering the Navigator toward irrelevant regions. This finding emphasizes the importance of high-quality descriptions for effective navigation.

**Diagnosis Agents.** We evaluated various public and private LLMs as baselines for our Diagnosis Agent (detailed in Table 1 under LLM Prompting Baselines). Specifically, we used PathFinder with T5-Based Text-Conditioned Visual Navigator and Quilt-LLaVA Description Agent to generate multiple descriptions for each WSI and prompted LLMs to make the classification given the descriptions. To view our prompt, please see Section 4.2 in Appendix.

Finally, it is worth noting that without the Triage Agent, the performance of the best Pathfinder-variant dropped be-

low baselines, likely due to Quilt-LLaVA’s train dataset’s bias toward malignant cases.

The evaluation of the baseline models are similarly done using the majority voting over 10 runs. PathFinder achieves 8% improvement compared to the best baseline approaches, ABMIL with CONCH [33] and UNI2-h [4] backbones. Considering that GPT-2 is a relatively small LLM compared to the current state-of-the-art, we believe that utilizing larger LLMs could further improve diagnostic outcomes.

Lastly, to investigate the impact of the number of trajectories on model performance, we evaluated the model using between 1 and 20 trajectories for majority voting, as well as the effect of varying trajectory lengths. Figure 3 illustrates this analysis, indicating optimal performance with 5 trajectories and 10 patches per trajectory. We run every experiment for 10 rounds and report the mean and standard deviation.



## 6. Discussion

This study presents PathFinder, a multi-modal, multi-agent AI framework designed to emulate the multi-scale, iterative diagnostic approach of expert pathologists for histopathology whole slide images (WSIs). By integrating Triage, Navigation, Description, and Diagnosis Agents, PathFinder collaboratively gathers evidence to deliver accurate, interpretable diagnoses with natural language explanations. Notably, it surpasses state-of-the-art methods and the average performance of human experts in melanoma diagnosis, setting a new benchmark in AI-driven pathology.

PathFinder has the potential to accelerate diagnostic workflows, reducing the reliance on manual examination and enabling timely patient care in clinical settings. Its natural language descriptions provide interpretability, facilitating the validation of AI-generated diagnoses by pathologists. Moreover, its integration of vision-language models (VLMs) and large language models (LLMs) highlights the promise of multi-modal AI in delivering scalable, specialized diagnostic tools that could improve access to pathology expertise.

**Limitations.** Despite its strengths, PathFinder has limitations. The framework relies on pre-existing datasets and significant computational resources, posing challenges in resource-constrained environments. Additionally, the complexity of the Navigation Agent’s decision-making process and occasional hallucinations by the Description Agent could affect transparency and accuracy of the decision-making process. Future work should address these issues by enhancing dataset diversity, computational efficiency, and patch selection strategies, further advancing PathFinder’s potential as a transformative tool in AI-assisted pathology.

**Acknowledgements.** We thank Dr. Oliver Chang and Dr. Kristin Shaw for their participation in our clinical evaluation. We also acknowledge Microsoft for providing OpenAI credits, Department of Defense (W81XWH-20-1-0798), National Cancer Institute (U01CA231782, and R01CA200690), and partial funding through a Population Health Initiative at University of Washington.

## References

- [1] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024. 2
- [2] Patricia A Carney, Lisa M Reisch, Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan Knezevich, Berta M Geller, Gary Longton, and Joann G Elmore. Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified delphi method. *Journal of cutaneous pathology*, 43(10):830–837, 2016. 3
- [3] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2
- [4] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 2, 7, 8
- [5] Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Bin Zhang, Nana Pei, Rongshan Yu, Yu Qiao, et al. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. *arXiv preprint arXiv:2410.11761*, 2024. 2
- [6] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. 2
- [7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6
- [8] Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, et al. An interactive agent foundation model. *arXiv preprint arXiv:2402.05929*, 2024. 2
- [9] Joann G Elmore, Raymond L Barnhill, David E Elder, Gary M Longton, Margaret S Pepe, Lisa M Reisch, Patricia A Carney, Linda J Titus, Heidi D Nelson, Tracy Onega, et al. Pathologists’ diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *bmj*, 357, 2017. 2, 3, 8
- [10] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*, 2024. 2
- [11] Fatemeh Ghezloo, Pin-Chieh Wang, Kathleen F Kerr, Tad T Brunyé, Trafton Drew, Oliver H Chang, Lisa M Reisch, Linda G Shapiro, and Joann G Elmore. An analysis of pathologists’ viewing processes as they diagnose whole slide digital images. *Journal of Pathology Informatics*, 13:100104, 2022. 2
- [12] Fatemeh Ghezloo, Oliver H. Chang, Stevan R. Knezevich, Kristin C. Shaw, Kia Gianni Thigpen, Lisa M. Reisch, Linda G. Shapiro, and Joann G. Elmore. Robust roi detection in whole slide images guided by pathologists’ viewing patterns. *Journal of Imaging Informatics in Medicine*, 2024. 7, 8
- [13] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang’Anthony’ Chen. Augmenting pathologists with navipath: design and evaluation of

- a human-ai collaborative navigation system. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023. 2
- [14] Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 755–764. Springer, 2023. 2
- [15] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhao Xu, and Chaoyang He. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024. 2
- [16] Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Vision and the road ahead. *arXiv preprint arXiv:2404.04834*, 2024. 2
- [17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 6
- [18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. 2018. 4
- [19] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms’ gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024. 2
- [20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6
- [21] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024. 5
- [22] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, and Linda Shapiro. Multi-modal masked autoencoders learn compositional histopathological representations. *arXiv preprint arXiv:2209.01534*, 2022. 2
- [23] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023. 2, 4, 7, 8
- [24] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 7, 8
- [25] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018. 1
- [26] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 2
- [27] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, 2021. 1
- [28] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 6
- [29] G Li, HAAK Hammoud, H Itani, D Khizbullin, and B Ghanem. Camel: communicative agents for” mind” exploration of large scale language model society.(2023). *arXiv preprint arXiv:2303.17760*, 2023. 2
- [30] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yungwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024. 2
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 4, 1
- [32] Kechun Liu, Wenjun Wu, Joann G Elmore, and Linda G Shapiro. Semantics-aware attention guidance for diagnosing whole slide images. In *MICCAI*, 2024. 2, 7, 8
- [33] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 7, 8
- [34] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahromg Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, pages 1–3, 2024. 2
- [35] Sachin Mehta, Ximing Lu, Wenjun Wu, Donald Weaver, Hannaneh Hajishirzi, Joann G Elmore, and Linda G Shapiro. End-to-end diagnosis of breast biopsy images with transformers. *Medical image analysis*, 79:102466, 2022. 2
- [36] Ezgi Mercan, Linda G Shapiro, Tad T Brunyé, Donald L Weaver, and Joann G Elmore. Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers. *Journal of digital imaging*, 31:32–41, 2018. 2
- [37] Jesper Molin, Morten Fjeld, Claudia Mello-Thoms, and Claes Lundström. Slide navigation patterns among pathologists with long experience of digital review. *Histopathology*, 67(2):185–192, 2015. 2
- [38] Tracy Onga, Raymond L Barnhill, Michael W Piepkorn, Gary M Longton, David E Elder, Martin A Weinstock, Steven R Knezevich, Lisa M Reisch, Patricia A Carney, Heidi D Nelson, et al. Accuracy of digital pathologic analysis vs traditional microscopy in the interpretation of melanocytic lesions. *JAMA dermatology*, 154(10):1159–1166, 2018. 3, 4, 5
- [39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 6
- [40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a

- unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. [5](#)
- [41] Lucia Roa-Peña, Francisco Gómez, and Eduardo Romero. An experimental study of pathologist’s navigation patterns in virtual microscopy. *Diagnostic pathology*, 5:1–11, 2010. [2](#)
- [42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18, pages 234–241. Springer, 2015. [5](#), [1](#)
- [43] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. [7](#)
- [44] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024. [2](#), [5](#)
- [45] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. [2](#), [4](#), [1](#)
- [46] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*, 2024. [2](#)
- [47] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019. [2](#)
- [48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. [1](#)
- [49] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, pages 1–12, 2024. [2](#)
- [50] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. 2024. [2](#)
- [51] Wenjun Wu, Sachin Mehta, Shima Nofallah, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:163526–163541, 2021. [2](#)
- [52] Wenjun Wu, Sachin Mehta, Shima Nofallah, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:163526–163541, 2021. [2](#), [7](#), [8](#)
- [53] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024. [2](#)
- [54] Zhaochang Yang, Ting Wei, Ying Liang, Xin Yuan, Ruitian Gao, Yujia Xia, Jie Zhou, Yue Zhang, and Zhangsheng Yu. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *bioRxiv*, pages 2024–05, 2024. [2](#)
- [55] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022. [2](#)
- [56] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. *arXiv preprint arXiv:2403.08967*, 2024. [2](#)



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

## Supplementary Material

### 1. Triage agent

Figure 1 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 1, 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
<b>Triage Agent</b>	<b>0.57</b>	<b>0.95</b>	<b>0.91</b>

Table 1. Comparison of Triage Agent with benchmark methods on Class 1 vs. Non-Class 1 classification

### 2. 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

#### 2.1. 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.

#### 2.2. 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

#### 2.3. 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).

### 3. 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 2 illustrates the prompt used and a sample of the data generated for fine-tuning the Description Agent.

### 4. Evaluation and experiments

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

#### 4.1. 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



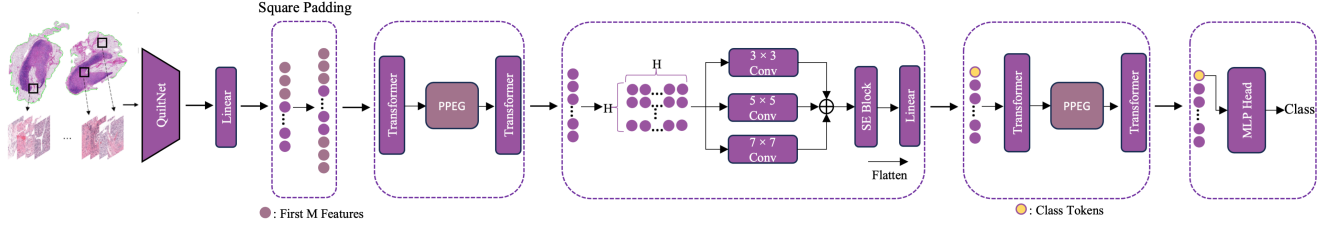


Figure 1. Overview of the Triage Agent architecture. Definitions of  $M$  and  $H$  can be found in Section 4.

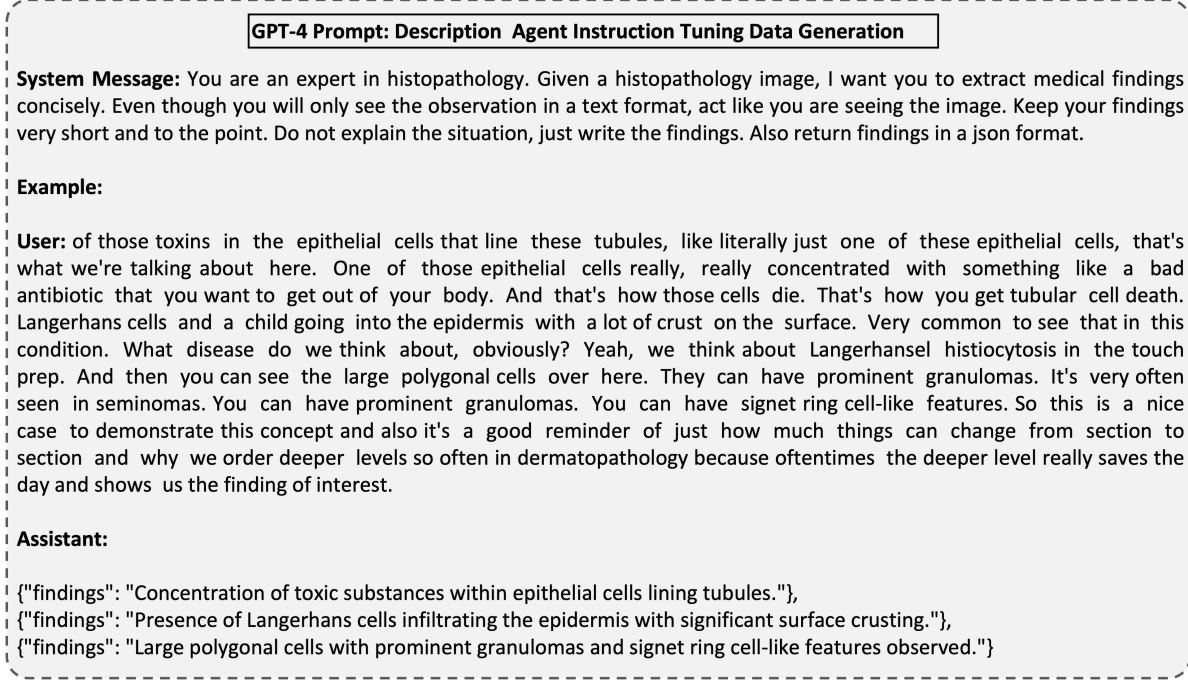


Figure 2. GPT-4 prompt to generate instruction-tuning dataset for the Description Agent.

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. 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, randomized 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 3 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, inher-

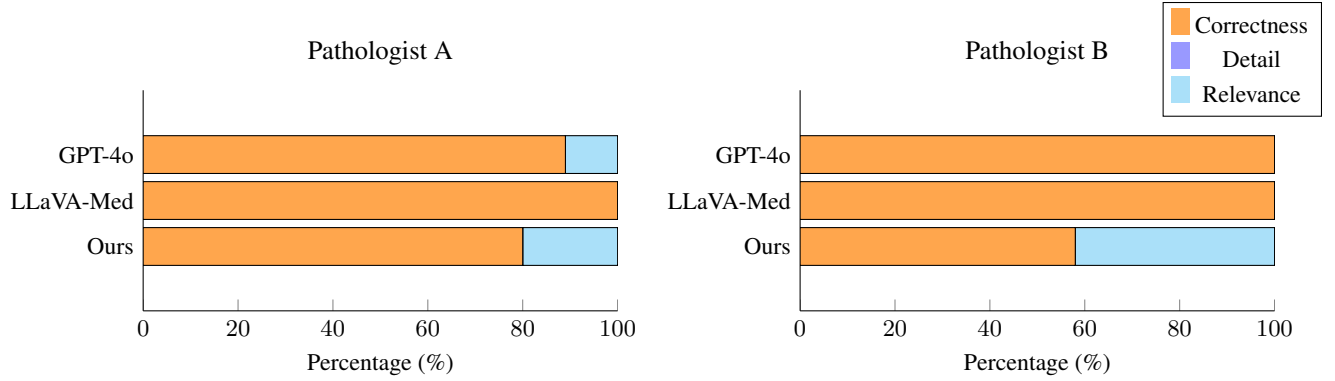


Figure 3. Expert human pathologist preferences for each model, segmented by the reasons for their choices. Each subplot corresponds to one pathologist and shows their ratings for PathFinder (Ours), LLaVA-Med, and GPT-4o.

ently limiting detailed information. The majority of preferences were based on the correctness of the descriptions.

#### 4.2. 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.