PathInteract: A Video Analysis Suite for Capturing and Analyzing Pathology Slide Reviewing Interactions

[1]Jun Jiang Ph.D., [2]Qiangqiang Gu Ph.D., [3]Xin Zhou MD. Ph.D., [4]Ruifeng Guo MD. Ph.D., [1]Sunyang Fu Ph.D., [1]Andrew Wen M.S., [1]Liwei Wang MD. Ph.D., [2]Nianyi Li MD. Ph.D., [1]Qiuhao Lu Ph.D., [5]Rongzhen Zhang M.D.,Ph.D., [5]Alexexander Banerjee DO, [6]Peiliang Lou Ph.D., [6]Chen Wang Ph.D., [2]Yanshan Wang Ph.D., [1]Hongfang Liu Ph.D.

[1]Department of Health Data Science and Artificial Intelligence, McWilliam School of Biomedical Informatics, UTHealth Houston, TX
[2]Computational Pathology & AI Center of Excellence, Department of Pathology, University of Pittsburgh School of Medicine, Pittsburgh, PA.
[3]Department of Pathology, University of Colorado, Denver.
[4]Department of Laboratory Medicine and Pathology, Mayo Clinic, FL
[5]Department of Pathology and Laboratory Medicine, McGovern Medical School, UTHealth Houston, TX
[6]Department of AI and Informatics Research, Mayo Clinic, Rochester, Minnesota

Corresponding author:
Hongfang Liu, PhD
Postal address: 7000 Fannin Street #Suite 600, Houston, TX 77030
E-mail: hongfang.liu@uth.tmc.edu
Telephone: 713-500-3900

Abstract:
Digital pathology has transformed how pathologists review and interpret tissue specimens, enabling remote access, efficient storage, and advanced visualization. Systematically analyzing pathologists' slide reviewing interactions can uncover opportunities to design AI-assisted tools that integrate seamlessly into routine practice. We present an analysis suite, PathInteract, which extracts experts' interactions with pathology slides based on video recording. Mouse cursor movements, viewport actions (i.e., zooming and panning), and verbal narratives are extracted using multiple deep learning and computer vision moduals. Unlike prior methods requiring specialized software or equipment, our approach operates on screen recordings, enabling broader applicability. We developed PathInteract using QuPath-recorded diagnostic sessions with view-tracking logs and applied it to ten educational YouTube videos. Results revealed distinct viewing patterns based on pathologist experience, tissue type, and use context. Cursor tracking and viewport detection achieved strong agreement with ground truth, while speech analysis highlighted differences in cell-level focus across diseases. PathInteract enables scalable analysis of reviewing interactions in digital pathology through video recording and supports repurposing of existing pathology videos towards building interpretable, multimodal pathology AI datasets.

Key words: slide reviewing interactions, pathology diagnostic pattern, view action tracking

Introduction

Histopathological diagnosis is an inherently complex process that requires pathologists to visually inspect all the slide images, integrate subtle tissue patterns, and apply domain expertise and cognitive reasoning to arrive at a final interpretation[1, 2]. Studying these diagnostic dynamics can yield valuable insights into decision making, with the potential to improve diagnostic accuracy, consistency, and training [3, 4]. The widespread adoption of digital pathology has made it possible to systematically capture how pathologists navigate slides, opening new opportunities for quantitatively analysis of diagnostic practice. Recent studies have demonstrated the utility of analyzing viewing behavior data for developing automated feedback systems and decision-support tools to support both training and clinical practice [5].

Despite these advances, effectively capturing and analyzing how pathologists review whole slide images (WSIs) remains challenging. Current approaches often rely on workstation-installed software to track user interactions. For instance, HD View SL, a customized tool derived from Microsoft's open-source Silverlight gigapixel image viewer, has been used to monitor viewport movements, cursor activity, and keystrokes, providing detailed behavioral insights[1, 6-8]. However, these solutions face significant implementation barriers: they require administrative privileges for installation, generate proprietary data, and are often inaccessible outside controlled environments. Such limitation hinders broader adoption in clinical practice, particularly in healthcare enviornments, usually with strict IT governance. Eye-tracking has also been employed to capture pathologists' reviewing behavior [9, 10], but its reliance on specialized hardware and complex data analysis pipelines makes it difficult to generalize and scale.

An alternative and scalable strategy is to analyze video recordings of digital pathology practice. These recordings may originate from routine diagnostic reviews, care discussions, or educational sessions, and can capture rich behanioral signals such as viewport navigation, cursor trajectories, spoken commentary, and visual focus at the frame level. While plenty studies exist on the technical modules for similar tasks, few studies have systematically evaluated slide-reviewing patterns in day-to-day clinical practice. This gap presents an opportunity to establish a practical, video-based approach for analyzing diagnostic interactions, thereby advancing our understanding of pathologist behavior and supporting the development of AI-assisted tools that integrate seamlessly into real-world workflows.

During digital pathology slide review, multiple information streams can be extracted from videos to characterize diagnostic behavior. 1) *mouse cursor* movements often reflect pointing behavior and can indicate the pathologist's focus on specific regions or even delineation of diagnostic areas [11]. However, detecting the cursor in pathology videos is technically challenging due to its small size and the heterogeneous, often complex, slide background. Traditional computer vision techniques such as template matching [12] are limited in effectiveness because cursor appearance may vary across videos. Deep learning-based object detection models[13, 14] offer greater adaptability, but their application in this

specific context remains largely unexplored. 2) *Viewport actions*, primarily zomming and panning, reveal how pathologists navigate digital slides, providing insights into areas of interest, diagnostic focus, and workflow patterns[15]. These actions are typically inferred through object tracking methods such as optical flow [16], which estimate motion by comparing the positions of key points across successive video frames. However, the accuracy of such methods has not been rigorously evaluated in the context of pathology video analysis. 3) *Audio* recordings captured during review sessions often include verbal commentary or think-aloud protocols that reflect diagnostic reasoning and inspection strategies[17]. Speech recognition tools can transcribe these narratives into text, enabling the use of large language models (LLMs) to extract clinically relevant information. 4) Individual video *frames* themselves can be leveraged to study visual sequences and highlight key diagnostic transitions [18]. When combined with text information, either extracted from accompanying audio or sourced from external platforms such as YouTube, these image-text pairs can serve as training data for multimodal deep learning models [19, 20].

While these diverse information streams provide novel opportunities to investigate cognitive and behavioral patterns in diagnostic practice, studies that systemsally extract and analyze reviewing behaviros from pathology videos remain scarce. Existing efforts have typically focused on isolated components, often failing to capture the entire diagnostic workflow or to validate video-based extraction methods comprehensively. Consequently, the reliability of video-based approaches for accurately characterizing diagnostic behavior remains uncertain.

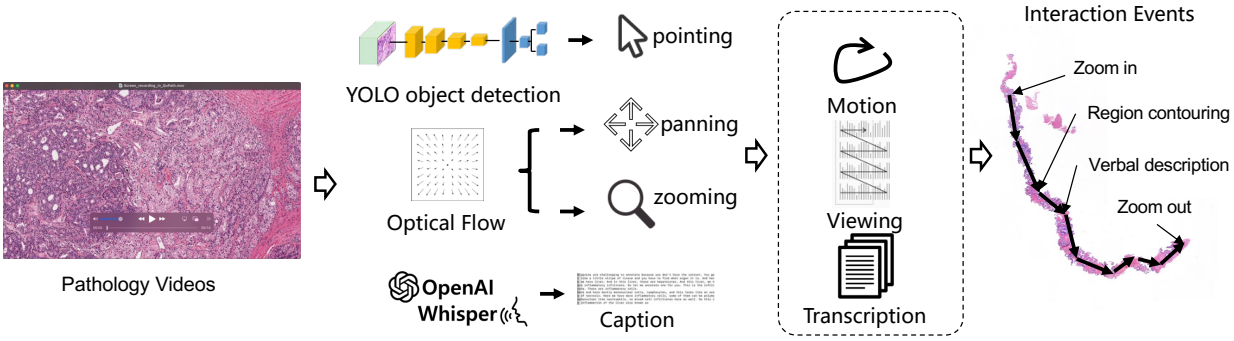


Figure 1. Overview of PathInteract. Interaction events, such as mouse cursor motions, viewport actions and audio descriptions were extracted from pathology slide review recordings.

To address these challenges, we introduce an integrative analysis suite, PathInteract, designed to capture pathologists' dynamic interactions during the slide review process. PathInteract combines computer vision, speech recognition, and deep learning models to extract and analyze key inspection patterns, including cursor movements, viewport actions, and audio commentary. (**Figure 1**). Our main contributions are as follows:

1) Integrated video-based suite: we developed a unified system that extracts diagnostic behavior patterns directly from pathology videos, eliminating the need for specialized software installation on workstations.
2) Systematic evaluation: we established a robust evaluation strategy for the suite using QuPath's view-tracking functions as ground truth, enabling rigorous assessment of the accuracy of video-based behavior capture.
3) Behavioral insights: we conducted downstream analyses across multiple diseases and pathologists using the fraemwork, uncovering patterns of diagnostic behavior and providing new insights into interobserver variability and decision-making dynamics.
4) AI utility: we demonstrated that PathInteract can extract clinically meaningful information, such as regional annotations and descriptive commentary, that can be repurposed for training and refining AI models.

Datasets

We created two primary datasets to support different components of our study. The first dataset consists of 20 QuPath view tracking from two pathologists and was designed to evaluate the accuracy of individual modules within PathInteract. The second dataset consists of 10 pathology-related videos downloaded from YouTube, which was used for downstream analysis to demonstrate the effectiveness of PathInteract. In addition, independent task-specific datasets were constructed for modules such as cursor detection to train and evaluate corresponding deep learning models. Further details regarding each dataset are provided in the relevant methods and results sections.

QuPath view-tracking datasets

Using the view-tracking feature in QuPath, we generated a collection of videos paired with detailed tracking records. Two pathologists were instructed to replicate their routine diagnostic behavior by reviewing 10 whole-slide images in QuPath, simulating real-world practice. The images included five bulk tissue samples from the Cancer Genome Atlas Program (TCGA)[21] and five prostate needle biopsies from the Prostate cANcer graDe Assessment (PANDA) grand challenge dataset [22]. With the view-tracking function enabled, QuPath periodically recorded zoom levels, viewport locations, and cursor positions. These actions were then replayed within QuPath, and screen recordings were captured to serve as ground truth for evaluating the accuracy of our methods.

Online pathology videos

To evaluate the overall utility of PathInteract, we curated 10 publicaly available videos (720p resolution) from two pathology-related YouTube channels. Five of these videos were dermatology residency training recordings, while the remaining five featured hematology case discussions. All videos were used in compliance with YouTube's Terms of Service and under fair use for academic research (see disclosure). PathInteract was applied to these videos, and statistical analyses were performed on the extracted pathology slide-view patterns to examine diagnostic dynamics across different scenarios.

Methods

Mouse cursor tracking

The mouse cursor tracking module detects mouse curors within the video frames. We employed YOLOv8[23], an efficient object detection framework to achieve this goal. To train this model, a synthetic dataset was generated from mouse cursor images and pathology images. Within this dataset, each sample was created by overlaying a mouse cursor image to a tissue slide image at a random location. The location of the cursor was saved as the ground truth for YOLOv8 training. To improve the model's generalizability, the pathology images, used as the background for cursor detection, were randomly sampled from different regions of multiple WSIs at various resolutions. Additionally, a diverse set of mouse cursor images was collected to account for the variations in cursor appearance across different videos. The generated dataset was formatted to match the input requirements of Ultralytics YOLO[24], and split into training, validation and testing dataset for YOLOv8 training and evaluation. Precision, Recall, F1 score and $mAP_{0.75}$ were calculated as the quantitative evaluation metrics.

During the referencing phase, the cursor tracking results were also compared with the cursor locations recorded in the QuPath view tracking logs. To evaluate the tracking accuracy in QuPath screen recordings, we set the video resolution to 720p, as the resolution (720p, 360p, 4k) influences the detected cursor location. Similarly, we fixed the video frame rate (fps) at 20, as the frame rate affects the number of tracked instances. Under these settings, we used the Fréchet Distance [25] as the evaluation metric to measure the consistency between our detection results and the QuPath view tracking.

Viewport zooming and panning

To detect the zooming and panning actions within online videos, optical flow [26] was used. This method automatically captures key points within current (t=0) and next (t=1) video frames. By calculating the relative locations of the corresponding key points, the zooming and panning signals can be detected. To exclude perturbations within zooming and panning detections, threshold values were set respectively (zooming threshold =1±0.002, panning threshold=±1). The detection mechanism can be summarized as formulars as below:

$$Z = \frac{\frac{1}{N}\sum_{i=1}^{N}|P_{t1}{}^{i} - \overline{P_{t1}}|}{\frac{1}{N}\sum_{i=1}^{N}|P_{t0}{}^{i} - \overline{P_{t0}}|}, if \begin{cases} Z > 1.002, & Zoom\ out \\ Z < 0.998, & Zoom\ in \end{cases} \quad (1)$$

$$P(\Delta x, \Delta y) = \frac{1}{N}\sum_{i=1}^{N}(P_{t0} - P_{t1}), if \begin{cases} \Delta x > 1, & Pan\ left \\ \Delta x < -1, & Pan\ right \\ \Delta y > 1, & Pan\ up \\ \Delta y < -1, & Pan\ down \end{cases} \quad (2)$$

In which the P_t denotes the key points, N denotes the number of key points, x and y denotes the coordinates of the key points. t0 and t1 denote the timestamp of the data frame.

The viewport action tracking module was evaluated based on the view-tracking logs from QuPath. We used the gradient of zoom factors (Z) and view port locations (x, y) at each timestamp to determine the zooming and panning actions during the slides reviewing. The columns named "Downsample factor" (denotes zoom factors), "X" and "Y" (denote viewport locations) within the view tracking logs were processed according to the following formular:

$$\Delta Z = Z_{t0} - Z_{t1}, if \begin{cases} \Delta Z > 0.01, & Zoom\ out \\ \Delta Z < -0.01, & Zoom\ in \end{cases} \quad (3)$$

$$\Delta x = x_{t0} - x_{t1}, if \begin{cases} \Delta x > 0.01, & Pan\ left \\ \Delta x < -0.01, & Pan\ right \end{cases} \quad (4)$$

$$\Delta y = y_{t0} - y_{t1}, if \begin{cases} \Delta y > 0.01, & Pan\ up \\ \Delta y < -0.01, & Pan\ down \end{cases} \quad (5)$$

Speech to text

The speech to text module in PathInteract aims to convert pathologists' audios to text for further downstream analysis. While speech recognition technology has been thoroughly researched with numerous mature solutions available for general applications, the accuracy of these systems in specialized medical contexts, particularly for pathology-specific terminology, has received limited investigation. Given that this module functions largely independently within PathInteract, our evaluation concentrates primarily on measuring its precision in recognizing pathology-related medical terminology.

Whisper[27], a weakly supervised speech recognition model developed by OpenAI, has gained widespread adoption for automatic speech recognition in medical communication[28]. We employed Whisper (specifically the whisper-medium model) to automatically transcribe diagnostic comments from pathology reports of 97 prostate cancer patients obtained from The Cancer Genome Atlas (TCGA) database. The textual descriptions extracted from each case report's comment section were first converted to audio format, then processed through the Whisper model, with transcriptions saved as text in a CSV file (Google Drive in Supplementary).

To evaluate the automatic speech recognition performance on these prostate cancer diagnostic comments (N=97), we conducted quantitative assessments using word error rate (WER)[29], character error rate (CER)[30], match error rate (MER)[30], word information preserved (WIP)[29], word information lost (WIL)[29], and BLEU score metrics [31].

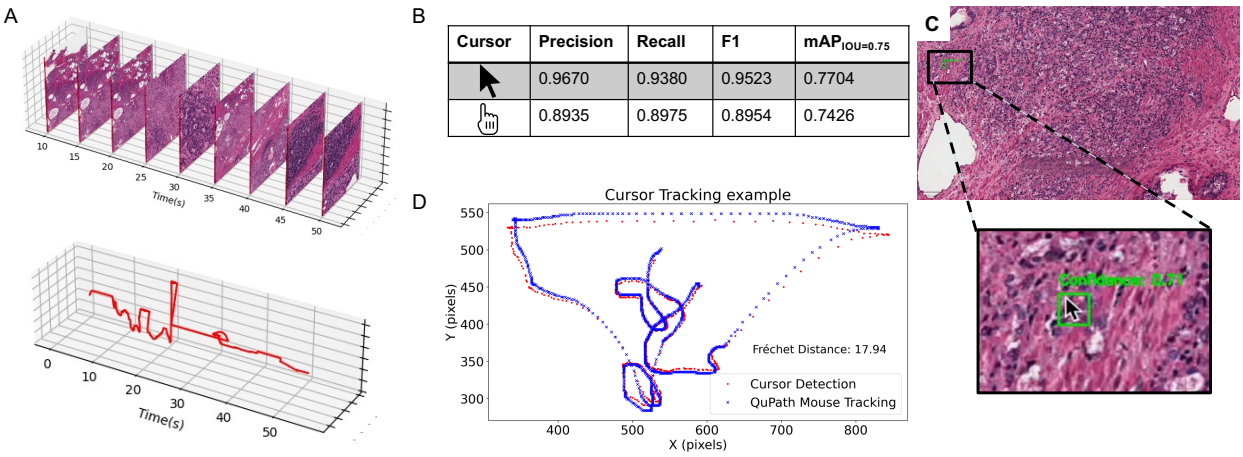
Results

1. The slide reviewing actions were successfully captured by our PathInteract.

a. Mouse cursor movement tracking

A synthetic dataset was created to train a YOLOv8 model for mouse cursor detection. The two most common cursor types observed in pathology review videos, “arrow” and “hand” (**Figure 2B**), were used for this purpose. Background images were randomly sampled from five bulk tissue whole slide images (WSIs), each stored at four resolution levels. For each synthetic sample, a 1920 × 1280-pixel region of interest (ROI) was randomly selected, and one of the two cursor types was randomly overlaid.

In total, 20,000 images were generated and divided into training (60%), validation (20%), and testing (20%) subsets. The YOLOv8 model was trained with a batch size of 8 and an initial learning rate of 0.001. Additional training details are provided in the accompanying source code. The model converged within 85 epochs, and its detection performance is summarized in **Figure 2B**. During both training and inference, only the top prediction was retained per image, provided its detection confidence exceeded 50%. Detected cursors were visualized using bounding boxes annotated with confidence scores (**Figure 2C**). The model demonstrated high accuracy and consistency across test samples, confirming the feasibility of using synthetic data to train effective cursor detection models for digital pathology workflows.



Cursor	Precision	Recall	F1	$mAP_{IOU=0.75}$
	0.9670	0.9380	0.9523	0.7704
	0.8935	0.8975	0.8954	0.7426

Figure 2. A) Mouse cursor movement tracking results, including representative key frames and the cursor locations across all time points. B) Quantitative evaluation of mouse cursor tracking accuracy. C) Example frame showing detected mouse cursor with bounding box and confidence score. D) Comparison of detected and ground-truth cursor coordinates (time frame: 7–13s, corresponding to subfigure A); the Fréchet Distance between the two trajectories is 17.94.

We then applied the mouse cursor tracking model to a QuPath recorded video to evaluate the effectiveness of our detection module. Each video frame was extracted and processed using the trained YOLOv8 model to capture cursor location variations over time (**Figure 2A**). However, we observed discrepancies between our model’s predictions and the cursor positions recorded in QuPath’s view tracking logs (**Figure 2D**). One notable difference is that cursor positions in the QuPath logs are sampled at uniform time intervals, whereas our video-based detections are sparser, constrained by the video’s frame rate. Additionally,

slight spatial offsets were observed between the detected cursor positions and those from the QuPath logs, likely due to minor differences in aspect ratios between the original QuPath display window and the exported video format. These findings highlight the challenges of achieving precise alignment between video-derived detections and log-based tracking data. We anticipate that applying temporal interpolation to the video-based cursor predictions prior to evaluation could improve alignment and enhance the accuracy of quantitative comparisons.

b. Viewport actions

During our implementation, we observed several limitations of directly using optical flow as the primary method for detecting view actions. Specifically, because the image viewer's user interface (UI) remains static across video frames, key point detection can be misleading, asmany detected points are located on non-informative UI elements such as the toolbar. This issue is exacerbated when the viewport displays homogeneous regions of the pathology slide, where few meaningful key points can be identified and tracked. To address this, we computed the standard deviation across frames to identify and exclude regions with minimal variation,typically corresponding to the static UI. By masking out these low-variance areas using averaged Otsu threshold value to 100.70, the input to optical flow was limited to the regions that the slide images were actually being inspected (**Figure 3A**).

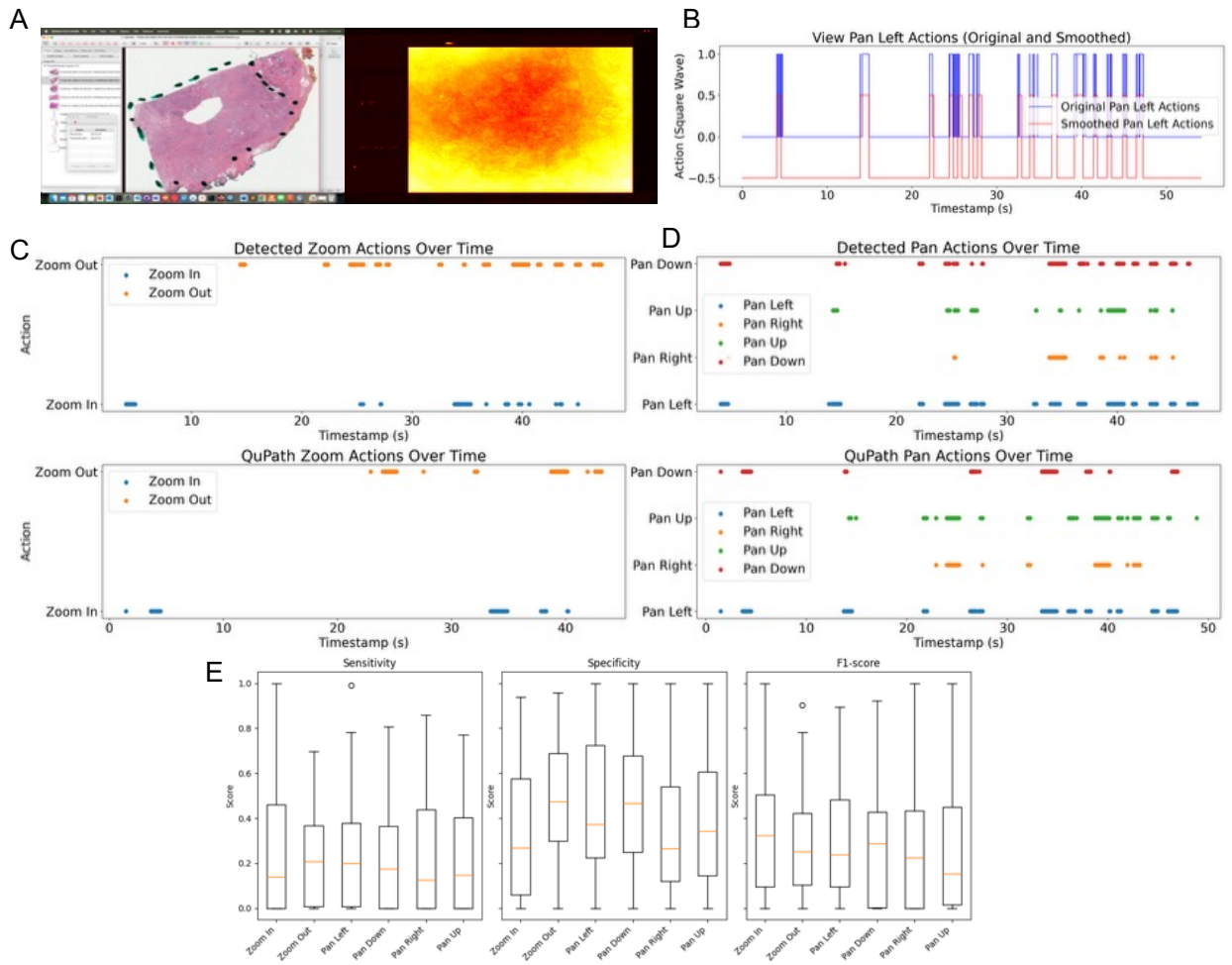


Figure 3. A) An example heatmap showing the content variation in the slide viewer. Higher values (hotter colors) indicate the pathologist spent more time on inspecting that region. B) An example of smoothing to the detected viewing action. C) An example of zoom in/out actions, detection results (first row) vs. ground truth (second row). D) An example of panning actions, detection results (first row) vs. ground truth (second row). E) Boxplot of the event-based sensitivity, specificity and F1 scores that measures the accuracy of viewport action detection.

Since the viewport actions are temporal events, standard performance evaluation metrics, such as sensitivity and specificity, may not always be the most appropriate and can even be misleading. A python library “timescoring”[32], developed to measure the performance of epilepsy detection from EEG signals, was used to measure the performance of our viewport action detection modual. Important parameters for the assessment including tolerance to start and end time (set to 0.5s and 1s respectively), minimum overlap (set to 0s), maximum event duration (set to 8s), and minimum duration between events (set to 1s). With these settings, the event-based sensitivity, specificity and F1 scores were calculated for quantitative measurements.

To restrain signal jittering in the detected actions, we smoothed the signals by filtering out fast flippings before quanlitative and quantitative assessment (**Figure 3B**). The zooming and panning actions were visually compared with the back-end tracking in QuPath(**Figure 3C, 3D**), showing comparable viewport actions, especially for panning. The quantitative evaluation (sensitivity, specificity and F1 scores) results demonstrated that event-based scorings are encouraging (**Figure 3E**).

c. Speech to text

Our evaluation primarily focuses on assessing its accuracy in recognizing pathology-related medical terms. The evaluation results yielded mean ± standard deviation values of 0.42 ± 0.32 for WER, 0.18 ± 0.24 for CER, 0.40 ± 0.27 for MER, 0.47 ± 0.28 for WIP, 0.53 ± 0.28 for WIL, and 0.41 ± 0.31 for BLEU score (**Figure 4A**).

To better understand transcription error patterns and their underlying causes, we randomly selected 19 diagnostic comments from the total dataset for manual evaluation by an internal medical resident physician (Google Drive in Supplementary). This manual assessment revealed four predominant error patterns: 1) occasional substitution of medical terms with phonetically similar but incorrect alternatives; 2) consistent difficulty recognizing pathology-specific terminology such as "Gleason pattern"; 3) errors in transcribing complex directional descriptors and spatial orientation terms; 4) and content omission in repetitive sequences where similar sentence structures resulted in significant intervening content being dropped (**Figure 4B**).

These transcription error patterns highlight critical areas for future enhancement of the speech recognition module, particularly regarding specialized medical vocabulary and complex sentence structures. These findings suggest that implementing a fine-tuned, medical or pathology specific large scale speech recognition model may substantially improve the performance of this module[33].

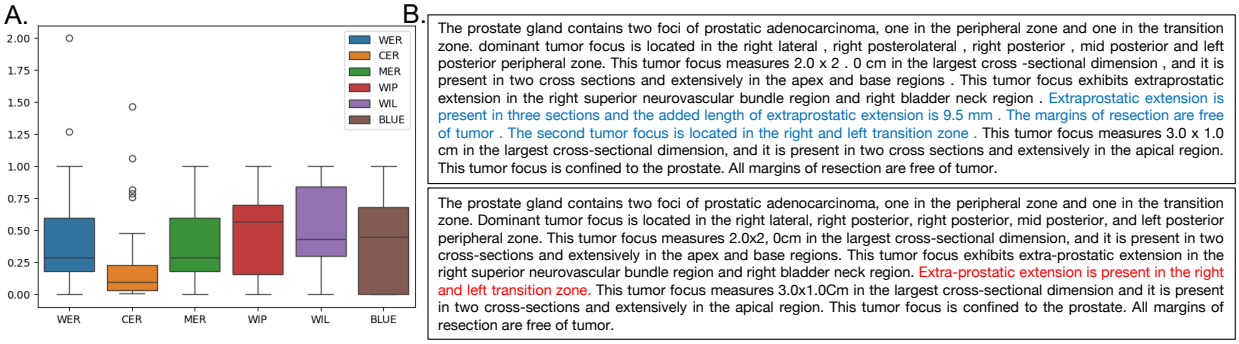


Figure 4. A) Quantitative evaluation results of speech recognition. B) An example of original text (upper) vs. transcription (lower) from Wisper. The highlighted content are the sentences with errors. The quantitative evaluation metrics for this case are WER: 0.36, CER: 0.18, MER: 0.36, WIP: 0.56, WIL: 0.44, BLUE: 0.5.

2. Distinct viewing patterns were observed across pathologists and tissue types

With the QuPath view-tracking records, we analyzed slide review behaviors across two pathologists with different levels of clinical experience, focusing on two tissue types: bulk resections and needle biopsies (**Figure 5A**). The results revealed notable differences in viewing patterns related to case review time, zoom level, and mouse cursor movement speed (**Figure 5B**).

The more experienced pathologist (Pathologist 1) spent less average time per case, suggesting more efficient navigation and decision-making. Additionally, Pathologist 1 tended to use lower magnification levels when reviewing bulk tissue slides, possibly relying more on global structural patterns rather than high-resolution detail. However, when reviewing needle biopsy slides, both pathologists used similar resolution levels, indicating that the diagnostic demands of smaller tissue samples may drive consistent zoom behavior (**Figure 5C**).

We also found that Pathologist 1 engaged in fewer zoom-in/zoom-out actions compared to Pathologist 2 across both tissue types, suggesting a more stable viewing strategy. While the panning behavior (horizontal and vertical movements within the slide) was generally comparable between the two pathologists for the same tissue type, there were significant differences in panning patterns between bulk tissue and needle biopsy slides (**Figure 5D**). This likely reflects the fact that needle biopsy specimens are mounted in a more uniform, linear orientation, which constrains navigation paths, whereas bulk tissues are larger and more heterogeneous, requiring more extensive scanning across regions.

These findings highlight the influence of both pathologist experience and tissue type on digital slide navigation behaviors and offer insights into how viewing strategies adapt to different diagnostic contexts. More example videos and our

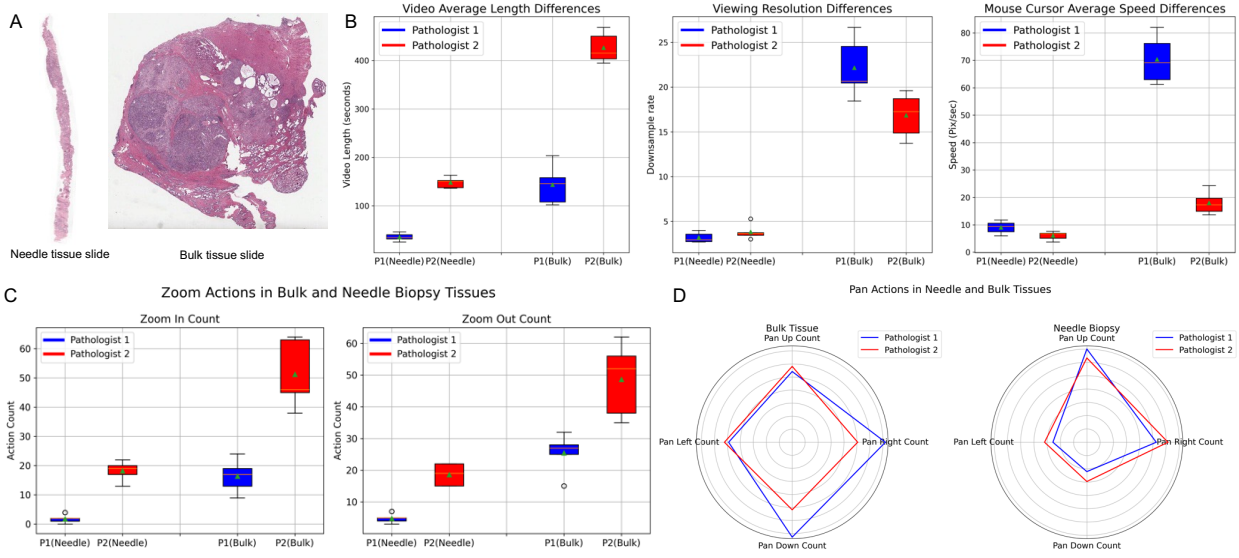


Figure 5. A) Example of needle biopsy and bulk tissue. B) View pattern differences between two pathologists on two different tissue types, including inspection time length, viewing

resolution and average mouse cursor speed. C) Zooming and D) Panning actions in bulk and needle biopsy tissues between two pathologists.

3. Online videos demonstrated distinct reviewing patterns across different diseases

We applied PathInteractto the online videos downloaded from YouTube. While educational video series from the same YouTube creator often exhibited consistent patterns across episodes, we observed distinct slide reviewing patterns across different diseases. For instance, mouse cursor moves much faster, and the detected mouse cursor moving events were more frequent in dermatology related YouTube Videos(**Figure 6A**). Moreover, zoom in/out actions were significantly less frequent in the YouTube videos compared to the QuPath tracking logs (**Figure 6B**). This discrepancy is likely due to the educational nature of the YouTube content, where the primary goal is teaching rather than clinical efficiency. In contrast, the QuPath tracking recordings captured pathologists actively mimicking a real-world diagnostic workflow, which naturally involves more intensive and efficient navigation of the slides. Meanwhile, we found that in hematology videos, most of the frames were regions cropped from slide scans, while nearly all the dermatology videos used whole slide images.

Another notable distinction between educational or case-sharing pathology videos and clinical recordings is the presence of verbal narratives. In clinical practice, pathologists typically do not speak aloud while reviewing slides, as doing so may disrupt efficiency and workflow. In contrast, pathologists in YouTube-based educational videos frequently verbalize their thought processes to aid in teaching and knowledge dissemination. Using our PathInteract to convert the audio into text, and exclude the stopwords (ie. is, look), we found that the word “cell” consistently appeared with the highest frequency across videos, regardless of the specific diseases being discussed (**Figure 6C**). The high frequency of “plasma”, “lymphocyte”, “mean” and “sort” can indicated the cell morphology and arrangement play important role in diagnosis, while “blood vessel” and “follicle” suggested that anatomy structures are significant for dermatology diagnosis. Further word tree analysis showed that pathologists focus on plasma cell in hematology, while endothelial cells were emphersized in dermatology (**Suppelimentary Figure 1**). The finding suggests that cell-level features are fundamental to the diagnostic reasoning shared by expert pathologists in educational contexts. The strong emphasis on cellular morphology highlights its importance in pathology education and underscores its potential value for training AI models and simulating diagnostic behavior.

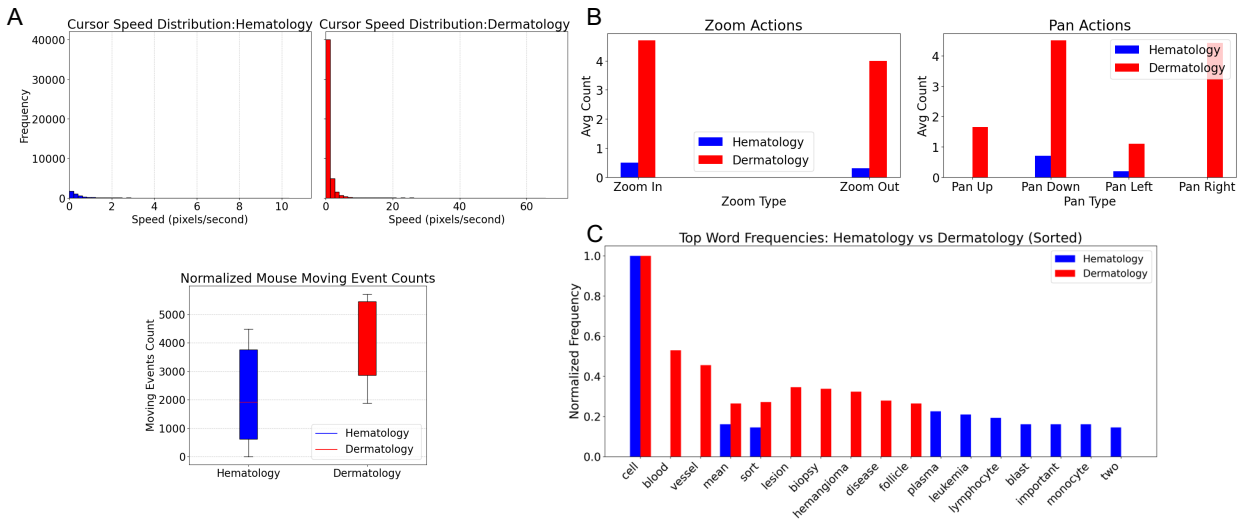


Figure 6. A) Mouse curor moving speed and number of detected mouse cursor moving events in two YouTube Channels. B) Averaged count of viewport actions, including Zoom in and out, Pan left, right, up and down. C) Word frequency histogram, top 10 most frequently words were included.

4. PathInteract facilitate extracting meaningful annotations from pathology videos

Annotated datasets provide significant enhancement for supervised AI model training. We evaluated the feasibility of leveraging information extracted by PathInteract to generate annotations that can be used for training AI models. Specifically, we explored two complementary sources of annotation: (1) mouse cursor tracking to detect regional annotations and (2) text captions to identify pathologists' descriptive statements and link them to corresponding video frames.

To identify regional annotations from mouse cursor tracking, we adopted a method similar to that used for detecting delineation behavior from trajectory data [34]. This approach computes the first and second derivatives of mouse cursor coordinates to capture both the direction and change in direction of movement, thereby estimating the curvature of the trajectory within a given time frame. The curvature metric

$$K = \frac{|x'y'' - y'x''|}{(x'^2 + y'^2)^{3/2}} \qquad (6)$$

quantifies the likelihood that the cursor movement corresponds to a regional annotation event. We applied this method to an example mouse tracking dataset, and intervals with a mean curvature above the threshold (mean curvature > 0.15) were highlighted (**Figure 7A, B**). By mapping the coordinates back to the original image, these cursor tracking results corresponded to two regions annotated by the pathologist (**Figure 7C**), suggesting that regional annotations can be identified based on mouse trajectory tracking results from our pipline.

To identify text descriptions associated with regional or cellular annotations, we used ChatGPT-4 to analyze the speech-to-text captions extracted by PathInteract. Two main prompts were applied: (1) Highlight sentences that are associated with pathologists'

regional or cell annotations and (2) Separate multiple annotations within the paragraph based on contextual clues. ChatGPT-4 successfully identified and separated annotation related sentences, as illustrated in **Figure 7D**. Based on the speech recognition timestamps, we then aligned these annotations with the corresponding video frames (**Figure 7E**).
Together, these results demonstrate that both mouse cursor tracking and caption analysis can be integrated to extract annotations (region delieantions and image–text pairs) from pathology videos, offering a rich source of data that could be used for AI model training and validation.

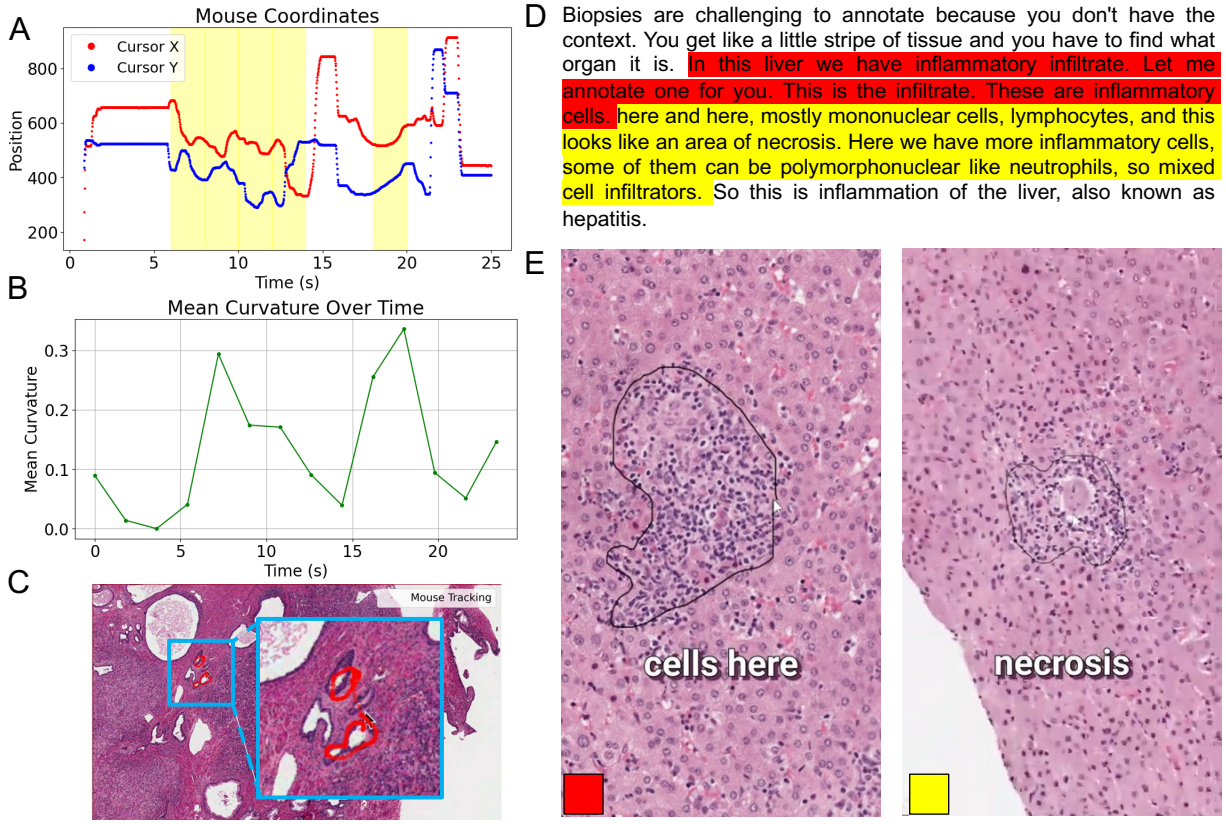


Figure 7. A) An example of mouse cursor tracking result. Red and blue lines denote the x and y coordinates over time. The intervals (6~14s and 17~20s) were highlighted based on mean curvature in B). C) Mouse cursor tracking highlights two circled regions in the image, indicating areas of focused interaction. D) ChatGPT highlighted caption sentences (red and yellow) from PathInteract as two annotations associate with liver cells and regions. E) Two video frames that corresponds to the descriptive sentences (red and yellow).

Discussion

We have developed an end-to-end suite, PathInteract, to extract pathologists' slide review behaviors, including mouse cursor movements, viewport zooming and panning, and verbal narratives. By integrating state-of-the-art methods within each module, PathInteract

achieves high accuracy across multiple behaviroal signals, providing a systematic approach to capture and analyze the dynamic, multimodal interactions that occur during digital slide review. By incorporating a broader range of viewing pattern recognition techniques, we have demonstrated that it is possible to repurpose pathology video datasets for advanced AI research. We also evaluated modules within PathInteract leveraging QuPath-recorded sessions and incorporating publicly available educational YouTube videos, strengthening the generalizability of our findings and reflecting real-world diagnostic practice. Nonetheless, we recognize that standardized reading tasks in controlled settings would provide additional value by helping disentangle the effects of expertise and personal habits on diagnostic decision-making.

Despite these contributions, several technical limitations remain. Mouse cursor tracking is challenged by the variability of cursor shapes across platforms and users, particularly the "cross" cursor common in some viewers. Current tracking focuses on frame-level cursor localization without analyzing tracking accuracy at different frame rate and cursor moving speed. Viewport action detection is constrained by limitations in optical flow: low texture or visual contrast regions can cause missed or inaccurate identification of zooming and panning events. Rapidly viewport changes can reduce detection reliability. Future improvements may include sequence-to-sequence (seq2seq) models [35] to encode temporal frame embeddings and dedicated zoom-level detection to identify zoom peaks, corresponding to critical inspection points for focused analysis. Speech recognition of pathologists' verbal narratives remains challenging due to domain-specific terminology. Future improvements could involve the use of pathology-specific language models or fine-tuning existing speech recognition models on pathology lecture data to improve accuracy.

Future work will also expand current suite to detect higher-level behavioral events. While we have demonstrated the feasibility of identifying delineation behaviors via cursor tracking, and sycronizing verbal descriptions to frames using large language models (e.g., ChatGPT-4), performance requires validation on larger and more diverse datasets. Detecting general viewing behaviors, such as zoom peaks, fixations, and slow panning intervals, could provide additional references for extracting annotations from pathology videos. Zoom peak detection, for instance, can help determine optimal resolution levels for downstream deep learning models, improving computational efficiency. Similarly, identifying fixation points and slow panning movements may reveal diagnostically significant regions, thereby enhancing model interpretability and clinical relevance.

PathInteract has significant potential to enhance both clinical practice and pathology education. In clinical practice, behavior-informed supervisory signals derived from cursor paths, zoom peaks, and viewport dynamics could be used to train AI models that mimic expert workflows, prioritize diagnostically relevant areas, and support more explainable decision-making. Such systems may reduce oversight, increase diagnostic efficiency, and foster trust in AI-assisted pathology. In medical education, PathInteract provides objective metrics to assess trainees' slide review strategies, including time spent on key regions and alignment with expert navigation patterns. These metrics could underpin adaptive feedback systems to acclerate trainee skill acquisition. Moreover, curated video-language datasets

generated by PathInteract can train next-generation vision-language models, enabling AI systems to learn both from static images and the reasoning process of expert pathologists.

Data availability

The source code of our suite is available in our GitHub repo at:
https://github.com/smujiang/PathInteract
Examples of slide reviewing recordings, QuPath view tracking videos, videos demonstrating speech recognition and other resources can be found in our Google Drive at:
https://drive.google.com/drive/folders/1zOF2Dbh5f-CQ1yXGlEqoFZZ74w3lTdVL

Ethics Statement and Patient Consent

This study utilized publicly available pathology image datasets that contain no patient-identifiable information. As such, institutional review board approval and informed patient consent were not required. All data were handled in accordance with relevant ethical guidelines and regulations.

Disclosure

The authors acknowledge the use of Claude Sonnet 4 to assist with language editing during the manuscript preparation process. This assistance was implemented under the authors' direct oversight and control, and all content was carefully reviewed and edited by the authors prior to submission.

The authors acknowledge the use of publicly available educational pathology videos from the YouTube channels "Cockerell Dermatopathology" (hosted by Clay Cockerell, MD) and "Jerad Gardner, MD" (hosted by Jerad Gardner, MD). Videos from these two channels were used exclusively for research purposes to analyze diagnostic review behaviors. The content was not modified in any way, nor was it redistributed outside of its original platform. This use is in accordance with YouTube's Terms of Service and policies regarding non-commercial, academic, and transformative use of publicly accessible content. The analysis results do not represent any clinical judgment or diagnostic evaluation of the cases presented in the videos. Furthermore, the findings do not reflect any opinion or professional practices of any individuals featured in the videos.

Acknowledgement

Research reported in this publication was supported by the National Center for Advancing Translational Science of the National Institutes of Health under award number U01TR002062, by the National Library of Medicine under award number R01LM011934, National Institute of Aging grant RF1AG072799, and by the Cancer Prevention Institute of Texas (CPRIT) under award number RR230020. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Advancing Translational Science, the National Library of Medicine, the National Institutes of Health, or the State of Texas.
Since some of the slide images were downloaded from TCGA, the results presented in this work here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

1. Brunyé, T.T., et al., *Machine learning classification of diagnostic accuracy in pathologists interpreting breast biopsies.* Journal of the American Medical Informatics Association, 2024. **31**(3): p. 552-562.
2. Ghezloo, F., et al., *An analysis of pathologists' viewing processes as they diagnose whole slide digital images.* Journal of Pathology Informatics, 2022. **13**: p. 100104.
3. Lopes, A., et al., *The Effect of Window Size on Pathologists' Search for Rare Elements in a Digital Pathology Setting.* Archives of Pathology & Laboratory Medicine, 2025.
4. Foucar, E., *Diagnostic decision-making in anatomic pathology.* Pathology Patterns Reviews, 2001. **116**(suppl_1): p. S21-S33.
5. Sudin, E., et al., *Digital pathology: the effect of experience on visual search behavior.* Journal of Medical Imaging, 2022. **9**(3): p. 035501-035501.
6. Mercan, E., et al., *Localization of Diagnostically Relevant Regions of Interest in Whole Slide Images: a Comparative Study.* J Digit Imaging, 2016. **29**(4): p. 496-506.
7. Ghezloo, F., et al., *Robust roi detection in whole slide images guided by pathologists' viewing patterns.* Journal of Imaging Informatics in Medicine, 2025. **38**(1): p. 439-454.
8. Onega, T., et al., *Accuracy of Digital Pathologic Analysis vs Traditional Microscopy in the Interpretation of Melanocytic Lesions.* JAMA Dermatol, 2018. **154**(10): p. 1159-1166.
9. Lopes, A., A.D. Ward, and M. Cecchini, *Eye tracking in digital pathology: A comprehensive literature review.* Journal of Pathology Informatics, 2024. **15**: p. 100383.
10. Alamudun, F.T., *Eye tracking methods for analysis of visuo-cognitive behavior in medical imaging*. 2016.

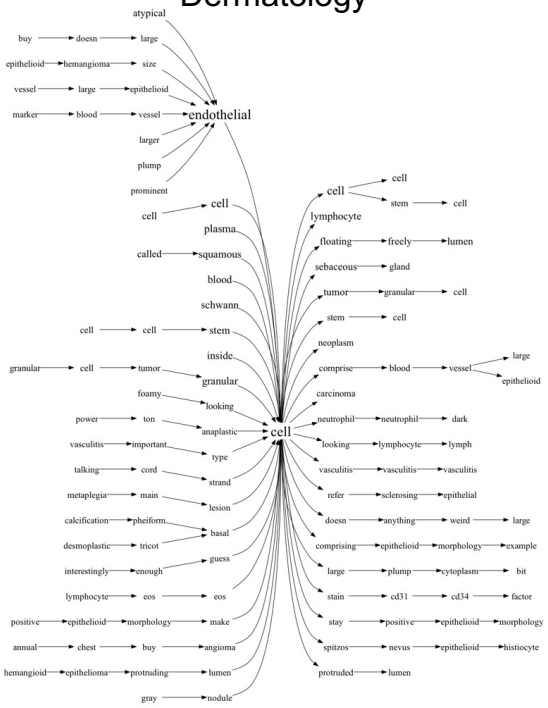
11. Gu, H., et al. *Augmenting pathologists with NaviPath: design and evaluation of a human-AI collaborative navigation system*. in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023.
12. Lewis, J.P. *Fast template matching*. in *Vision interface*. 1995. Quebec City, QC, Canada.
13. He, K., et al. *Mask r-cnn*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
14. Jiang, P., et al., *A Review of Yolo algorithm developments.* Procedia computer science, 2022. **199**: p. 1066-1073.
15. Corvò, A., M.A. van Driel, and M.A. Westenberg. *PathoVA: A visual analytics tool for pathology diagnosis and reporting*. in *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*. 2017. IEEE.
16. Horn, B.K. and B.G. Schunck, *Determining optical flow.* Artificial intelligence, 1981. **17**(1-3): p. 185-203.
17. Crowley, R.S., et al., *Development of visual diagnostic expertise in pathology: an information-processing study.* Journal of the American Medical Informatics Association, 2003. **10**(1): p. 39-51.
18. Zhang, H., et al., *PathNarratives: Data annotation for pathological human-AI collaborative diagnosis.* Frontiers in Medicine, 2023. **9**: p. 1070072.
19. Ikezogwo, W.O., et al., *Quilt-1M: One Million Image-Text Pairs for Histopathology.* Adv Neural Inf Process Syst, 2023. **36**(DB1): p. 37995-38017.
20. Sun, Y., et al., *Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration.* arXiv preprint arXiv:2407.00203, 2024.
21. Weinstein, J.N., et al., *The cancer genome atlas pan-cancer analysis project.* Nature genetics, 2013. **45**(10): p. 1113-1120.
22. Bulten, W., et al., *Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge.* Nature medicine, 2022. **28**(1): p. 154-163.
23. Varghese, R. and M. Sambath. *Yolov8: A novel object detection algorithm with enhanced performance and robustness*. in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. 2024. IEEE.
24. Jocher, G., et al., *ultralytics/yolov5: v3. 0.* Zenodo, 2020.
25. Alt, H. and M. Godau, *Computing the Fréchet distance between two polygonal curves.* International Journal of Computational Geometry & Applications, 1995. **5**(01n02): p. 75-91.
26. Koenderink, J.J., *Optic flow.* Vision research, 1986. **26**(1): p. 161-179.
27. Radford, A., et al. *Robust speech recognition via large-scale weak supervision*. in *International conference on machine learning*. 2023. PMLR.
28. Zolnoori, M., et al., *Decoding disparities: evaluating automatic speech recognition system performance in transcribing Black and White patient verbal communication with nurses in home healthcare.* JAMIA open, 2024. **7**(4): p. ooae130.
29. Morris, A.C., V. Maier, and P.D. Green. *From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition*. in *Interspeech*. 2004.

30. Graves, A., et al. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*. in *Proceedings of the 23rd international conference on Machine learning*. 2006.
31. Papineni, K., et al. *Bleu: a method for automatic evaluation of machine translation*. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002.
32. Dan, J., et al., *SzCORE: seizure community open-source research evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms*. Epilepsia, 2024.
33. Roushan, R., et al. *Optimizing Speech Recognition for Medical Transcription: Fine-Tuning Whisper and Developing a Web Application*. in *2024 IEEE Conference on Engineering Informatics (ICEI)*. 2024. IEEE.
34. Flash, T. and N. Hogan, *The coordination of arm movements: an experimentally confirmed mathematical model*. Journal of neuroscience, 1985. **5**(7): p. 1688-1703.
35. Sutskever, I., O. Vinyals, and Q.V. Le, *Sequence to sequence learning with neural networks*. Advances in neural information processing systems, 2014. **27**.

Supplementary materials

Supplementary Figure 1. Word tree created from the transcription from dermatology (left) and hematology (right)YouTube videos.

Dermatology



Hematology