



# Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos

Mehmet Saygin Seyfioglu<sup>♣,†\*</sup>      Wisdom O. Ikezogwo<sup>♣,‡</sup>      Fatemeh Ghezloo<sup>♣,‡</sup>  
 Ranjay Krishna<sup>†,‡</sup>      Linda Shapiro<sup>†</sup>

<sup>‡</sup>University of Washington

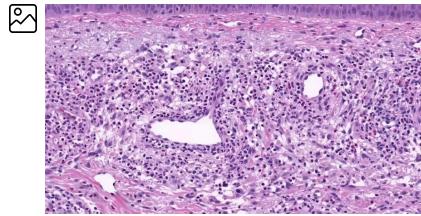
<sup>†</sup>Allen Institute for AI (AI2)

## Abstract

*Diagnosis in histopathology requires a global whole slide images (WSIs) analysis, requiring pathologists to compound evidence from different WSI patches. The gigapixel scale of WSIs poses a challenge for histopathology multi-modal models. Training multi-modal models for histopathology requires instruction tuning datasets, which currently contain information for individual image patches, without a spatial grounding of the concepts within each patch and without a wider view of the WSI. To bridge this gap, we introduce QUILT-INSTRUCT, a large-scale dataset of 107,131 histopathology-specific instruction question/answer pairs, grounded within diagnostically relevant image patches that make up the WSI. Our dataset is collected by leveraging educational histopathology videos from YouTube, which provides spatial localization of narrations by automatically extracting the narrators' cursor positions. QUILT-INSTRUCT supports contextual reasoning by extracting diagnosis and supporting facts from the entire WSI. Using QUILT-INSTRUCT, we train QUILT-LLAVA, which can reason beyond the given single image patch, enabling diagnostic reasoning across patches. To evaluate QUILT-LLAVA, we propose a comprehensive evaluation dataset created from 985 images and 1283 human-generated question-answers. We also thoroughly evaluate QUILT-LLAVA using public histopathology datasets, where QUILT-LLAVA significantly outperforms SOTA by over 10% on relative GPT-4 score and 4% and 9% on open and closed set VQA<sup>1</sup>.*

## 1. Introduction

In histopathology, analyzing giga-pixel-level whole slide images (WSIs) requires a nuanced understanding beyond examining isolated patches taken within the WSI. Patholo-



 User: Can you provide a concise description of the histopathology image shown?

 Quilt-LLaVA: Upon examination of the histopathology image, it is evident that there is a significant presence of lymphocytes. These cells are densely packed and exhibit a high degree of pleomorphism, indicating a possible pathological condition. The lymphocytes are irregular in shape and size, with some displaying enlarged nuclei and prominent nucleoli. In addition to the lymphocytes, there is also a noticeable presence of eosinophils. These cells are characterized by their bright pink cytoplasm and bilobed nuclei. The eosinophils are scattered throughout the tissue, suggesting an ongoing immune response. The overall cellular architecture appears disrupted, with the normal tissue architecture replaced by a dense infiltrate of lymphocytes and eosinophils. This could be indicative of an inflammatory or immune-mediated process. Further diagnostic tests may be required to confirm the exact nature of this pathological condition.

 User: Given the presence of a mixture of lymphocytes and eosinophils, what can be deduced from this single image regarding a possible diagnosis?

 Quilt-LLaVA: The presence of a mixture of lymphocytes and eosinophils in this patch suggests an inflammatory response. This could be seen in a variety of conditions, including infections, autoimmune diseases, and certain types of skin lesions. However, without additional context, it's challenging to make a definitive diagnosis based solely on this feature. It's important to note that the presence of these cells alone does not necessarily indicate a malignant condition. In fact, these cells can be seen in a variety of benign conditions as well. Therefore, a comprehensive evaluation of the whole slide image, including the overall architecture of the lesion, the presence of a junctional and dermal component, and the absence of atypia or mitotic activity, would be necessary to make a more accurate diagnosis.

Figure 1. QUILT-LLAva is capable of **describing** the prominent medical regions within a histopathology patch. Additionally, it can be utilized to **reason** towards a diagnosis based on the current observations. Note: The image includes eosinophils and lymphocytes, and is sampled from a WSI showing rare benign dermatitis.

gists typically analyze these WSIs holistically, compounding evidence from various regions to form a diagnosis. While

\*msaygin@cs.washington.edu; ♦ Denotes equal contribution.

<sup>1</sup>Our code, data, and model is publicly accessible at [quilt-llava.github.io](https://quilt-llava.github.io).

navigating the WSI regions, pathologists engage in a critical reasoning process over a multi-turn conversation, determining what morphological concepts within the regions confirm or rule out specific diagnoses, making the process dynamic and context-sensitive. For example, in liver tissue WSI, isolated steatosis (fat accumulation) might not be indicative of significant pathology, but if accompanied by ballooning of hepatocytes and lobular inflammation, it strongly suggests non-alcoholic steatohepatitis, a more serious condition that could lead to cirrhosis or liver cancer [5]. This approach involves understanding the diagnostic relevance of histological visual features, spatial relationships, and aggregating insights across the WSI [1]. Although current histopathology multi-modal models can analyze isolated image patches effectively, they often lack the capability to reason beyond that patch to determine next steps, thereby limiting their utility.

Meanwhile in natural images, the success of the Large Language and Vision Assistant (LLaVA) [14] demonstrates the possibility of multi-modal models as chat partners. LLaVA is trained by extracting instruction-tuning data with Large Language Models (LLMs), such as GPT-4 [19]. Moreover, photos usually include multiple objects which can be referred to using bounding boxes, which can be used to create spatially grounded instruction-tuning data [14]. Recently, multi-modal models have also begun employing video content, allowing for more complex reasoning beyond single images [13]. However, this progress has not translated to histopathology yet, where existing multi-modal models often rely on PubMed articles for image-caption pairs [14, 18]. These articles lack two critical elements: **1)** visually grounded captions necessary for effective visual grounding of histopathology concepts, and **2)** the broader contextual information from the WSI, essential for reasoning beyond single image patches, limiting their effectiveness in detailed histopathological analysis.

To fill this void, we present QUILT-INSTRUCT, an instruction-tuning dataset of 107,131 histopathology-specific question/answer pairs. Similar to QUILT [9], QUILT-INSTRUCT capitalizes on educational histopathology videos from YouTube, featuring pathologists narrating their examination of WSIs. **1)** To alleviate the limited spatial awareness, we extract narrators' mouse cursors from videos, grounding histopathology concepts within each patch using spatio-temporal clustering. **2)** To enable reasoning, we propose novel instruction-tuning QA prompting techniques for histopathology: Complex reasoning, and iterative abductive reasoning, which incorporates the global WSI diagnosis and its supporting facts with image captions to ground factual information, preventing hallucinations (see Fig. 18).

Using QUILT-INSTRUCT, we train QUILT-LLAVA, a multi-modal model for histopathology, with its capabilities illustrated in Fig 1. QUILT-LLAVA undergoes a two-stage training process (see Fig 3). First, it is aligned with the

histopathology domain using 723K image-text pairs from QUILT [9], and then it is further instruction-tuned with QUILT-INSTRUCT. QUILT-LLAVA analyzes given images in detail, spatially localizes medical concepts, and reasons beyond the given image patch by guiding users on what further evidence is needed to validate or rule out certain diagnoses, and can even be used as an educational tool that, instead of directly revealing a diagnosis, subtly hints at it, aiding pathology students in their training (Appendix Section 2.1.)

Finally, educational videos are valuable for their interactivity, as narrators often engage viewers by asking and answering questions throughout their presentations. Leveraging this, we propose QUILT-VQA, an organic evaluation dataset extracted from naturally occurring questions and answers from QUILT videos with the help of GPT-4 and manual verification, to evaluate QUILT-LLAVA's reasoning capabilities. We evaluate QUILT-LLAVA on QUILT-VQA and two public histology VQA test sets on both open and closed questions. Using red circle  [22] marking the area of interest in the image, we can prompt QUILT-LLAVA to focus on specific regions of the image patch. We outperform the SOTA by 4% and 9% on open and closed set VQA tasks. To further evaluate the reasoning capabilities of QUILT-LLAVA, we utilize GPT-4 to score the model's generated response against two other SOTA multi-modal models: LLAVA [16] and LLAVA-MED [14]. QUILT-LLAVA outperforms LLAVA and LLAVA-MED by over 16% and 7%, respectively, and upon increasing the size of instruction-tuning data and only pre-training for three epochs, we achieve even better results of 10.8% over LLAVA-MED.

## 2. Related work

We built our work from the expanding body of literature in visual instruction-tuning and video-based dataset generation, with a specific focus on their application in the areas of medical image analysis and histopathology.

**Visual instruction-tuning in natural images.** Thanks to the open-source availability of LLMs [3, 11, 23, 24], studies in the general vision-language domain have advanced the training of multi-modal models by harnessing implicit visual understanding through generating an instruction-tuning dataset from image captions. Prior work [4, 17, 32] demonstrated significant capabilities, with LLaVA-1.5 matching GPT-4's performance in certain multi-modal tasks [16].

**Visual instruction-tuning in medical images.** Visual Med-Alpaca [6] created 54K question-answer pairs for instruction-tuning using GPT-3.5. PMC-VQA [31] curated a larger multiple choice answer based dataset from general medical domains using PubMed, yet its coverage of histopathology is limited. For instance, LLaVA-Med [14] employs image captions from PubMed articles and, in cases of brief captions, supplements them with sentences from the article, which may not directly pertain to the referenced figure. Within

LLaVA-Med, a subset of 17k images relates to histology, yielding 49K question-answer pairs. Furthermore, these works rely on isolated image-caption pairs for constructing instruction-tuning datasets, which limits GPT-4’s capacity to reason beyond its context or, if attempted, raises its likelihood of hallucination.

**Video-based image-text datasets.** Multiple works have attempted to curate datasets from videos in the natural domain [13, 25–28]. MIMIC-IT [13] uses video data to create various prompts, including some reasoning-based ones, to create instruction-tuning datasets. Video Localized Narratives [25] involve annotators verbally describing an image while concurrently moving their mouse cursor over the regions they refer to, yielding dense visual groundings. Acquiring a dataset in the same way is expensive in histopathology. PathNarratives [29] employed eight pathologists to manually annotate spatially grounded pathology data; however, it is not available for open access. More recently, QUILT [9] was proposed, which curates large-scale data from educational histopathology content from YouTube to create image-caption pairs, however, it lacks spatially grounded captions. We expand upon QUILT’s video content to generate grounded histopathology data. To the best of our knowledge, our work is the first to use videos for creating spatially grounded instruction-tuning datasets in the medical domain.

### 3. Curating QUILT-INSTRUCT

To construct a comprehensive histopathology instruction-tuning dataset with visual groundings in WSIs, we harness the rich narrative content of educational YouTube videos.

#### 3.1. Data preparation

We make the key observation that, in educational videos, narrators often pause while exploring large-scale WSIs before indicating salient areas with their cursor [10, 12, 20, 25]. Our process uses three steps to convert unstructured videos into usable visually-grounded instruction data: First, we **localize narrators’ cursors** in videos. Then, we perform **spatio-temporal clustering of cursor location** to visually ground histopathology concepts within images. These two steps are illustrated in Fig. 2. Finally, using the extracted grounded captions, we use an LLM to **generate our instruction-tuning dataset - QUILT-INSTRUCT**. This process involves prompting techniques, from independent prompts generating diverse Q/A pairs for each image patch to reasoning-based prompts combining information across patches in WSIs, creating Q/A pairs that reason towards diagnoses.

**Data source.** We construct QUILT-INSTRUCT from 4149 educational YouTube videos totaling over 1,000 hours of content. These videos are part of a recent histopathology dataset, QUILT [9]. The videos feature a *narrative style*, where pathologists articulate histopathological concepts while interacting with WSIs. QUILT provides rich image-text data

but does not offer spatial annotations to connect text descriptions with specific regions in the images—a key aspect often absent in most medical datasets. Fortunately, narrators’ mouse cursors highlight morphological features, offering visual grounding to bridge this gap. We next detail the process of extracting these cursors from raw video footage.

**Localizing Cursors.** Extracting the cursor location from histopathology images poses a significant challenge due to the homogeneity in color and texture, alongside the presence of other dynamic elements in videos, such as minor pixel variations and variations in the narrators’ cursor movement speed and style. Despite these obstacles, we have made a critical observation: Narrators typically pause their exploration of a WSI before gesturing with their cursor, guiding the audience’s attention. Therefore, if the background is stable for several consecutive frames, it is highly likely that the narrator may be using their cursor to mark medically relevant regions. Leveraging this observation, we employ a relatively straightforward method to detect mouse cursors. First, we isolate segments in each video where the background is mostly static, termed stable chunks shown in the leftmost section of Fig. 2. To detect the stable chunks, we utilize a simple algorithm (Algorithm 1 in the Appendix) that detects stable chunks of frames by employing a frame-differencing approach to detect chunks with minimal background movement. Our algorithm computes the absolute difference between consecutive frames and then applies a Gaussian filter for adaptive thresholding to pinpoint frames with minor changes. Given the typically uniform coloration of histopathology backgrounds due to histopathology staining, simple pixel-wise difference yields a high rate of false positives, namely, chunks that had a change occur in them are identified as stable. To alleviate this, we incorporate a perceptual metric, Structural Similarity Index Measure (SSIM). By evaluating SSIM on randomly sampled patches, we further verify whether a change has occurred between two frames. For each stable chunk where the background remains consistent, we calculate a median frame in the pixel domain to represent the background. This median frame is then subtracted from the frames within the stable chunk, and a thresholding is applied to mitigate noise or other artifacts, effectively isolating the cursor as illustrated in the middle section of Fig. 2. Subsequently, we capture the cursor by identifying the coordinates of the maximum pixel value. However, this approach assumes the absence of other slight movements with respect to pixel changes, which is not always the case, as there may be movements like the narrator’s facial expressions within the scene. We address this by applying a face detection model [21] to mask such distractions, ensuring our analysis concentrates solely on cursor movement. This algorithm offers a surprisingly generalizable way to collect cursor traces from any educational videos with similar *narrative styling* at a meager computational cost.

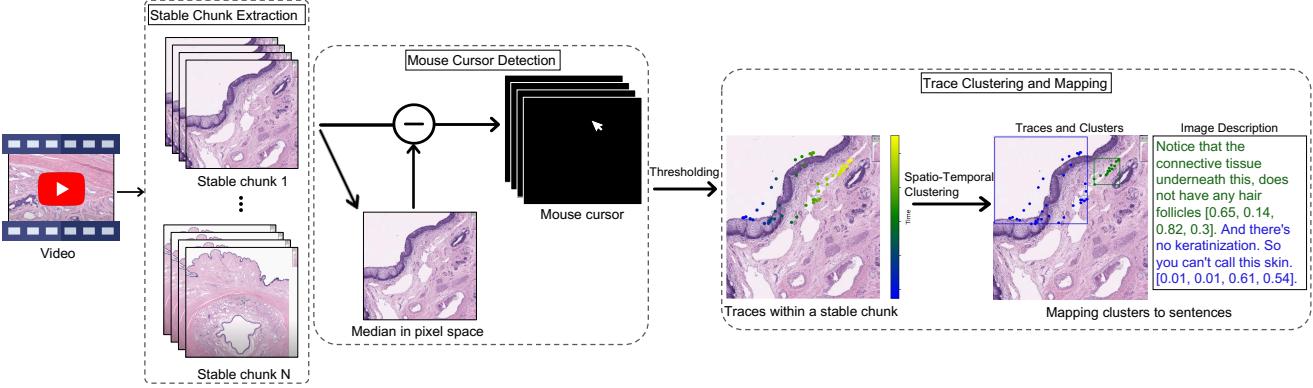


Figure 2. To create QUILT-INSTRUCT, we first identify stable chunks within the video. For each chunk, we compute a median frame in the pixel domain and subtract it from every frame within the chunk. We then apply a threshold to reduce noise and take the maximum value to capture the mouse cursor points. These cursor points are then clustered to localize medical content in image captions. Please note that color encodes time in the "Trace Clustering and Mapping" part of the figure.

After pinpointing these candidate stable chunks, we retain those with a minimum duration of three seconds. Subsequently, we utilize a histology classifier as described in [9] to eliminate non-histological segments using each stable chunk’s background frame. This methodology generates 132K image-caption pairs, represented as  $\langle I_j^v, c_j^v \rangle$ , where  $I_j^v$  is the median frame of the  $j^{th}$  stable chunk within the  $v^{th}$  video, and  $c_j^v$  is the corresponding caption for  $I_j^v$ , which is obtained by converting the narrator’s speech to text similar to the strategy outlined in [9]. Each pair spans an average of 24 seconds, corresponding to 55 words in the caption. From these 132K image caption pairs, 44,163 exhibits active cursor movement where we preserved cursor position as  $(x_j^t, y_j^t)$ , where  $x \in [0, W]$  and  $y \in [0, H]$ , with  $W$  and  $H$  representing the image width and height, respectively,  $t$  spans from 0 up to the total duration of the  $j^{th}$  stable chunk. Next, we demonstrate our approach for clustering extracted mouse cursors, which serves to visually ground our captions.

**Spatio-temporal clustering of the mouse cursor.** Next we proceed to cluster the tracer points  $(x_j^t, y_j^t)$  for each stable chunk to spatially ground narrators’ words. For spatio-temporal clustering, we transformed our inputs into  $(x_j^t, y_j^t, t, w_j^t)$ , where  $t$  is time and  $w_j^t$  denotes the cumulative word count at each cursor position, to encode the pace of speech better, and we have normalized the inputs to ensure each input is scaled within the standardized range of 0 to 1. Given this input format, to achieve a balance between spatial and temporal aspects — specifically, to determine the relative importance of spatial proximity and temporal proximity in cluster formation — we apply an exponential decay to the spatial coordinates,  $e^{-\lambda t}$ , where  $\lambda$  is set at 0.05, ensuring that points closer in time are preferentially clustered together. The algorithm dynamically selects the number of clusters based on the number of words in  $c_j^v$ . To refine the

process of mapping words to clusters, we first determine the temporal center of each cluster. This is achieved by calculating the mean timestamp of the tracer points within a cluster, which gives us a ‘temporal midpoint.’ Subsequently, we assign words to clusters based on their temporal proximity to these midpoints, ensuring that each word matches the cluster whose average time is closest to the word’s occurrence time. Finally, the resulting clusters are represented by bounding boxes, denoted in the format  $[x_1, y_1, x_2, y_2]$ , where each coordinate is a floating-point number normalized between 0 and 1. These coordinates specify the boxes’ top-left ( $x_1, y_1$ ) and bottom-right ( $x_2, y_2$ ) corners. A sample is shown in Fig. 2. Next, we show how we used our grounded captions to generate our instruction-tuning data QUILT-INSTRUCT.

### 3.2. Generating QUILT-INSTRUCT

We generate two sets of question-answer types. First, we employ Conversations and Detailed Descriptions prompts, akin to those in [17], that take as input single patch-level grounded-text to generate Q/A pairs constrained by the independent input sample; hence we call these prompts – Independent prompts. Secondly, we leverage the contextual continuity inherent in our dataset – that is, for a video reviewing a single WSI, we leverage the sequential unraveling of concepts/clues toward a final diagnosis, by introducing novel Reasoning-based Prompts which receive as input patch-level grounded-text and global WSI-level information. These enable an LLM (GPT-4) to extend its reasoning beyond its immediate context while still being anchored by the factual insights derived from the entire video, reducing its hallucinative behavior.

### 3.2.1 Independent Prompts

Following the approach of [17], we generate Q/A pairs from each image caption  $c_j^v$ , creating an array of questions that aim to explain the context presented in the image. However, these prompts are designed to elicit answers based solely on the information within the caption and do not attempt to extrapolate beyond it.

**Conversation.** The objective of the Conversation prompt is to construct a dialogue between an AI assistant and an individual inquiring about a histopathology image. We generate a diverse set of Q/A pairs based on the image caption  $c_j^v$ , which includes bounding boxes of medical concepts. The assistant, equipped with descriptions and bounding box coordinates, responds as if directly observing the image, discussing elements like cell types, activities, and relative positions of medical concepts within the image. The dialogue includes around 3-4 Q/A pairs, maintaining a conversational tone while focusing on the visual content of the images.

**Detailed Descriptions.** Similar to [17], we instruct GPT-4 to use  $c_j^v$  to generate a detailed description for a given image. Subsequently, questions are randomly selected from a pre-compiled list given in the Appendix Fig.19.

### 3.2.2 Reasoning-based Prompts

Most image-caption pairs used to create instruction-tuning datasets in other works, such as PubMed or COCO [15], exhibit contextual isolation, meaning  $\langle I_i, c_i \rangle$  and  $\langle I_j, c_j \rangle$  where  $i \neq j$  do not share context with each other. Consequently, Q/A pairs generated by GPT-4 will be constrained to the context of a single image. Given the gigapixel nature of histopathology images, deriving a comprehensive diagnosis based on a single image patch is often inadequate. The model should ideally reason beyond the given image, guiding the human user on subsequent steps or areas to scrutinize. This may be attempted using GPT-4's inherent medical knowledge. However, coaxing GPT-4 to extrapolate beyond a single caption without an interconnected context could lead to hallucinations, compromising data quality.

In contrast, our approach capitalizes on video-extracted image-caption pairs derived from a single patient's WSI. To that end, we manually reviewed the entire video content of QUILT identifying 2066 videos that feature only a single WSI for our reasoning-based prompts. By utilizing the entire text from these single whole-slide videos, we initially used GPT-4 (see Fig. 10 in Appendix) to infer the final diagnosis  $d_j^v$  and the supporting facts  $f_j^v$ . Utilizing  $f_j^v$  and  $d_j^v$  enables us to craft reasoning-based prompts that implicitly direct GPT-4 towards abductive reasoning. This approach fosters extrapolations more anchored in context, reducing hallucination tendency. We developed two prompt types, **Complex Medical Reasoning** and **Iterative Abductive Reasoning**, for reasoning-based instruction-tuning dataset generation.

**Complex Medical Reasoning.** Given a caption  $c_j^v$ , along with a diagnosis  $d_j^v$  and contributory facts  $f_j^v$ , we prompt GPT-4 in a diagnostic reasoning task designed to extrapolate beyond the immediate context of  $c_j^v$ . More broadly, we instruct GPT-4 to utilize its inherent medical knowledge to interpret the contents of a single image caption  $c_j^v$ , while subconsciously incorporating the diagnosis and supporting facts extracted from the entire video. If the observations from  $c_j^v$  suffice for making a diagnosis based on GPT-4's general medical knowledge, it proceeds to provide a diagnosis along with the facts leading up to it. If not, it extrapolates using  $f_j^v$  to suggest what to look for in the WSI to validate or rule out certain diagnoses. We term this approach "Subconscious Knowledge Injection," as it allows GPT-4 to leverage its medical knowledge while being subtly guided by the provided  $d_j^v$  and  $f_j^v$ , effectively constraining its context and ensuring focused reasoning and fewer hallucinations.

**Iterative Abductive Reasoning:** We simulate a conversation between two GPT-4 agents, mimicking a scenario where a professional pathologist uses our model to ask longer medically intricate questions about an image. This contrasts with the shorter questions typically found in other prompts that we use. The first agent, termed Human-GPT, is provided with an image patch caption  $c_j^v$  and is tasked with abductively reasoning about the possible diagnoses and the facts used to arrive at these conclusions. This is presented in the format: *User: {[Abduction: xxx}, {Facts Used: xxx}]*. The second agent, referred to as the AI Assistant GPT, is privy to the diagnosis  $d_j^v$  and facts  $f_j^v$ , simulating someone who has viewed the WSI of this particular patient. The AI Assistant evaluates the accuracy of the abduction derived by Human-GPT and provides comments or hints at potentially overlooked details using its inherent medical knowledge while utilizing  $d_j^v$  and  $f_j^v$ . This is communicated in the format: *GPT: {[Comments: xxx}, {Hint: xxx}]*. If Human-GPT has exhausted all information from  $c_j^v$  and a diagnosis cannot be made, the AI Assistant suggests exploring other patches for additional evidence, such as "Consider looking for evidence of X in other patches to validate your diagnosis.". Furthermore, leveraging Iterative Abductive Reasoning-based instruction-tuning data, QUILT-LLAVA can serve as an educational tool, aiding human users in brainstorming by providing hints about the next steps, without fully revealing the answer. (See Fig. 16 in Appendix). The conversation between the AI Assistant and Human-GPT proceeds with an upper limit of iterations, randomly selected between two to four interchanges per agent, where after each exchange, the start of a new exchange incorporates the post-exchange history, allowing GPT to retain the memory of past conversations. The dialogue may terminate before reaching this limit if Human-GPT arrives at a conclusive diagnosis, or if the AI Assistant determines that Human-GPT has fully exhausted all relevant information from  $c_j^v$ , and then guides it on what to do next using  $f_j^v$ .

**Dataset Statistics.** We extracted 162,566 image-caption pairs from QUILT. To refine this data, we filtered out captions with fewer than 20 words and those with more than 150 words. This process resulted in a dataset of 114,343 pairs, with an average caption length of 55 words. From this, we created QUILT-INSTRUCT, comprising 107,131 question/answer pairs where, on average, we have questions with 16.5 words and answers with 101 words. For reasoning-based prompts, we manually reviewed 4,149 videos and selected 2,066 that focused on a single WSI from a single patient.

## 4. Training QUILT-LLAVA & evaluating with QUILT-VQA

In this section, we detail using QUILT-INSTRUCT to train QUILT-LLAVA. Next, we curate QUILT-VQA independently of QUILT-INSTRUCT, to evaluate QUILT-LLAVA. Finally, we generate an Instruction Following Test Set from QUILT-VQA to assess QUILT-LLAVA’s ability to follow instructions.

### 4.1. Training QUILT-LLAVA

We embrace the LLAVA autoregressive model architecture for its simple yet efficient design. Additionally, our selection of the LLAVA architecture aimed to ensure consistency in evaluation against our baselines LLAVA [17] and LLava-MED [14], both of which utilize a curriculum-learning strategy [14] on instruction tuning sets. Overall, LLAVA integrates a vision module, an LLM, and an MLP connector, allowing the LLM to process visual information. Initially, the MLP —serving as a projector— is trained until it converges. During this stage, both LLM and the vision module kept frozen. Subsequently, both the MLP and the LLM are fine-tuned with instruction-following data to align the model with human pathologists. Our overall architecture is shown in Fig. 3. LLAVA typically uses a pre-trained CLIP image encoder; for our domain, we use the pre-trained CLIP model trained using public histopathology datasets such as QUILT-NET [9] and PLIP [8]. We also run ablations with various image encoders, training strategies, and visual prompts.

**Histopathology domain alignment.** First, we align our vision and language models within the histopathology domain. To that end, we extract 723K image-text pairs from QUILT and convert the captions into the Q/A format. To do that, we randomly select an instruction (Question) and prepend it to the caption (Answer) to create Q/A pairs. The instructions, drawn from a predefined list (See Appendix Fig. 18), are designed to variably describe the visual information present in the images. At this stage, we freeze vision and language models and only train the MLP layer, which is tasked to project the embeddings from the image encoder to enable the language model predict the images’ caption given the question. This stage aligns the histology image embeddings to their corresponding text embeddings.

**Histopathology instruction-tuning.** Finally, we fine-tune our model with QUILT-INSTRUCT. At this stage, we keep the visual encoder weights frozen and continue to train the MLP layer and the language module.

### 4.2. Evaluation Data Generation: QUILT-VQA

In the field of histopathology, researchers rely on evaluation datasets like PathVQA [7] and PMC-VQA [31] to assess the performance of their models. However, these datasets exhibit notable shortcomings, including significant repetitiveness due to paraphrasing the same question. Even worse, often-times, there are contradictory answers to the same question (See Appendix Section 3.4). In contrast, educational video content provides a valuable resource: the interactive element introduced by narrators who often pose questions during their presentations and subsequently provide the answers themselves. For instance, a narrator says “*Do you know what kind of organ we’re dealing with?*” and then proceeds to elaborate by “*Yes, this is a colon.*”. This Q/A format within the videos offers a rich organic Q/A dataset to be extracted and repurposed for evaluation.

To harness this potential, we map the “?”s in the video’s transcript to our stable chunks. If a “?” appears within a 45-second time frame of a stable chunk, we expand the stable chunk’s caption to encompass the complete sentence that includes the question mark. This method ensures that the questions are related to the visual content displayed. Following the data pre-processing and the strategic mapping of question marks, we prompt GPT-4 to extract the question and answer pairs directly given in a text. We input GPT-4 the stable chunk’s text, as well as any sentences from the text that end in a question mark, indicating a posed question. Fig. 11 in the Appendix demonstrates our GPT-4 prompt to generate QUILT-VQA. After the initial extraction by GPT-4, we perform a manual verification, ensuring that each question-answer pair in our dataset is not only medically relevant but also correctly corresponds to the content provided in the stable chunk’s text. See table 7 in the Appendix for statistics of QUILT-VQA. Additionally, we categorize questions into two groups: Image-dependent (1055 Q/A pairs), referencing the narrator’s specific image, and general-knowledge (228 Q/A pairs), rooted in broader medical understanding. This division lets us craft evaluation sets that comprehensively test the model’s image analysis and medical knowledge.

### 4.3. Evaluation data generation: Instruction Following Test Set

In addition to QUILT-VQA, which is focused on assessing QUILT-LLAVA’s medical knowledge, we also aimed to evaluate the model’s capability in following instructions during multi-modal conversations. To that end, we constructed a set of 326 questions, including 256 conversational and 70 detailed description questions, all derived from image-text

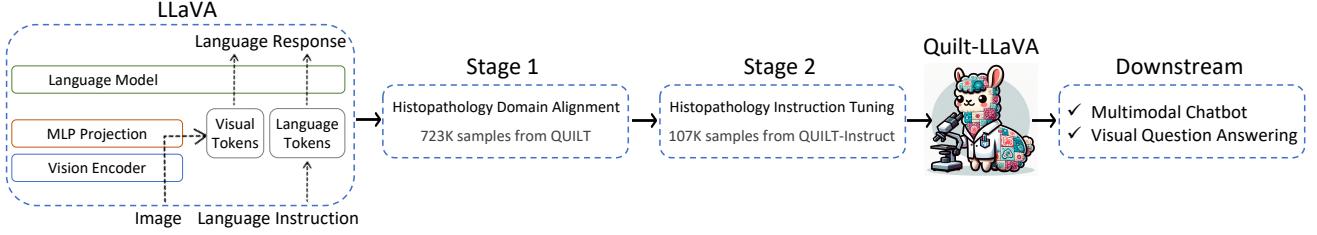


Figure 3. QUILT-LLAVA was initialized with the general-domain LLaVA and trained for two stages: Histopathology Domain Alignment on QUILT and instruction-tuning on QUILT-INSTRUCT. We evaluated QUILT-LLAVA on visual conversation and question answering tasks.

pairs extracted from unseen videos within QUILT-VQA. To generate this evaluation set, we followed the same Conversation and Detailed Description based prompts that we used when generating QUILT-INSTRUCT.

## 5. Experiments

This section covers the performance of QUILT-LLAVA against existing SOTA multi-modal models on histopathology VQA benchmarks using a variety of metrics. First, using GPT-4 we conducted an alignment of output generations with the ground truth answer. Second, we run open and closed set VQA tasks. Finally, we ablate with visual promptings as well as different trained models.

**Oracle (GPT-4) Alignment Evaluation.** To assess the effectiveness of QUILT-LLAVA in multi-modal conversations, using the Instruction Following test-set (see section 4.3), we leverage the GPT-4 (language-only model) to evaluate the *helpfulness, relevance, accuracy, and level of details* of the responses from the two assistants (candidate model and GPT-4), and give an overall score on a scale of 1 to 10 of the answers generated from the two assistants, where a higher score indicates better overall performance, while also providing comprehensive explanation of its evaluation, for us to better understand the models. We then compute the relative score using GPT-4 reference score for normalization [14, 17], as seen in Table 1a across 14 sub-pathologies and 2 QA types. See Fig. 12 in the Appendix for the exact prompt used. All QUILT-LLAVA models outperform baselines, with QUILT-LLAVA trained on single epoch stage-1 and a balanced subset of 40K instruction-tuning pairs for stage-2 fine-tuning outperforming LLaVA and LLaVA-MED by over 16% and 7% respectively, and upon increasing the size of instruction-tuning data and pre-training for three epochs we achieve even better results of 10.8% over LLaVA-MED.

**Visual Question Answering.** We evaluated QUILT-LLAVA on three histopathology VQA datasets, the details of which are provided in Table 7 in the Appendix. These datasets feature a mix of open-ended and close-ended Q/A pairs. For closed-set questions, accuracy is used as the metric to determine the proportion of correct answers given by the model.

In contrast, for open-set questions, we focus on recall to assess how frequently the model’s responses include ground-truth tokens [14]. We compare all versions of QUILT-LLAVA against the medical domain LLaVA-MED, and general domain LLaVA in Table 1b. All variants of QUILT-LLAVA outperform LLaVA. We see the difference in text encoder initialization from Vicuna [3] in the first set of experiments paired with pre-trained open-sourced QUILTNET models doing better on open-set questions with an average of 4% over SOTA and the text encoder initialization from LLaVA doing better on closed-set questions with an average of 9% over SOTA. Performance of QUILT-LLAVA on closed-set questions showcases our model’s instruction following capacity (in both binary *yes/no* and true multi-choice *{A/B/C/D}*) QA unlike LLaVA-MED which underperforms on the latter due to lack of following such multi-choice instructions), whilst outperforming both baselines on PathVQA and QUILT-VQA, albeit performance is limited on PMC-VQA-Subset, we believe it stems from the PMC-VQA’s issues as discussed in Section 3.4 of the Appendix. QUILT-LLAVA outperforms both baselines on Open-set evaluation on PathVQA and QUILT-VQA, with significant margins. We also run ablations with multiple image encoders and training durations in Stages 1 and 2 summarized in Table 2a and qualitative examples in Tables 3-6 in the Appendix.

**VQA with Visual Prompting.** We adopt the visual prompting methodology from [22] for evaluating our model. This involves utilizing the subset of QUILT-VQA with bounding boxes to create ellipses that encapsulate the concepts highlighted by these boxes (See Appendix Fig. 17). Our QUILT-LLAVA surpasses the two baselines in both open and closed-set visually prompted VQA, achieving over 4% in open-set and 10% in closed-set accuracy. We attribute this enhanced performance to our grounded instruction-tuning dataset, underscoring QUILT-LLAVA’s capability to reason within a region of interest, a critical aspect in histopathology.

**Instruction-tuning Ablation Studies.** To study the impact of QUILT-INSTRUCT on performance, we train several models: **I** – Only balanced independent prompts data 3.2.1 model, **R** – Only balanced reasoning prompts data 3.2.2, and **IR** –



## References

- [1] Tad T Brunyé, Agnes Balla, Trafton Drew, Joann G Elmore, Kathleen F Kerr, Hannah Shucard, and Donald L Weaver. From image to diagnosis: Characterizing sources of error in histopathologic interpretation. *Modern Pathology*, 36(7):100162, 2023. [2](#)
- [2] Hyaline Change. Cellular responses to stress and toxic insults: Adaptation, injury, and death. *Robbins and Cotran Pathologic Basis of Disease, Professional Edition E-Book*, page 1, 2009. [2](#)
- [3] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023. [2, 7](#)
- [4] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. [2](#)
- [5] Shah Giashuddin and Mouyed Alawad. Histopathological diagnosis of nonalcoholic steatohepatitis (nash). In *Non-Alcoholic Steatohepatitis: Methods and Protocols*, pages 1–18. Springer, 2022. [2](#)
- [6] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023. [2](#)
- [7] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. [6, 2, 3](#)
- [8] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. [6](#)
- [9] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfoglu, Fatehmeh 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, 3, 4, 6, 8](#)
- [10] Julie A Jacko. Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications. 2012. [3](#)
- [11] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. [2](#)
- [12] Daniel Kahneman. *Attention and effort*. Citeseer, 1973. [3](#)
- [13] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. [2, 3](#)
- [14] 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. *arXiv preprint arXiv:2306.00890*, 2023. [2, 6, 7, 8, 4, 18](#)
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. [5](#)
- [16] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. [2, 8, 4](#)
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. [2, 4, 5, 6, 7](#)
- [18] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. [2](#)
- [19] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023. [2](#)
- [20] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020. [3](#)
- [21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. [3](#)
- [22] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. [2, 7](#)
- [23] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023. [2](#)
- [24] 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. [2](#)
- [25] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471, 2023. [3](#)

- [26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.
- [27] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.
- [28] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. [3](#)
- [29] Heyu Zhang, Yan He, Xiaomin Wu, Peixiang Huang, Wenkang Qin, Fan Wang, Juxiang Ye, Xirui Huang, Yanfang Liao, Hang Chen, et al. Pathnarratives: Data annotation for pathological human-ai collaborative diagnosis. *Frontiers in Medicine*, 9:1070072, 2023. [3](#)
- [30] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreeet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pre-training for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. [4](#)
- [31] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. [2](#), [6](#), [3](#)
- [32] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [2](#)



# Quilt-LLaVA: Visual Instruction Tuning by Extracting Localized Narratives from Open-Source Histopathology Videos

## Supplementary Material

### 1. QUILT-INSTRUCT and QUILT-VQA

#### 1.1. Total Cost

We used GPT-4 to create the QUILT-INSTRUCT dataset, spending \$8858 in total where \$6600 is spent on input tokens and \$2204 on output tokens, and \$54 on evaluations. In detail, Complex Reasoning yielded 25,838 Q/A pairs, with an average of 13.16 words per question and 267 words per answer. Detailed Descriptions produce 39,620 samples, each having 11.91 words per question and 130.2 words per answer. For Conversations, we generated 34,242 rows, with an average of 3.51 Q/A pairs per row, totaling 44.74 words for questions and 221 words for answers. Finally, for Iterative Abductive Reasoning, there are 14,642 rows with multiple Q/A pairs, averaging 2.84 Q/A pairs per row, comprising 91.94 words per question and 201 words per answer. Furthermore, we spent \$140 on generating 1283 samples in QUILT-VQA with an average of 10.43 words per question and 19.16 words per answer. This results in \$8998 total cost.

---

#### Algorithm 1 Stable Video Chunk Detection Algorithm

---

```
1: procedure DETECTSTATICFRAMES(video, starttime, endtime)
2:   video = video[starttime:endtime]
3:   fixedFrames ← ∅
4:   ValidFrames ← ∅
5:   prevFrame ← first frame in video
6:   for frame ∈ rest of frames in video do
7:     absDiff ← absolute difference between frame and prevFrame
8:     absDiffThresh ← apply adaptive thresholding to absDiff
9:     meanVal ← mean value of absDiffThresh
10:    if meanVal < 10 then
11:      fixedFrames ← fixedFrames ∪ frame
12:    else
13:      if length of fixedFrames ≥ minimum duration then
14:        subclip ← extract sub-clip of frames with
15:          constant background from fixedFrames
16:        for randomly sampled patch ∈ each frame of subclip do
17:          SSIMVal ← calculate SSIM of patch
18:          if SSIMVal > threshold then
19:            ValidFrames ← ValidFrames ∪ frame
20:          end if
21:        end for
22:      end if
23:      fixedFrames ← ∅
24:    end if
25:    prevFrame ← frame
26:  end for
27:  staticTimestamps ← get start & end times from ValidFrames
28:  return staticTimestamps
29: end procedure
```

---

#### 1.2. Prompts

Figures 5, 6, 7, 8, and 9 demonstrate the prompts used for generating conversational, detailed description, com-

plex medical reasoning, and iterative abductive reasoning (assistant and student) Q/A pairs in QUILT-INSTRUCT, respectively. In addition, Figure 10 includes the prompt used to extract the diagnosis and supporting facts from single WSI cases. Figure 11 shows the prompt used for generating QUILT-VQA, our evaluation set described in 4.2. And finally, Figure 12 shows the prompt used for evaluating the conversational chat abilities of QUILT-LLAVA using GPT-4.

### 2. Use Case and Limitations

#### 2.1. Use Case of QUILT-LLAVA

Despite our best efforts, QUILT-LLAVA still produces erroneous information, meaning it cannot be fully trusted as an automated diagnostic tool. Instead, it should serve as a support tool for pathologists or other medical professionals who are familiar with the field, enabling them to correct or revise QUILTNET’s responses. Additionally, since QUILT-LLAVA is capable of reasoning beyond its immediate context, it could assist medical practitioners in challenging cases where they might be at an impasse. Furthermore, thanks to its Iterative Abductive Reasoning-based Q/A data, QUILT-LLAVA can function as a brainstorming tool. Rather than providing immediate answers, it can offer hints, which is particularly beneficial for pathology students or junior pathologists looking to learn and develop their skills. An example is given in Fig. 16.

#### 2.2. QUILT-INSTRUCT and QUILT-VQA Limitations

**Limitations of Hand-Crafted Video Processing Algorithms.** When aligning narrators’ speech with images, there can be instances where the narrator discusses something irrelevant or refers to previously observed matters, potentially resulting in unaligned image-text pairs. However, our empirical observations indicate that these occurrences are rare. Also, the images extracted from the video might inherently contain noise, such as artifacts like text, drawings, faces of the narrators’, etc., as illustrated in Fig. 20. Additionally, in some cases, the patient’s medical history, race, gender, etc., are mentioned in the videos, which may or may not be inferred from the aligned image. Moreover, the clustering algorithm relies on simple rules and thresholds, which may not yield the optimal clustering for some cases. Finally, our dataset currently includes only English, limiting the QUILT-LLAVA to operating in just a single language.

## Limitations of Generating QUILT-INSTRUCT Using LLMs.

As we derive visual instruction tuning data from captions using GPT-4, we inevitably inherit its biases and limitations. Additionally, there are instances where GPT-4 does not fully adhere to commands instructing it to generate Q/A datasets as if it were directly observing the image; instead, it sometimes produces responses that reference the description, generating phrases that state "it is mentioned in the description, etc.". Moreover, despite being provided with spatially grounded captions, GPT-4 may occasionally overlook this information, resulting in data that lacks spatial context. Most critically, GPT-4 is still prone to generating inaccurate information, leading to instances of QUILT-LLAVA producing incorrect statements or hallucinations (Please see Failure Cases in Section 4.1). Regarding legal matters, our commitment to making our dataset publicly available requires strict adherence to OpenAI's terms. These terms clearly stipulate that the QUILT-INSTRUCT is to be utilized exclusively for research purposes.

## Limitations of Generating QUILT-VQA Using LLMs.

Limitations of QUILT-VQA also encompass those mentioned for QUILT-INSTRUCT above. In addition, despite padding chunks with question marks to ensure that answers fall within the context, there remain cases where narrators provide answers to these questions far beyond the chunk's timeframe. This can lead to instances where the answers are not adequately captured.

### 2.3. Visual Prompting Details and Limitations

The ellipses are drawn in red , with their size set to 0.05% of the largest dimension of the image, as recommended by [22].

In our visual prompt testing, we used red, which may not stand out against the purple-pink hues common in histopathology images as seen in Figure 17. For better contrast, future work could explore other colors like green or blue, similar to the inks used by pathologists to mark regions of interest in WSIs.

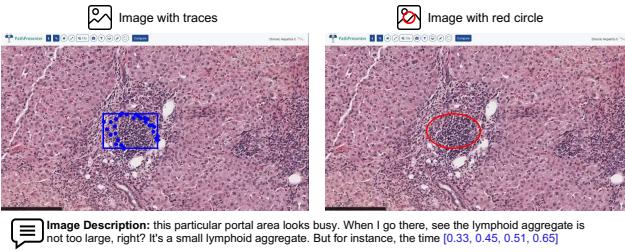


Figure 17. A sample of visual prompting w/ drawing a red circle around the clustered traces.

Category	Details and Example
Wrong answers	PathVQA includes questions with wrong answers. E.g., Q: Did early ischemic injury increase eosinophilia? A: No. (Correct answer is "Yes" according to Figure 1-9 in [2].)
Inconsistent Q/A pairs	PathVQA has similar questions with inconsistent answers. E.g., Q: Does early ischemic injury show surface blebs, increase eosinophilia of cytoplasm, and swelling of occasional cells? A: Yes. (Contradicts with example in "Wrong answers".)
Ambiguous questions	PathVQA includes ambiguously phrased questions. The question below starts with "What is showing" and the answer provides a diagnosis. The question should have been more clear as "what is the diagnosis of" or "what has the following symptoms?" E.g., Q: What is showing increased eosinophilia of cytoplasm? A: early (reversible) ischemic injury.
Repetitive Q/A pairs	Both PathVQA and PMC-VQA generate multiple Q/A pairs for a single image-text pair, leading to repetition. E.g., two different questions in PathVQA about early ischemic injury. (Refer to examples in "Wrong answers" and "Inconsistent Q/A pairs".)
General-domain Q/A pairs	Most VQA datasets focus on general-domain biomedical questions. E.g., PMC-VQA is generated from PubMed articles and we extracted a histopathology subset from it for evaluation purposes. In addition, PathVQA includes many non-histopathology images.

Table 8. Summary and examples of issues in public VQA datasets.

## 3. Visual Question Answering

### Visual Question Answering Datasets for Histopathology

#### 3.1. PathVQA

[7] is a dataset of 32799 question-answer pairs extracted from 4998 pathology image-caption pairs within textbooks and digital libraries. The questions are categorized as open-ended and closed-ended questions where the former consists of what, where, when, whose, how, how much/how many questions and the latter includes yes/no questions. We use 6761 samples from the evaluation set.

#### 3.2. PMC-VQA

[31] version-2 includes a test set of 34823 VQA pairs of non-compound images that cover various modalities or diseases. This dataset was curated from image-caption pairs in PMC-OA articles and are in multiple-choice format. We retrieved PMC-VQA-Subset including 2318 histopathology VQA pairs from this dataset.

**Image**

**Description**

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture. [0.31, 0.27, 0.46, 0.71]

**Diagnosis**

Invasive Ductal Carcinoma

**Supporting Facts**

- 1. Irregular, infiltrative blue areas in the breast tissue.
- 2. Presence of irregular, angulated glands infiltrating into the fibrous and fatty stroma.
- 3. Comparison of benign lobule and malignant invasive breast gland showing irregularity and distortion in latter.
- 4. Nuclear pleomorphism with variation in size and shape
- 5. Presence of prominent nucleoli.
- 6. Frequent mitotic figures indicating rapid growth.

[...continues]

**Conversation**

**User:** Can you describe the main features visible in this histopathology image?

**Assistant:** The image primarily showcases a lobule, which is a small rounded glandular formation typically found in breast tissue. This lobule exhibits a classic benign structure with two distinct layers. The outer layer is composed of myoepithelial cells, identifiable by a single row of outer nuclei. [...continues]

**User:** Where is this lobular architecture located within the image?

**Assistant:** The lobular architecture is prominently displayed in the central region of the image, extending towards the upper right quadrant. The bilayered structure of the lobule, with its myoepithelial and luminal cells, is clearly visible in this area.  
[...conversation continues]

**Detailed Description**

**User:** In a few words, what does the histopathology image depict?

**Assistant:** In the central to lower right portion of the histopathology image, there is a clear depiction of a lobule, a characteristic structure in benign breast parenchyma. This lobule is composed of small, rounded glandular formations, demonstrating a classic bilayered appearance. The outer layer of the lobule is made up of myoepithelial cells, which can be identified by a single row of nuclei. Adjacent to this, there is an inner layer composed of luminal cells, also identifiable by a single row of nuclei. The space between these luminal cells forms the lumen. This bilayered structure, with its distinct outer myoepithelial layer and inner luminal cell layer surrounding the lumen, is a hallmark of benign breast tissue. The lobular architecture is well-preserved and clearly visible in this region of the image, further supporting the benign nature of the tissue.

**Complex Medical Reasoning**

**User:** Can we make a diagnosis based on this single patch of histopathology image?

**Assistant:** The patch under examination reveals a lobule with a typical bilayered structure. This structure consists of an outer myoepithelial layer and an inner luminal cell layer, surrounding a central lumen. The presence of these two distinct layers of cells and the well-preserved lobular architecture are characteristic features of benign breast parenchyma. The myoepithelial cells, represented by the outer row of nuclei, play a crucial role in maintaining the integrity of the breast tissue. They act as a natural barrier against potential invasion of cancer cells into the surrounding stroma. The luminal cells, represented by the inner row of nuclei, are the ones that usually undergo malignant transformation in most breast cancers. However, in this patch, the lobule appears to maintain its normal architecture without any signs of distortion or irregularity. There are no visible signs of nuclear pleomorphism, which would indicate variation in size and shape of the nuclei, a common feature in malignancies. Also, there are no evident mitotic figures, which would suggest rapid cell division and growth, another hallmark of cancer. Based on these observations, this patch seems to represent benign breast tissue. However, it's important to note that this conclusion is based solely on this single patch. Breast cancer, such as invasive ductal carcinoma, often presents as irregular, infiltrative areas in the breast tissue, with angulated glands infiltrating into the fibrous and fatty stroma. [...continues]

**Iterative Abductive Reasoning**

**User:** I see a lobule with small rounded glandular formations, exhibiting a bilayered appearance with an outer myoepithelial layer and an inner luminal cell layer. Could this be indicative of benign breast parenchyma?

**Assistant:** Your observation of the bilayered appearance of the lobule is correct and it is indeed indicative of benign breast parenchyma. However, it's important to remember that the presence of benign structures doesn't rule out the possibility of malignancy elsewhere. Hint: While you've correctly identified benign structures, consider looking for signs of irregularity or distortion in the glandular formations. [...continues]

**User:** I notice some irregularity and distortion in the glandular formations. There also seems to be nuclear pleomorphism and prominent nucleoli. Could this suggest the presence of a malignant process, possibly breast carcinoma?

**Assistant:** Your observations are indeed suggestive of a malignant process. The presence of nuclear pleomorphism and prominent nucleoli are key features of malignancy. [...continues]

**User:** I see the glands infiltrating into the surrounding stroma and frequent mitotic figures. Could this suggest an invasive ductal carcinoma of the breast?

**Assistant:** CORRECT! Your deductions are spot on. [...continues]

Figure 18. A sample from QUILT-INSTRUCT. Only Complex Medical Reasoning and Iterative Abductive Reasoning prompts implicitly utilize Diagnosis and Supporting Facts (which are extracted from the entire video). Note that some of the content is truncated due to space limitations. Please see Figs. 13-14-15 for a more complete overview of QUILT-INSTRUCT.

### 3.3. Quilt-VQA

consists of 1961 VQA pairs, which we curated from naturally-occurring Q/A by the narrators of the videos in QUILT. Please see Sec. 4.2 for details. Figure 21 shows some sample Q/A pairs.

	PathVQA	PMC-Subset	QUILT-VQA
# Images	858	2209	985
# QA Pairs	6761	2318	1283
# Open	3370	-	940
# Closed	3391	2318	343

Table 7. VQA datasets statistics. For PathVQA, we report the evaluation set statistics. PMC-Subset is the histopathology subset retrieved from PMC-VQA.

### 3.4. Visual Question Answering Dataset Issues

Current established VQA datasets have limitations and issues. Table 8 summarizes a few of these issues and provides examples from PathVQA [7] and PMC-VQA [31] datasets.

## 4. Evaluation

### 4.1. Qualitative Evaluation

#### Success Cases

Tables 3, 4, 5, and 6 demonstrate successful examples of QUILT-LLAVA chat abilities compared to LLAVA and LLAVA-MED.

#### Failure Cases

Tables 9, 10, 11, and 12 demonstrate successful examples of QUILT-LLAVA chat abilities compared to LLAVA and LLAVA-MED across a few Q/A types (e.g observations, staining, localization etc.).

### 4.2. Omission of GPT-4V in evaluation

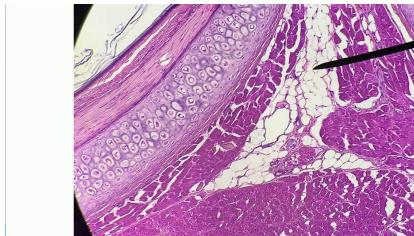
Due to the nature of the test datasets being crawled from videos (e.g QUILT-VQA) and articles (e.g. PMC-VQA-Subset) the images most often contain confounding artifacts which can often be indicative of diagnosis, or image features in text, we find that GPT-4V, unlike LLAVA models, are better capable at extracting tiny textual confounding artifacts that describe the image, when such telling artifacts are within the image and would otherwise fail when the artifacts are removed, See Table 13, based on this observation we exclude

Instruct	QUILT-LLAVA Model Variants		PathVQA		PMC-Subset		QUILT-VQA		QUILT-VQA	
	Stage 1	Stage 2	Open	Closed	Closed	Open	Closed	Open (w/o ● w/)	Closed (w/o ● w/)	
<i>QUILTNET ViT-B-32 Vision Encoder [9], 7B Language Model</i>										
107K	1	0	14.34	53.78	27.05	47.69	56.56	49.62 ● 54.13	55.56 ● 58.33	
107K	1	1	14.24	58.42	19.63	59.82	64.43	58.81 ● 61.08	68.52 ● 70.37	
107K	1	3	12.79	56.30	17.21	57.62	63.55	56.21 ● 58.32	65.74 ● 69.44	
107K	3	1	15.30	54.93	16.01	60.97	60.64	59.24 ● 64.06	56.48 ● 59.26	
<i>QUILTNET ViT-B-16 Vision Encoder [9], 7B Language Model</i>										
107K	1	1	15.14	58.24	15.83	59.61	61.52	57.46 ● 61.04	63.89 ● 63.89	
<i>CLIP Vision Encoder (ViT-L-14 @336px) [30], 7B Language Model</i>										
107K	0	1	15.95	59.89	15.84	58.43	69.10	56.19 ● 59.44	70.37 ● 62.04	
107K	0	3	15.29	61.81	17.29	55.61	68.81	54.28 ● 57.44	71.30 ● 65.74	
107K	1	1	15.19	58.54	16.18	59.01	62.10	57.01 ● 57.98	61.11 ● 57.41	
<i>LLAVA [16] checkpoint, 7B Language Model</i>										
107K	0	1	14.93	56.97	27.83	55.36	69.97	54.64 ● 59.88	70.37 ● 73.15	
107K	0	3	13.39	53.20	32.56	53.94	62.97	52.44 ● 55.19	57.41 ● 77.78	
107K	1	0	11.65	54.03	33.91	55.80	58.02	54.77 ● 59.56	51.85 ● 60.19	
107K	1	1	15.06	58.68	28.56	55.39	68.81	54.24 ● 59.83	71.30 ● 75.00	
<i>Baselines</i>										
LLaVA-Med [14] 7B	0	0	11.97	56.15	1.34	54.81	61.22	52.58 ● 53.97	69.44 ● 64.81	
LLaVA [16] 7B	0	0	11.65	54.02	33.91	55.81	57.73	54.74 ● 59.96	51.85 ● 60.19	

(a)

Table 2. Additional quantitative results on three established histopathology VQA datasets. For open-set questions, we report the recall for our free-form text generation method in column *Open*. For closed-set questions, we report the accuracy in column *Closed*. Red indicates the best-performing model.

#### Visual input example, Histopathology image:



User	What is the connective tissue visible in the image?
QUILT-LLAVA	The connective tissue in the image is identified as the perichondrium. It's a layer of dense irregular connective tissue that surrounds the cartilage. The perichondrium is visible in the upper left quadrant of the image.
LLaVA-Med	The connective tissue visible in the image is lamellar keratinized epithelium. This type of tissue is composed of multiple layers of keratinized cells, which are cells that produce keratin, a tough, fibrous protein that provides protection and structural support.
LLaVA	The connective tissue visible in the image is a piece of human muscle tissue.
Answer	The connective tissue visible in the image is cartilage.

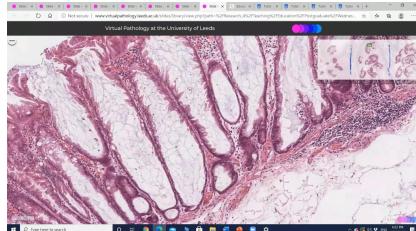
Table 3. **Success case** : Example comparison of histopathology visual chat and reasoning capabilities. Note: QUILT-LLAVA correctly understands the histology features and the localization of said features.

GPT-4V from our evaluations and leave its evaluation to future work.

---

**Visual input example, Histopathology image:**


---



User	What is visible in the middle of the image?
QUILT-LLAVA	In the middle of the image, there is a significant presence of mucin. Mucin is a protein that forms mucus, and its presence in large amounts can be indicative of certain conditions.
LLaVA-Med	In the middle of the image, you can see a histopathological section of the human prostate. Histopathology refers to the microscopic examination of tissue to study the manifestations of disease.
LLaVA	In the middle of the image, there is a close-up view of a human organ, specifically the stomach.
Answer	Mucus is visible in the middle of the image.

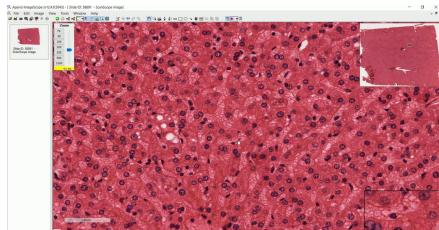
---

Table 4. **Success case** : Example comparison of histopathology visual chat and reasoning capabilities. Note: QUILT-LLAVA can correctly answer spatial questions.

---

**Visual input example, Histopathology image:**


---



User	What type of cells are the majority in the liver as seen in the image?
QUILT-LLAVA	The image shows that the majority of cells in the liver are hepatocytes, which are the primary cell type in the liver. These cells are located in the central to the right middle region of the image.
LLaVA-Med	The majority of the cells in the liver as seen in the image are foamy macrophages.
LLaVA	The majority of cells in the liver as seen in the image are red blood cells.
Answer	The majority of the cells in the liver, as seen in the image, are hepatocytes.

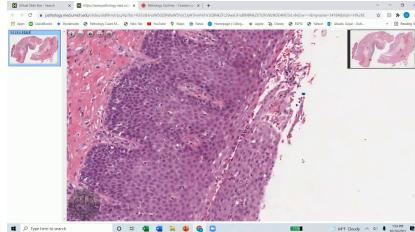
---

Table 5. **Success case** : Example comparison of histopathology visual chat and reasoning capabilities. Note: QUILT-LLAVA correctly interprets the histologic features in the image and their localization of features.

---

**Visual input example, Histopathology image:**

---



User	What are the three criteria visible in the image that suggest eosinophilic esophagitis?
QUILT-LLAVA	The image shows three key features that are indicative of eosinophilic esophagitis. Firstly, there is a noticeable increase in the number of eosinophils, which are a type of white blood cell. Secondly, there is a significant amount of eosinophilic infiltration, which refers to the accumulation of these cells in the tissue. Lastly, there is evidence of basal cell hyperplasia, which is an increase in the number of cells in the basal layer of the tissue. These three features together strongly suggest a diagnosis of eosinophilic esophagitis.
LLaVA-Med	The three criteria visible in the image that suggest eosinophilic esophagitis are: 1. Increased intraepithelial eosinophils (H&E, x40) 2. Eosinophilic infiltrates in the lamina propria (H&E, x40) 3. Presence of eosinophils in the muscularis mucosa (H&E, x40) These criteria are based on the histological features of the tissue sample, which can help in the diagnosis of eosinophilic esophagitis.
LLaVA	In the image, there are three criteria visible that suggest eosinophilic esophagitis (EoE). First, the esophageal tissue is inflamed, as indicated by the red and pink coloration. Second, there are numerous eosinophils present in the tissue, which are white blood cells that are typically elevated in EoE cases. Finally, the tissue appears to be thickened, which is another characteristic of EoE. These three criteria together suggest that the image depicts a case of eosinophilic esophagitis.
Answer	The three criteria visible in the image that suggest eosinophilic esophagitis are the presence of eosinophils, basal cell hyperplasia, and spongiosis.

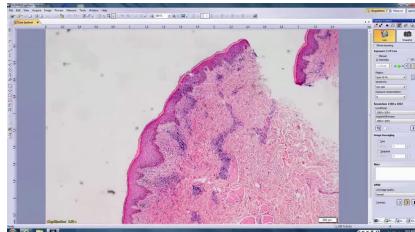
---

Table 6. **Success case** : Example comparison of histopathology visual chat and reasoning capabilities. Note: QUILT-LLAVA correctly extracts the morphology features to answer the question.

---

**Visual input example, Histopathology image:**

---



User	What can be observed in the epidermis in the image?
QUILT-LLAVA	The epidermis in the image appears to be quite normal. It's located towards the upper left quadrant of the image, extending towards the center. The epidermis is the outermost layer of the skin and is responsible for providing a barrier against environmental damage.
LLaVA-Med	In the epidermis of the control group, you can observe the presence of a stratum corneum, which is the outermost layer of the skin. Additionally, there is a keratin layer, which is a protein that provides structural support and protection to the skin.
LLaVA	In the image, the epidermis is observed to be affected by a disease, possibly cancer, as it appears to be infected or affected by a growth.
Answer	Spongiosis and erythrocytes can be observed in the epidermis in the image.

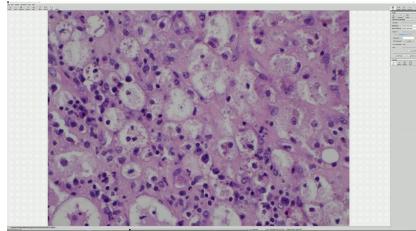
---

Table 9. **Failure case** : Example comparison of histopathology visual chat and reasoning capabilities.

---

**Visual input example, Histopathology image:**

---



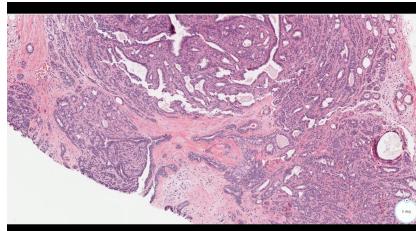
User	What stain can be used to highlight the histiocytes in the image?
QUILT-LLAVA	The histiocytes in this image can be highlighted using a Ziehl-Neelsen stain. This stain is particularly useful for identifying acid-fast bacilli, which are often associated with certain types of infections.
LLaVA-Med	The image uses hematoxylin and eosin (H&E) staining, which is a common staining technique used to highlight histiocytes. Histiocytes are a type of immune cell that plays a role in the body's defense against pathogens and foreign substances.
LLaVA	In the image, the histiocytes can be highlighted using a stain called hematoxylin. This stain is commonly used in histology to stain various cell types, including hematopoietic cells, and it helps to differentiate them from other cell types in the tissue. The presence of hematoxylin in the image indicates that the histiocytes are being stained, making them more visible and easier to identify.
Answer	The S100 stain can be used to highlight the histiocytes in the image.

Table 10. **Failure case** : Example comparison of histopathology visual chat and reasoning capabilities.

---

**Visual input example, Histopathology image:**

---



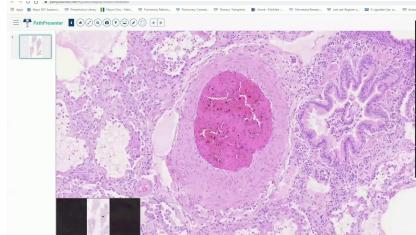
User	How would you describe the stroma visible in the image?
QUILT-LLAVA	The stroma in the image appears to be quite cellular, indicating a high density of cells. This is a significant observation as it can be a sign of certain pathological conditions.
LLaVA-Med	The stroma in the image appears to be moderately cellular. The stroma is the supportive tissue surrounding the tumor cells, and its cellularity refers to the density of the cells within the stroma. In this case, the stroma is described as moderately cellular, which means it has a moderate amount of cells compared to a highly cellular or poorly cellular stroma.
LLaVA	The stroma in the image is described as being "pink" and "red," which suggests that it is a mixture of pink and red tissue.
Answer	The stroma visible in the image is described as sclerotic.

Table 11. **Failure case** : Example comparison of histopathology visual chat and reasoning capabilities.

---

**Visual input example, Histopathology image:**

---



User	How would you characterize the structures lining the alveolar spaces in the image?
QUILT-LLAVA	The structures lining the alveolar spaces in the image are quite distinctive. They are not typical of what you would expect to see in a healthy lung tissue. Instead, they resemble a certain type of growth or formation, which is not commonly seen in normal lung histology.
LLaVA-Med	The structures lining the alveolar spaces in the image are characterized as delicate and flattened.
LLaVA	The structures lining the alveolar spaces in the image are described as "pink" and "purple." These colors suggest that the structures are likely composed of blood vessels, which are responsible for delivering oxygen to the body's cells. The presence of these blood vessels in the image indicates that the tissue is actively involved in the process of respiration and oxygen exchange.
Answer	The structures lining the alveolar spaces in the image are isoendophilic hyaline membranes. They are classic features of a diffuse alveolar damage process.

---

Table 12. **Failure case** : Example comparison of histopathology visual chat and reasoning capabilities.

**GPT-4 Prompt: Conversational instruction-following data generation**

**System Message:** You're an AI assistant specialized in histopathology image interpretation. Given descriptions and bounding box coordinates [x1, y1, x2, y2] (normalized between 0-1) related to histopathology images, create a dialogue as if you're directly observing and analyzing the image.

Guidelines:

Observation: Respond as if directly viewing the image.

Spatial Reference: Interpret the spatial layout from the bounding box details but avoid mentioning the coordinates.

Tone: Maintain a conversational style, emulating a visual AI assistant's perspective. At the end of the final answer, indicate that you are an AI assistant and not a real medical doctor.

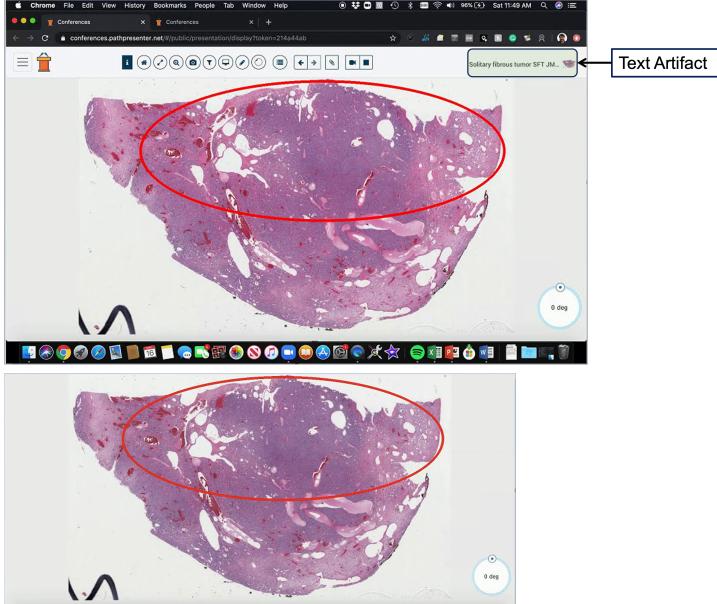
Aim for a coherent dialogue, approximately 3-4 Q&A pairs, not exceeding 500 words. Treat each Q&A pair independently; do not assume context from previous pairs.

Figure 5. The GPT-4 prompt used to generate conversational instruction-following data.

---

**GPT-4V example, Histopathology image:**


---



User	What is the unique vascular pattern seen in the image? Tip: Focus on image areas highlighted by the red circle(s)
QUILT-LLAVA	The image reveals a unique vascular pattern that is quite distinctive. The red circles, which are highlighted, seem to be pointing towards a specific area of interest. This area is characterized by a dense network of blood vessels, which is a common feature in certain types of tumors. The vessels appear to be irregularly shaped and densely packed, which is not typically seen in normal tissue. This vascular pattern could be indicative of a highly vascularized tumor, which is often associated with aggressive growth and poor prognosis.
GPT-4V w/o Cropping	The image you've provided appears to be a histological slide of tissue, most likely viewed under a microscope and stained for better visualization of the structures. The red circle highlights an area of interest within the tissue. The unique vascular pattern within the red circle seems to be characteristic of a "staghorn" or "hemangiopericytoma-like" vascular pattern. This pattern is often associated with solitary fibrous tumors (SFTs), which are rare tumors that can occur in various parts of the body. The vessels in this pattern tend to be elongated and branched, resembling the antlers of a stag, hence the name "staghorn." Solitary fibrous tumors are typically composed of spindle-shaped cells and have this distinctive vascular pattern. They can be benign or malignant and may require further clinical correlation and immunohistochemical staining to confirm the diagnosis and plan appropriate treatment. Please note that the interpretation of histological images should be performed by a trained pathologist, and this assessment is based on the visual information provided in the image.
GPT-4V w/ Cropping	The image appears to be a microscopic slide of a tissue sample, which has been stained for better visualization of the structures within the tissue. The red circle drawn on the image seems to highlight a particular area of interest, possibly for educational or diagnostic purposes. The tissue exhibits various shades of purple and pink, indicating different types of cells and extracellular components. The presence of white spaces could represent adipose (fat) tissue, voids where cells are not present, or artifacts from the slide preparation process. The label "0 deg" in the corner suggests that the image may be part of a series taken at different rotations or angles. Without more context, it's difficult to determine the exact type of tissue or the purpose of the analysis.
Answer	The unique vascular pattern seen in the image is characteristic of solitary fibrous tumors, previously known as hemangiopericytomas. This pattern includes dilated, branching, tortuous vessels.

Table 13. Example evaluation GPT-4V: The textual confounding artifact highlighted with green box: *solitary fibrous tumor SFT JM..* is picked up by GPT-4 and forms the basis of its description w/o cropping, however, when the text is cropped out of the image, it is clear that GPT-4V lacks histology image understanding as it simply describes the colors and other non-confounding artifacts missing the histology features within the image e.g the vessel structures.

**GPT-4 Prompt: Detailed Description instruction-following data generation**

**System Message:** You are a specialized AI in histopathology image interpretation. When provided with descriptions of histopathology images and, occasionally, bounding box coordinates [x1, y1, x2, y2] (normalized between 0 and 1), offer a detailed description as if you're directly observing the image.

Instructions:

- 1) Always describe as if you are directly looking at the histopathology image, without any reference to the text description provided.
- 2) If bounding box coordinates are provided, use them to guide the spatial layout of your description without explicitly mentioning the coordinates.
- 3) If no bounding boxes are provided, and no spatial explanations in the captions are found, avoid making spatial inferences.
- 4) Provide explanations in a formal manner suitable for someone with a medical education.
- 5) Never pose questions; only provide explanations.

Figure 6. The GPT-4 prompt used to generate detailed description instruction-following data.

**GPT-4 Prompt: Complex Medical Reasoning instruction-following data generation**

**System Message:** You're an AI assistant specialized in histopathology image interpretation. Given descriptions related to histopathology images, create dialogue as if you're directly observing and analyzing the image.

Imagine you are directly looking at a single histopathology patch from a whole slide image and you are provided with a report from another doctor who analyzed the same patient's whole slide image who has made the diagnosis and provided their clues from different patches from the whole slide image. So you will be provided with:

Diagnosis: [xxx]

Clues from Whole Slide: [xxx]

Single patch: [xxx]

Your task is to construct a question-answer dialogue, referring to yourself as "GPT" and the inquirer as "User." The user may ask questions like, "What can be deduced from this single image regarding a possible diagnosis?"

Guidelines for GPT's Response:

- 1) Maintain the perspective that you are unaware of the "Clues from Whole Slide" and the "Diagnosis" already established. Your analysis and reasoning should revolve solely around the information presented in the "Single patch" description.
- 2) Embark on the diagnostic reasoning process by discussing the specific features observed in the 'Single patch.' These may include cellular structures, any evident abnormalities, or other noteworthy characteristics. Relate how these elements typically correlate with certain pathological conditions.
- 3) If the single patch's details suggest a possible diagnosis based on common medical knowledge, you may hypothesize a probable condition. However, emphasize that this is based solely on the given patch and that a broader examination of the whole slide is standard procedure for confirmation.
- 4) In cases where the 'Single patch' information is inconclusive or insufficient for a diagnosis, state the necessity of further examination of additional patches from the whole slide. Do this without alluding to any specific 'Clues from Whole Slide' or 'Diagnosis' previously provided. Indicate that this comprehensive method is vital to identify consistent pathological hallmarks or observe any contrasting features.
- 5) Throughout your response, avoid directly referencing or inferring content from the 'Clues from Whole Slide' or the initial 'Diagnosis.' Any reasoning or conclusions should appear as though they are derived independently, utilizing only the data within the 'Single patch' and general medical knowledge.
- 6) Always formulate your answers like you are directly observing the image and not the text description. Never say things like "as mentioned in the description...", "from the description of the single patch we can see that...", or "the mention of X" etc. Always say things like "The presence of X", "Given these observations" etc.

Figure 7. The GPT-4 prompt used to generate complex medical reasoning instruction-following data.

**GPT-4 Prompt: Iterative Abductive Reasoning instruction-following data generation (Assistant)**

**System Message:** You are the AI histopathology expert, guiding a student through complex patient case evaluations for diagnostic purposes. Your role involves iterative discussions with the student, who only has access to a single patch image from a whole slide image at a time. You, however, have an overview of observations from various patches of the patient's whole slide image and know the final diagnosis. The information is presented to you in the format:

Diagnosis: [xxx]

Observations from different patches: [xxx]

Student's Image: [xxx]

Instructions:

Assessing Abductions:

Review the student's abductions and the factual basis they provide. Acknowledge that they are analyzing just one patch image at a time. First, evaluate that if a diagnosis can be made solely on what the student sees or they need extra evidence from different images (which you possess with "Observations from different patches:").

Determine the accuracy and completeness of the student's abductions based on their 'Facts Used.' This information will be presented as User:[{Abduction: xxx},{Facts Used: xxx}].

If the student's conclusions are fully correct, affirm with "CORRECT!!!" and conclude the dialogue by saying "End of Guidance". Otherwise, evaluate if a more accurate abduction could be derived from their observations. Provide targeted, insightful hints to redirect their focus within the same patch or suggest examining other patches. Avoid explicit diagnosis revelations; instead, guide through suggestive questioning or hinting at overlooked details.

Guide them towards validation strategies if they've exhausted the current patch's potential, suggesting, "Consider looking for evidence of X in other patches."

Providing Hints:

Frame your hints and feedback as if you've directly observed and memorized the images, maintaining the illusion of a first-hand, real-time analysis. Do not disclose any additional observations; the student must work with their current patch.

Refrain from statements like "you overlooked in the observations...", or "your abductions do not align with all the observations". Always know that the User cannot see those observations. Instead, nudge them towards correct inferences by suggesting what to focus on in future patch analyses.

Offer concise, constructive hints that deepen their understanding and encourage accurate deductions.

Engage in a focused dialogue that stimulates critical thinking and effective synthesis of the given information.

Respond using the format: GPT:[{Comments: xxx},{Hint: xxx}], and conclude your guidance with "End of Guidance" when you ascertain the student has gleaned all possible insights from the current image and has to move on to see another patch from the same whole slide image to make the correct diagnosis.

Ensure your responses do not exceed 170 tokens, maintaining efficiency in communication.

Your role is pivotal in enhancing the student's diagnostic acumen through this simulated, interactive learning experience.

Embody the mentor persona, leveraging your 'visual' insights to foster a challenging yet educational dialogue.

Give your answer in the format of GPT:[{Comments: xxx},{Hint: xxx}] for which you are the GPT.

After your first answer, the AI assistant will respond to you with User:[{Abduction: xxx},{Facts Used: xxx}], which will be appended to your first answer and so on. So you will see the conversation history between you (GPT) and the student (User).

Figure 8. The GPT-4 prompt used to generate iterative abductive reasoning (assistant) instruction-following data.

### GPT-4 Prompt: Iterative Abductive Reasoning instruction-following data generation (Student)

**System Message:** You are a human histopathology expert working alongside an AI expert to analyze patient cases. You'll receive text descriptions of patch images from patient's whole slide images. These descriptions belong to a histopathology image patches so interpret them as if you're viewing the actual images. Use your histopathology knowledge to make abductions about conditions or features in the given image. In this unique setup, you'll engage in a collaborative case analysis with AI expert, simulating professional diagnostic deliberations.

\* Instructions:

1) Making Abductions:

\*Transform the text 'images' into medical abductions. Mention what you 'see' and the resulting diagnoses. If uncertain, ask for more information.

\*Act as if you're directly observing and describing the images. Avoid phrases like "The description mentions..."

2) Interaction Structure:

\*Respond in this format: User:[{Abduction: xxx},{Facts Used: xxx}], summarizing your observations and diagnoses. Narrate 'live', using "I see..." or "The tissue shows..." .

\*Discuss your reasoning with the AI, anticipating feedback on your abductions. The AI will confirm correct responses or guide you with hints.

3) Dialogue Flow:

\*Engage with the Expert Histology AI after making abductions, acknowledging its hints with responses like, "That makes sense!"

If guided to other patches (signaled by "End of Guidance"), conclude with your final abduction summary and "End of Conversation."

\* Also when making your Abduction, try to pose it as a question "Could this suggest an ongoing inflammatory response? Possibly due to an infection or autoimmune condition" like as if you are asking to someone who knows this better than you.

4) Conciseness:

\*Limit abductions to 90 tokens max. Focus on clarity and brevity.

5) Format

\*Give your answer in the format of User:[{Abduction: xxx},{Facts Used: xxx}] for which you are the User.

\*After your first answer, the AI assistant will respond to you with GPT:[{Comments: xxx},{Hint: xxx}], which will be appended to your first answer and so on. So you will see the conversation history.

Remember, you're a firsthand observer. Your 'viewing' of the descriptions should mimic live, real-time analysis, crucial for an authentic interactive learning experience. Following is an example abduction

Given image: translate into that the angular spaces are filled with the macrophages so let's discuss about the histology so this is the first slide which is on the low power so in this slide we can see that first is the lesion is a diffuse process and the second is that the all the lesion it is it is in the one stage of the process.

You will response as:

User:[{'Abduction: I see a diffuse lesion that appears to be in one stage of a process. The angular spaces are filled with macrophages. Could this suggest an ongoing inflammatory response, possibly due to an infection or autoimmune condition?'},{Facts Used: The presence of a diffuse lesion in one stage of a process and the angular spaces filled with macrophages.'}]

Remember, GPT will reply to you like

GPT:[{'Comments: Your observation about the presence of macrophages is correct, as indicated by the CD68 stain. However, the disease you're considering might not necessarily be infectious. Remember, the pancreatin stain is positive on the alveolar septal cells but negative inside the air spaces.'},{'Hint: Consider the significance of the pancreatin stain being negative inside the air spaces. What could this suggest about the cells present there? Also, think about diseases that might involve macrophages but are not necessarily infectious. For example there are non-infectious types of Pneumonia.'}]

Then you will reply again making a new Abduction

Figure 9. The GPT-4 prompt used to generate iterative abductive reasoning (student) instruction-following data.

**GPT-4 Prompt: Extract Diagnosis and Supporting Facts of a histopathology case**

**System Message:** Drawing upon a vast medical knowledge base and as if you were a senior pathologist at the Mayo Clinic, evaluate the provided note meticulously. Specifically, answer the following:

Does the note contain any medical keywords or abbreviations commonly used in clinical settings?

Based on the content of the note, is there a discernible medical diagnosis?

Provide a simple 'yes' or 'no' answer for each question and NOTHING ELSE. Also know that there cannot be a case where diagnosis is True and medical content is False.

Figure 10. The GPT-4 prompt used to extract the diagnosis and supporting facts of a histopathology case.

**GPT-4 Prompt: Evaluation Question-Answer pair generation (Quilt-VQA)**

**System Message:** You are a senior pathologist at the Mayo Clinic. You are given a text where a pathologist is describing a histopathology image. You are given the sentences with a question mark from the given text. Generate well-defined question/answer pairs from these sentences.

Consider the following requirements to generate questions and answers:

- Only generate questions about information being seen in the image.
- Do not generate questions from the context if the question is not explicitly asked.
- Do not generate questions/answers based on information that can not be seen in the image being described like the patient's age, gender, medical history, or other studies/images outside of the current case.
- If the text does not explicitly provide the answer to a question, do not generate a question/answer pair.
- Do not answer the questions from information outside of the given text.
- Do not use phrases like "mentioned", "suggests" or "text" in the answers. Instead, refer to the information as being seen "in the image".
- Do not reveal answers in the questions.

Figure 11. The GPT-4 prompt used to generate question-answer pairs for evaluation set (QUILT-VQA).

**GPT-4 Prompt: Evaluating multimodal chatbot performance**

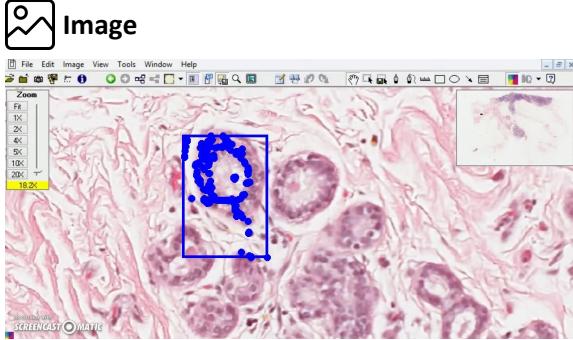
**System Message:** We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing an image. For your reference, the visual content in the image is represented with caption, describing the same image, which is embedded with bounding box coordinates of each object in the scene, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.

Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space.

In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Figure 12. The GPT-4 prompt used to evaluate the multi-modal model performance.



## Description

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]



## Conversation

**User:** Can you describe the main features visible in this histopathology image?

**Assistant:** The image primarily showcases a lobule, which is a small rounded glandular formation typically found in breast tissue. This lobule exhibits a classic benign structure with two distinct layers. The outer layer is composed of myoepithelial cells, identifiable by a single row of outer nuclei. The inner layer, on the other hand, consists of luminal cells, marked by another row of nuclei. The space between these cells forms the lumen. This bilayered appearance is a characteristic feature of benign breast parenchyma.

**User:** Where is this lobular architecture located within the image?

**Assistant:** The lobular architecture is prominently displayed in the central region of the image, extending towards the upper right quadrant. The bilayered structure of the lobule, with its myoepithelial and luminal cells, is clearly visible in this area.

**User:** What does this bilayered appearance indicate about the condition of the breast tissue?

**Assistant:** The bilayered appearance, with distinct myoepithelial and luminal cell layers, is indicative of benign breast parenchyma. This means that the breast tissue appears to be healthy and non-cancerous. However, it's important to note that a definitive diagnosis would require a comprehensive clinical evaluation and possibly additional tests.

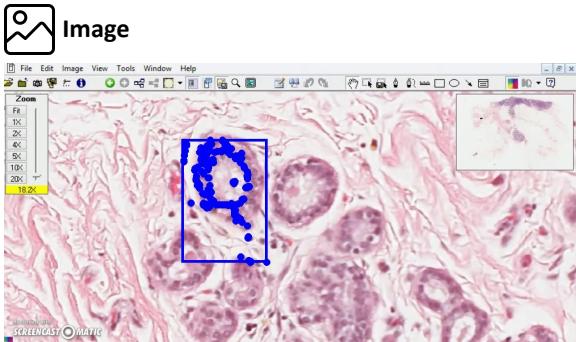


## Detailed Description

**User:** In a few words, what does the histopathology image depict?

**Assistant:** In the central to lower right portion of the histopathology image, there is a clear depiction of a lobule, a characteristic structure in benign breast parenchyma. This lobule is composed of small, rounded glandular formations, demonstrating a classic bilayered appearance. The outer layer of the lobule is made up of myoepithelial cells, which can be identified by a single row of nuclei. Adjacent to this, there is an inner layer composed of luminal cells, also identifiable by a single row of nuclei. The space between these luminal cells forms the lumen. This bilayered structure, with its distinct outer myoepithelial layer and inner luminal cell layer surrounding the lumen, is a hallmark of benign breast tissue. The lobular architecture is well-preserved and clearly visible in this region of the image, further supporting the benign nature of the tissue.

Figure 13. A full example of Conversation and Detailed Description types of Q/A in QUILT-INSTRUCT.



### Description

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]



### Diagnosis

Invasive Ductal Carcinoma



### Supporting Facts

1. Irregular, infiltrative blue areas in the breast tissue.
2. Presence of irregular, angulated glands infiltrating into the fibrous and fatty stroma.
3. Comparison of benign lobule and malignant invasive breast gland showing irregularity and distortion in latter.
4. Nuclear pleomorphism with variation in size and shape
5. Presence of prominent nucleoli.
6. Frequent mitotic figures indicating rapid growth.
7. Increased nuclear cytoplasmic ratios.
8. Poorly formed, irregular glands on low power view.
9. Infiltrative growth pattern where the glands invade into the surrounding stroma.



### Complex Medical Reasoning

**User:** Can we make a diagnosis based on this single patch of histopathology image?

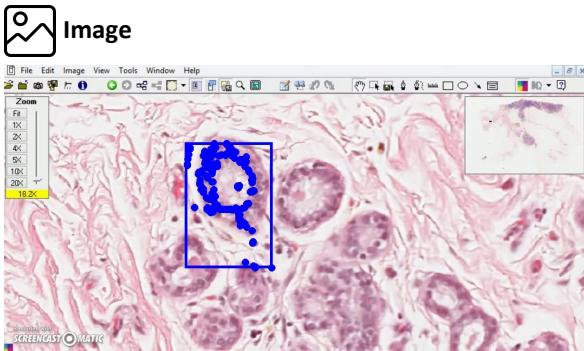
**Assistant:** The patch under examination reveals a lobule with a typical bilayered structure. This structure consists of an outer myoepithelial layer and an inner luminal cell layer, surrounding a central lumen. The presence of these two distinct layers of cells and the well-preserved lobular architecture are characteristic features of benign breast parenchyma.

The myoepithelial cells, represented by the outer row of nuclei, play a crucial role in maintaining the integrity of the breast tissue. They act as a natural barrier against potential invasion of cancer cells into the surrounding stroma. The luminal cells, represented by the inner row of nuclei, are the ones that usually undergo malignant transformation in most breast cancers.

However, in this patch, the lobule appears to maintain its normal architecture without any signs of distortion or irregularity. There are no visible signs of nuclear pleomorphism, which would indicate variation in size and shape of the nuclei, a common feature in malignancies. Also, there are no evident mitotic figures, which would suggest rapid cell division and growth, another hallmark of cancer. Based on these observations, this patch seems to represent benign breast tissue. However, it's important to note that this conclusion is based solely on this single patch. Breast cancer, such as invasive ductal carcinoma, often presents as irregular, infiltrative areas in the breast tissue, with angulated glands infiltrating into the fibrous and fatty stroma. These features are not evident in this patch, but they could be present in other areas of the tissue.

Therefore, while this patch appears benign, a definitive diagnosis would require a comprehensive examination of the whole slide image. This would allow for the identification of any areas of irregularity or distortion, increased nuclear cytoplasmic ratios, or poorly formed glands that might suggest an invasive growth pattern. Additionally, correlation with clinical information and possibly additional immunohistochemical studies would be necessary to confirm the diagnosis.

Figure 14. A full example of Complex Medical Reasoning type of Q/A in QUILT-INSTRUCT.



### Description

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]



### Diagnosis

Invasive Ductal Carcinoma



### Supporting Facts

1. Irregular, infiltrative blue areas in the breast tissue.
2. Presence of irregular, angulated glands infiltrating into the fibrous and fatty stroma.
3. Comparison of benign lobule and malignant invasive breast gland showing irregularity and distortion in latter.
4. Nuclear pleomorphism with variation in size and shape
5. Presence of prominent nucleoli.
6. Frequent mitotic figures indicating rapid growth.
7. Increased nuclear cytoplasmic ratios.
8. Poorly formed, irregular glands on low power view.
9. Infiltrative growth pattern where the glands invade into the surrounding stroma.



### Iterative Abductive Reasoning

**User:** I see a lobule with small rounded glandular formations, exhibiting a bilayered appearance with an outer myoepithelial layer and an inner luminal cell layer. Could this be indicative of benign breast parenchyma?

**Assistant:** Your observation of the bilayered appearance of the lobule is correct and it is indeed indicative of benign breast parenchyma. However, it's important to remember that the presence of benign structures doesn't rule out the possibility of malignancy elsewhere.

**Hint:** While you've correctly identified benign structures, consider looking for signs of irregularity or distortion in the glandular formations. Also, pay attention to the nuclear characteristics. Are there any signs of nuclear pleomorphism or prominent nucleoli? These could be indicative of a more sinister process.

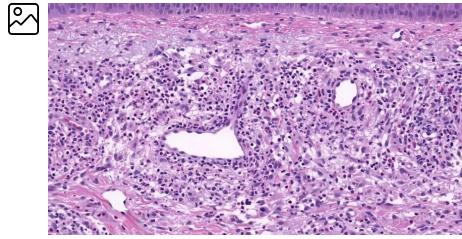
**User:** I notice some irregularity and distortion in the glandular formations. There also seems to be nuclear pleomorphism and prominent nucleoli. Could this suggest the presence of a malignant process, possibly breast carcinoma?

**Assistant:** Your observations are indeed suggestive of a malignant process. The presence of nuclear pleomorphism and prominent nucleoli are key features of malignancy. However, the specific type of breast carcinoma might still be unclear. **Hint:** Consider the growth pattern of the glands. Are they infiltrating into the surrounding stroma? Also, look for signs of rapid growth, such as frequent mitotic figures. These features could help you narrow down the type of breast carcinoma.

**User:** I see the glands infiltrating into the surrounding stroma and frequent mitotic figures. Could this suggest an invasive ductal carcinoma of the breast?

**Assistant:** CORRECT!!! Your observations and deductions are spot on. The presence of glands infiltrating into the surrounding stroma and frequent mitotic figures are indeed indicative of an invasive process, specifically invasive ductal carcinoma of the breast. **Hint:** None needed. You've made the correct diagnosis based on the evidence available.

Figure 15. A full example of Iterative Abductive Reasoning type of Q/A in QUILT-INSTRUCT.



User: Can you provide a concise description of the histopathology image shown?

Quilt-LLaVA: Upon examination of the histopathology image, it is evident that there is a significant presence of lymphocytes. These cells are densely packed and exhibit a high degree of pleomorphism, indicating a possible pathological condition. The lymphocytes are irregular in shape and size, with some displaying enlarged nuclei and prominent nucleoli. In addition to the lymphocytes, there is also a noticeable presence of eosinophils. These cells are characterized by their bright pink cytoplasm and bilobed nuclei. The eosinophils are scattered throughout the tissue, suggesting an ongoing immune response. The overall cellular architecture appears disrupted, with the normal tissue architecture replaced by a dense infiltrate of lymphocytes and eosinophils. This could be indicative of an inflammatory or immune-mediated process. Further diagnostic tests may be required to confirm the exact nature of this pathological condition.

User: The presence of scattered neutrophils throughout the tissue could suggest an acute inflammatory response. Could this be due to a skin condition such as a bug bite or a ruptured cyst?

Quilt-LLaVA: Your observation about the presence of neutrophils is correct, indicating an acute inflammatory response. However, the context of this inflammation is crucial. It's not just any skin condition, but a specific one. Hint: Consider the location of this inflammation. What conditions might cause an acute inflammatory response in this specific area? Also, think about the significance of the neutrophils being scattered throughout the tissue. What does this suggest about the nature of the inflammation?

Figure 16. An example illustrating where the QUILT-LLAVA is used as an educational companion, providing hints to human users to help train them as pathologists.

- o "Can you provide a concise description of the histopathology image shown?"
- o "How would you summarize the features of the histopathology image displayed?"
- o "What are the key observations in the presented histopathology image?"
- o "Briefly outline the characteristics of the histopathology image in view."
- o "In a few words, what does the histopathology image depict?"
- o "What essential information emerges from the histopathology image provided?"
- o "Could you distill the contents of the displayed histopathology image into a brief description?"
- o "Offer a succinct report on the histopathological aspects evident in the image."
- o "Identify and briefly discuss the components of the histopathology image."
- o "How would you encapsulate the findings presented in the histopathology image?"
- o "What insights can be concisely derived from the histopathology image on display?"

Figure 19. A list of instructions to variably describe the visual information present in the images in detail used to generate detailed description instruction-following QA pairs.

- o "Describe the histology image concisely."
- o "Provide a brief description of the given microscopic tissue image."
- o "Offer a succinct explanation of the pathology picture presented."
- o "Summarize the visual concepts of the histology image."
- o "Give a short and clear explanation of the subsequent cellular image."
- o "Share a concise interpretation of the histology image provided."
- o "Present a compact description of the pathology whole slide's key features."
- o "Relay a brief, clear account of the pathology picture shown."
- o "Render a clear and concise summary of the pathology whole slide photo."
- o "Write a terse but informative summary of the histopathology image."
- o "Create a compact narrative representing the histology image presented."

Figure 18. A list of instructions to variably describe the visual information present in the images used in the histopathology domain alignment. The list is taken from [14] and adapted for the histopathology content.

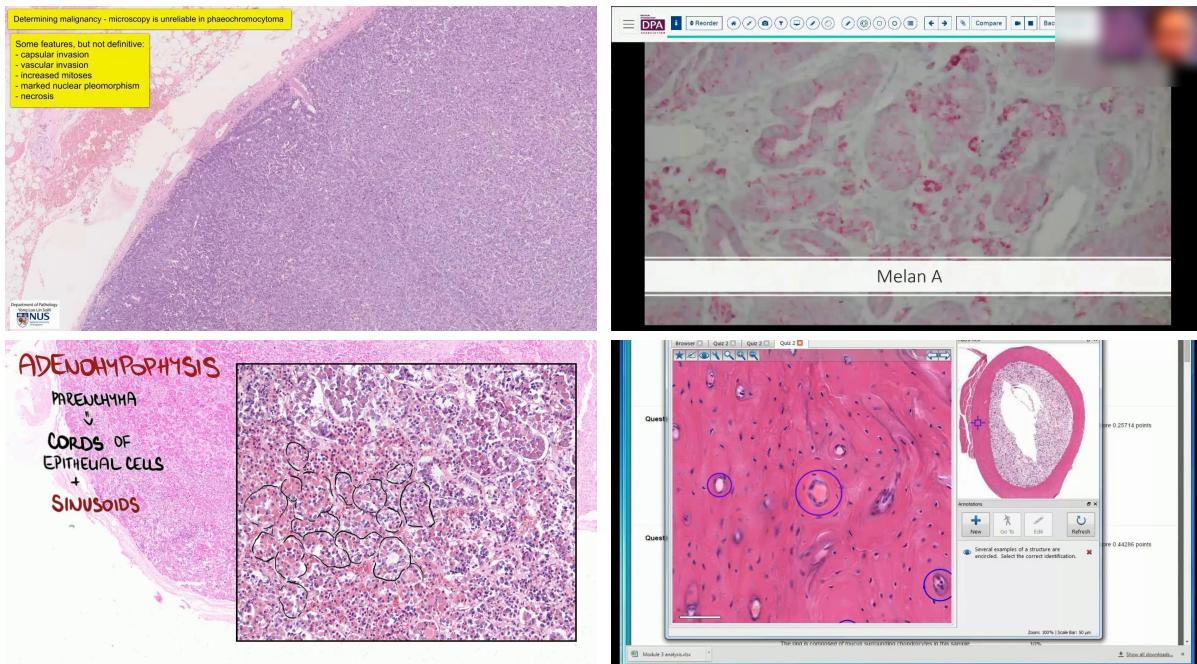


Figure 20. Some examples of noisy images include those where the median frame might contain artifacts such as text, drawings, or the narrator's face (seen in the top right figure). Although we completely mask the narrator's face in our process, for brevity, we have included an example here with a blurred face.)

	Image	Text	Question	Answer
Image Dependent		So they look relatively large even at low power. But at higher magnification, you can see that they're actually, there's some lymphocytes, but a lot of are histiocytes. Yes. And notice the arrangement of the cells. <b>How are they kind of arranged here?</b> Yeah, so they're arranged circumferentially around this more like lighter pale pink material. Yes, good.	How are the cells arranged in the image?	The cells are arranged circumferentially around the lighter pale pink material.
General Knowledge		Let's go on. There is a lot of morphology, I think, variable morphology. So what you see here. One area looks solid on cystic. And here. And there is lining epithelium. Yes, <b>what kind of epithelium?</b> I see, oh my God, this is mucoepidermoid. I think it is mucoepidermoid carcinoma. All right, so again we will see this is mucous cell.	What kind of epithelium is visible in the image?	The image shows mucoepidermoid epithelium.
		But once you start seeing fibrosis in the sinusoidal area, then it's perisinusoidal. <b>What significance it has?</b> So pericellular fibrosis is something that you see with steatohepatitis, meaning like you have a balloon hepatocyte and that eventually, you know, steatohepatitis eventually will lead to fibrosis and cirrhosis. So that's the starting point.	What is the significance of seeing fibrosis in the sinusoidal area?	Pericellular fibrosis in the sinusoidal area is something that you see with steatohepatitis. This indicates that there is a balloon hepatocyte that eventually will lead to fibrosis and cirrhosis. This is the starting point of the disease process.
		Low power, you can appreciate that there's a compact hyperkeratosis, both ortho and para. You can see here there's parakeratosis as well. And then you can see that there's acanthosis as well. And there seems to be some kind of abnormal maturation of the epithelium. It looks abnormally pink. We'll go on higher power. You immediately see that there's lichen sclerosis. It can be subtle. But this one has established lichen sclerosis. And then if you look at the epidermis itself, you begin to appreciate that there seems to be some basal cell atypia. <b>What is basal cell atypia?</b> Well, the ISSVD has some specific definitions that one should apply. So atypia is basically something that has any two of the following features: pleomorphism, enlargement, sort of two to three times the size of a lymphocyte. Most of the time, three times the size of a lymphocyte.	What is basal cell atypia?	Basal cell atypia refers to the abnormality that has any two of the following features: pleomorphism, enlargement, typically two to three times the size of a lymphocyte, most of the time, three times the size of a lymphocyte.

Figure 21. Some examples of QUILT-VQA, our evaluation dataset. Top two rows show image-dependent Q/A pairs and bottom two rows show general-knowledge Q/A pairs. The original question posed by the narrator of the video is highlighted in yellow.