




---

# Quilt-1M: One Million Image-Text Pairs for Histopathology

---

Wisdom O. Ikezogwo\* Mehmet S. Seyfioglu Fatemeh Ghezloo  
Dylan Geva Fatwir S. Mohammed Pavan K. Anand  
Ranjay Krishna Linda G. Shapiro  
University of Washington  
{wisdomik,msaygin,fghezloo,dgeva,pka4,ranjay,shapiro}@cs.washington.edu  
fatwir@uw.edu

## Supplementary material

We present the following items in the supplementary material section:

1. Data curation models, algorithms and parsing pipelines (Section A)
2. Exploratory analysis of the collected data (Section B)
3. Pretraining and downstream evaluation details (Section C)
4. Exploration of trained model representations (Section D)
5. A Datasheet [13] for our QUILT dataset (Section E)

## A Data curation models, algorithms and parsing pipelines


### A.1 Curating QUILT: an Overview

Creating a densely annotated vision-language dataset from videos is a significant undertaking, as it involves various handcrafted algorithms and machine learning models. In the following sections, we present more detailed information about the challenges of the data curation pipeline and algorithms used to address these challenges. To download QUILT-1M and its metadata and access the code to recreate the dataset and trained models, refer to our website.

**Collecting representative channels and videos.** The first challenge lies in obtaining relevant histopathology videos. We used a set of keywords (obtained from online histopathology glossaries<sup>2</sup>) to search for videos, resulting in  $\approx 65\text{K}$  potential matches. Figure 1 shows the word cloud of all keywords used for searching YouTube. However, filtering histopathology content based on thumbnail and title yields many false positives, often including general pathology videos. To address this, we process the frames of lower-resolution versions of each video to differentiate between histopathology and pathology content, narrowing the selection to  $\approx 9\text{K}$  videos.

**Filtering for narrative-style medical videos.** Among the  $\approx 9\text{K}$  videos, we sought videos with a "narrative style" where narrators freely explain whole slide images and streaks of similar frames occur, indicating an educational performance. To identify such content, we used a model that analyzed randomly sampled frames to determine if they maintained a consistent style over time. This process resulted in the selection of  $\approx 4\text{K}$  videos. Non-voiced videos are also filtered by using `inaSpeechSegmenter` [10] where the video endpoint does not provide the video language or transcript. To identify the audio language of a video, we first check YouTube's API. If the information is

---

\*Reach corresponding author at wisdomik@cs.washington.edu; : Equal contribution.

<sup>2</sup><https://lab-ally.com/histopathology-resources/histopathology-glossary>

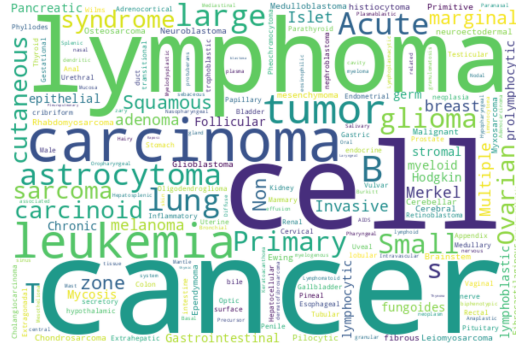


Figure 1: Word cloud of all keywords used for searching YouTube

unavailable through the API, we use OpenAI’s Whisper model [31] on the first minute of audio from the video.

To identify videos containing medical content, we employ a keyframe extraction process with a specific threshold to determine the minimum visual change required to trigger keyframes. For a new video, the thresholds for keyframe extraction are determined by linearly interpolating between the lowest threshold, 0.008 (5-minute video) and the highest 0.25 (200-minute video). Following the keyframe extraction process, we utilize a histopathology image classifier to identify histopathology content within the extracted keyframes. See A.3 for more details. To identify narrative-style videos, we randomly select a  $\min(\text{num\_of\_histo\_scene\_frames}, 20)$  keyframes from a video and utilize a pre-trained CLIP<sup>3</sup> (ViT-B-32) model to embed and compute a cosine similarity on the next three keyframes. If all three have similarity scores  $\geq$  a threshold of 0.9, we count the video as a narrative streak.

Table 1: Salvagable and Non-salvagable cases for ASR correction using an LLM.

Error due to	Raw output	Salvagable (because LLM can rephrase and/or extract contextually similar correction)	Non-salvagable (because the error losses all possible medical context and can lead to wrong entries)
Unfinetuned ASR	...look like the cranialomas I would expect in HP. They actually look more sarcooidal to me. The reason I say that is they, there’s a kind of <b>positive</b> of inflammatory cells associated with them. They’re really tight and well-formed. They’re very easy to see a low power. And so HP is in the <b>differential hypersensium nitose</b> , but I would be more worried about sarcoidosis.	<b>differential hypersensium nitose: hypersensitivity pneumonitis,</b> <b>cranialomas: granulomas</b>	<b>positive: paucity</b>
LLM	<b>high-larbidia-stinal lymphadenocathy</b> <b>lymphin-giatic pattern distribution</b>	returns <b>hilar lymphadenopathy</b> instead of a more appropriate <b>hilar mediastinal lymphadenopathy</b>	returns <b>lymphatic pattern distribution</b> instead of a more appropriate <b>lymphangitic pattern distribution</b>
Incomplete UMLS checker	... <b>picnotic</b>	-	LLM correctly returns <b>pyknotic</b> however, UMLS(2020) does not have the word <b>pyknotic</b> if fails to pass the UMLS check.

**Text extraction using ASR and text denoising.** Another challenge involves automatic speech recognition (ASR), as YouTube captions are often inadequate for medical vocabulary. To address this issue, we employed the Large-V2 open-source Whisper model [31] for speech-to-text conversion. However, general-purpose ASR models like Whisper can misinterpret medical terms, particularly when the speaker’s voice is choppy or accented. There are no straightforward trivial solutions due to: **1)** the absence of openly available medical ASR models or data for fine-tuning in the medical

<sup>3</sup><https://huggingface.co/sentence-transformers/clip-ViT-B-32>

domain; **2)** the inadequacy of medical named entity recognition models in detecting transcription errors, because these models are typically trained on correctly spelled words; **3)** the ineffectiveness of methods like semantically searching over a medical glossary, such as UMLS, which only prove effective when the erroneous text has significant similarity to the correct terms; and **4)** the inability of simpler methods like finding the longest common substring, which might work in finding a match in the glossary/ontology for replacement, but cannot identify the wrong words/phrases in the first place. To rectify ASR errors, we employed UMLS (a knowledge database) and a LLM (GPT-3.5). This, however, introduces a new challenge of identifying incorrectly transcribed words and determining which words were mistakenly "corrected" and correctly formatted by the LLM after error correction and resolving unintended parsing errors [1]. See Figure 3 in the main paper for LLM prompt examples of ASR correction and medical and ROI text extraction from the corrected ASR text. Refer to Table 1 for error examples of ASR correction using the LLM.



Figure 2: Representative Frame Identification. If a Stable frame is found by Algorithm 1 within the candidate regions, we use it as the representative frame. If not, we use the most dissimilar frames among unstable frames.

**Image frame extraction and denoising.** The image processing aspect of this task adds to its complexity, as it requires static frame detection, quality control for frames, and histology magnification classification. Each model utilized in these steps introduces its own biases and errors. We extract time-intervals (*chunks*) from each video from which we extract representative image(s). For each of the extracted *chunks* ( $t_n, t_{n+1}$ ), the static chunk detection algorithm 1 is used to extract sub-time-intervals with static frames within the chunk. If found, we save the median (in pixel space to prevent blurry outputs) of the stable frames, else (i.e no stable duration of frames) we leverage the structural similarity index (SSIM) method on histopathology key-frames to find the most dissimilar histopathology image to make up the representative images for the chunk, essentially de-duplicating the frames. Figure 2 demonstrates this process.

---

**Algorithm 1** Static Video Chunk Detection Algorithm
 

---

```

1: procedure DETECTSTATICFRAMES(video, starttime, endtime)
2:   video = video[starttime:endtime]
3:   fixedFrames  $\leftarrow \emptyset$ 
4:   SSIMValidatedFrames  $\leftarrow \emptyset$ 
5:   prevFrame  $\leftarrow$  first frame in video
6:   for frame  $\in$  rest of frames in video do
7:     absDiff  $\leftarrow$  absolute difference between frame and prevFrame
8:     absDiffThresh  $\leftarrow$  apply adaptive thresholding using a Gaussian filter to absDiff
9:     meanVal  $\leftarrow$  mean value of absDiffThresh
10:    if meanVal < 10 then
11:      fixedFrames  $\leftarrow$  fixedFrames  $\cup$  frame
12:    else
13:      if length of fixedFrames  $\geq$  minimum duration then
14:        subclip  $\leftarrow$  extract sub-clip of frames with constant background from fixedFrames
15:        for patch  $\in$  randomly selected patches in each frame of subclip do
16:          SSIMVal  $\leftarrow$  calculate SSIM of patch
17:          if SSIMVal > threshold then
18:            SSIMValidatedFrames  $\leftarrow$  SSIMValidatedFrames  $\cup$  frame
19:          end if
20:        end for
21:      end if
22:      fixedFrames  $\leftarrow \emptyset$ 
23:    end if
24:    prevFrame  $\leftarrow$  frame
25:  end for
26:  staticTimestamps  $\leftarrow$  extract start and end times from SSIMValidatedFrames
27:  return staticTimestamps
28: end procedure

```

---

**Aligning both modalities.** The alignment of the images with their corresponding text requires the implementation of unique algorithms. These algorithms are designed to reduce duplicate content and ensure accurate mappings between image and text. See Figures 3 and 4 and Table 2 for a demonstration of image-text alignment process. See Figure 5 for sample images and their corresponding medical and ROI texts and the sub-pathology classification provided by the LLM.

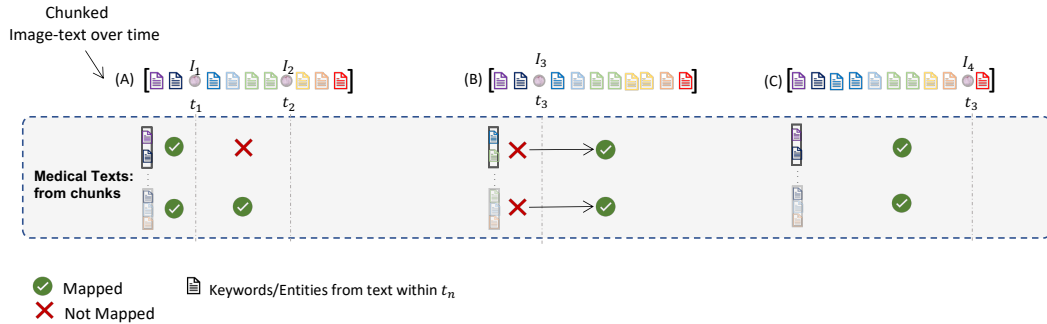


Figure 3: **Overview of use of timing and keywords for Alignment** Images within a video chunk, i.e. {A, B, C},  $I_n$  at  $t_n$  are aligned with medical texts extracted within the same chunk. The *raw\_keywords* within each example chunk is colour coded to illustrate matches with *keywords* extracted from the medical texts and only matching keywords allow for the pairing of texts containing said *keywords* to image frames with frame-times around *raw\_keywords* times.

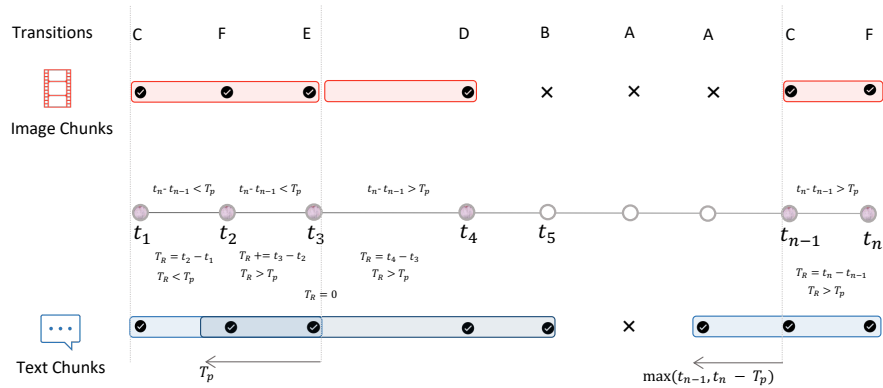


Figure 4: **Video Chunking algorithm illustrate.** With each transition tag explained in Table 2, we leverage predicted histopathology frames at times  $t_1, \dots, t_n$  to segment videos into chunks. Chunks at are minimum are  $T_P$  in duration, this value is estimated based on the word-per-second of the video with a minimum of 20 words being captured per chunk. Images within a chunk, unlike texts, are not overlapping with other chunks. Text overlap is done to provide needed context for LLM text correction and extraction.

Table 2: **All 6 (six) transition states for chunking narrative style videos.**  $p(H)_{t_n}$  is the binary histo image classifier prediction at the current frame’s time  $t_n$  and  $p(H)_{t_{n-1}}$  is the prediction at next frame’s time  $t_{n-1}$ , where  $T_R$  is the cumulative running time and  $T_P$  is the estimated minimum chunk time for the video, determined by the words per second of the video. Text and image chunks are implemented as an ordered list of time intervals and image indexes.

$P(H)@t_n$	$P(H)@t_{n-1}$	$t_n - t_{n-1} > T_P$	$T_r > T_P$	Text chunk	Image chunk	Tag
0	0	-	-	-	-	A
0	1	-	-	$end = t_n$ ; append( $s, e$ ); reset	append index to chunk state, if state is empty append prior index; reset state	B
1	0	-	-	$start = \max(t_{n-1}, t_n - T_P)$	append index to chunk state	C
1	1	1	-	$end = t_n$ ; append( $s, e$ ); reset state; $start = t_n - T_P$	append index to chunk state; reset state	D
1	1	0	1	$end = t_n$ ; append( $s, e$ ); reset state; $start = t_n - T_P$	append index to chunk state; reset state	E
			0	-	append index to chunk state	F

## A.2 Other data sources

### A.2.1 PubMed Open Access Articles

We searched the PubMed open-access from 2010 – 2022 with keywords (pathology, histopathology, whole-slide image, H&E, and 148 keywords from a histopathology glossary<sup>4</sup>). We utilized Entrez<sup>5</sup> to retrieved the top 10,000 most relevant articles for each keyword. This query yielded 109,518 unique articles with PMCIDs. We extracted 162,307 images and their corresponding captions. Using our histopathology classifier and cropping multi-plane figures as described in A.4, we extracted 59,371 histopathology image and caption pairs with an average caption length of 54.02 tokens. Figure 6 demonstrates the pipeline of collecting data from PubMed.

### A.2.2 Histopathology Image Retrieval from LAION

The Large-scale Artificial Intelligence Open Network (LAION-5B) [33] curated over 5 billion pairs of images and text from across the Internet, including a substantial volume of histopathology-related

<sup>4</sup><https://lab-ally.com/histopathology-resources/histopathology-glossary>

<sup>5</sup><http://www.ncbi.nlm.nih.gov/Entrez>

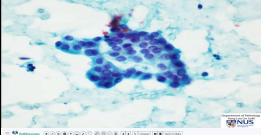
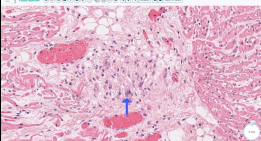
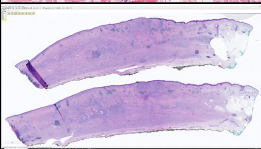
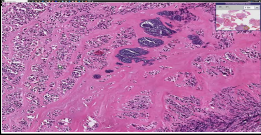
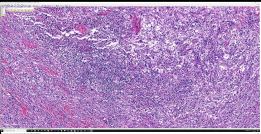
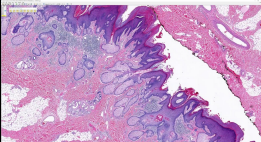
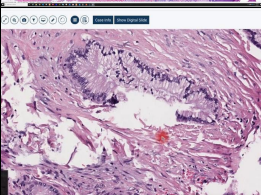
Image	Medical TEXT	ROI Text	Sub-pathology Classification
	['There are clusters of cells with micro-follicular formations.', 'Nuclear pseudo-inclusions, oval nuclei, nuclear grooves, and small nucleoli are present in some cells.']	['clusters of cells', 'micro-follicular formations', 'nuclear pseudo-inclusions', 'oval nuclei', 'nuclear grooves', 'small nucleoli']	['Endocrine', 'Cytopathology', 'Head and Neck']
	['Cluster of macrophages and T cells is characteristic of acute rheumatic fever.', 'Aschoff body is a characteristic feature of acute rheumatic fever.', 'Macrophages with elongated chromatin are called Anitschkow cells and are commonly seen in Aschoff bodies.', 'Pancarditis with Aschoff bodies is present.']	['Cluster of macrophages and T cells', 'Aschoff body', 'Macrophages with elongated chromatin', 'Anitschkow cells', 'Pancarditis']	['Cardiac', 'Hematopathology', 'Endocrine']
	['An 80-year-old man has a scar-like plaque on the scalp that has been called malignant on a biopsy.', 'The tissue affected by the plaque extends from the epidermis to the galea aponeurotica, near the periosteum of the skull.', 'The skin, dermis, and subcutis are all affected by the process.']	['scar-like plaque on the scalp', 'malignant on a biopsy', 'skin, dermis, and subcutis affected by the process']	['Dermatopathology', 'Soft tissue', 'Hematopathology']
	['Inflammatory cells surrounding cartilage can indicate acute chondritis, with neutrophils being the principal cell type.', 'Chronic chondritis may be diagnosed if lymphocytes are the predominant inflammatory cell type.']	['cartilage', 'inflammatory cells']	['Hematopathology', 'Bone', 'Dermatopathology']
	['Large histiocytes with abundant cytoplasm identified as Rosai-Dorfman histiocytes.', 'S100 stain showed perivascular cuffing.', 'Initial diagnosis of inflammatory pseudotumor of the orbit.', 'Rosai-Dorfman disease may burn out and leave behind fibrotic pockets.']	['Large histiocytes', 'perivascular cuffing', 'fibrotic pockets']	['Dermatopathology', 'Soft tissue', 'Hematopathology']
	['Epidermal acanthosis and papillomatosis resembling a wart or seborrheic keratosis.', 'Presence of large sebaceous glands that drain directly through their duct out to the skin surface, which is abnormal.', 'Presence of a demodex mite.']	['Epidermal acanthosis and papillomatosis', 'large sebaceous glands', 'demodex mite']	['Dermatopathology', 'Soft tissue', 'Hematopathology']
	['Histological description of glandular tissue with little atypia but located in a place where it does not belong can be a helpful criteria to discern the presence of malignancy.', 'Glands located on the periphery and infiltrating into adventitia and peripancreatic tissue may be malignant.']	['glandular tissue', 'pancreas,']	['Gastrointestinal', 'Pancreatic', 'Hematopathology']

Figure 5: A collection of sample images from our dataset, accompanied by corresponding medical text, ROI text, and the top three sub-pathology classifications derived from the ASR text using the LLM.

data. We tapped into this resource by retrieving the 3000 most similar LAION samples for each of the 1,000 pairs of images and text sampled from PubMed and QUILT, using a CLIP model pre-trained on the LAION data. The retrieval process utilized both image and text embeddings, with cosine similarity serving as the distance metric. Subsequently, we eliminated the duplicate images and removed all non-English pairs from the remaining pairs using LangDetect<sup>6</sup>. Consequently, the process yielded 22,682 image and text pairs.

### A.2.3 Twitter Data from OpenPath

We utilized a list of tweets curated by Huang et al. [14] which totaled up to 55,000 unique tweets and 133,511 unique image-text pairs. This exhibits a one-to-many relationship that leans towards the image side, differentiating our work from the OpenPath approach, where we had one image matching with multiple captions (as in the case of MS-COCO captions). In order to maintain comparability with OpenPath, we followed their text pre-processing pipeline given in [14].

<sup>6</sup><https://github.com/fedelopez77/langdetect>

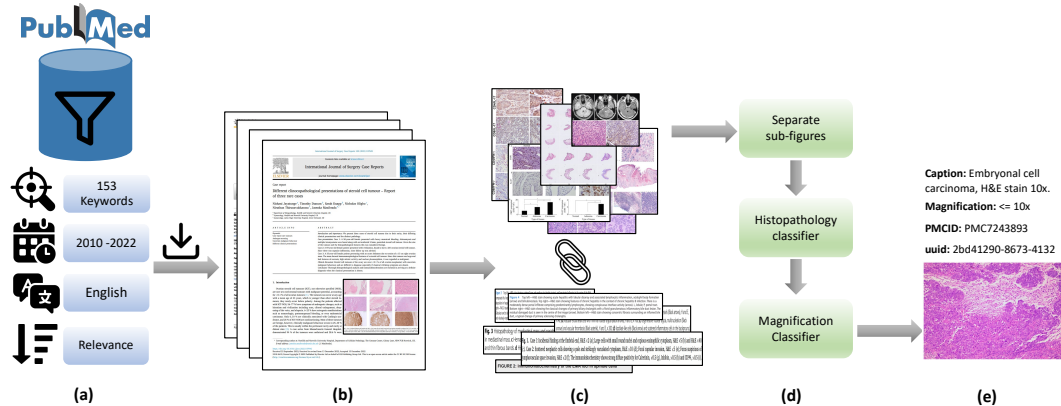


Figure 6: (a) Search PubMed open access database, filter based on keywords, date, language and sort by relevance. (b) Download paper and media for each search result. (c) Extract and pair figures and captions. (d) Separate multi-plane figures, find histopathology images and their magnification. (e) Final result.

### A.3 Histopathology and Magnification classifier

We use an ensemble of three histopathology image classifiers. To ensure robustness, our ensemble approach consists of two small Conv-NeXt models [26] and one linear classifier fine-tuned with DINO features [9]. This combination is necessary due to the homogenous appearance of histopathology images and the risk of false positives from similar pinkish-purple images. One Conv-NeXt model is trained in detecting non-H&E Immunohistochemistry (IHC) stained tissue images, while the other models are trained to handle all IHC stains and tissue types. The training data includes eight sub-groups of the TCGA WSI dataset and a mix of general-domain images, PowerPoint (slide) images, and scientific figure datasets. See Table 3 for details of these datasets.

For the magnification classifier, we finetune a pretrained ConvNeXt-Tiny model [26], with standard preset hyperparameters for a few epochs and select the best performing model on the validation set. To generate a training set for the magnification model, TCGA subsets were segmented into patches using a method similar to [41]. These patches were generated at various magnifications, which were then categorized into three labels: 0: {1.25x, 2.5x, 5x, 10x}, 1: {20x}, 2: {40x}. The TCGA subsets were chosen to ensure a diverse representation of tissue morphologies and cancer types, thereby ensuring robust and comprehensive model training. The model was also trained on cytopathology microscopy images and various IHC stains beyond H&E to enhance the model’s generalizability across different conditions. Only the ACROBAT and TCGA datasubsets are preprocessed to divide the WSIs into patches at various scales.

### A.4 Support Models, Ontology Databases and Algorithms

This section describes the support models, ontology databases and handcrafted algorithms utilized within our pipeline for both searching and parsing our data.

**Ontology databases.** We employ various ontologies, both specific to histopathology and general ones. Among them are OCHV [2], FMA [29], BCGO<sup>7</sup>, NCIT [11], MPATH [32], HPATH [40], and CMPO [18]. These ontologies serve a dual purpose. First, we used histopathology-specific ontologies (HPATH, MPATH, BCGO, and CMPO) to provide words/phrases to condition the LLM, enabling it to identify incorrect words. Second, all ontologies, in conjunction with UMLS, are used to obtain terms or phrases for validating the output of the LLM.

**Sub-pathology types.** The list of all 18 sub-pathology types used to prompt LLM on the text classification task are: *Bone, Cardiac, Cyto, Dermato, Endocrine, Gastrointestinal, Genitourinary, Gynecologic, Head and Neck, Hemato, Neuro, Ophthalmic, Pediatric, Pulmonary, Renal, Soft*

<sup>7</sup><https://bioportal.bioontology.org/ontologies/BCGO>

Table 3: Datasets used to train the histopathology image classifier. [ $\mu\text{m}$  per pixel - MPP]

Data Source	Subset	#WSI	#pathces	Train-Test	Magnification	Image-size
TCGA (H&E Stain)	GBM	19				
	LUSC	20			89,022 - 40x	
	LIHC	20				
	SARC	23	169,431	84715-16943	57,671 - 20x	384 x 384
	KIRC	16			16,660 - 10x	
	KICH	4			4,748 - 5x	
	BRCA	17			1,465 - 2.5x	
ACROBAT Weitz et al. [39]	SKCM	19			466 - 1.25x	
	H&E KI67	99	50589	28105-22484	(10x, 5x, 2.5x)	384 x 384
	ER , PGR, HER2	-				
BCI Liu et al. [24]	-	-	4,870		20x (0.46 MPP)	1024 x 1024
CESD Liu et al. [23]	-	-	686		100x/400x	2048 x 1536
Smear Hussain et al. [15]	-	-	963		400x	2048 x 1536
Celeb Liu et al. [25]	-	-	202,599	8,103-1,944	-	-
Places Zhou et al. [42]	-	-	36,550	2,109-1,372	-	-
A12D Kembhavi et al. [21]	-	-	4,903	0.7-0.3%	-	-
DocFig Jobin et al. [17]	-	-	33,004	0.8-0.2%	-	-
SciFig-pilot Karishma [19]	-	-	1,671	0.8-0.2%	-	-
SlideImages Morris et al. [28]	-	-	8,217	0.8-0.2%	-	-
TextVQA Singh et al. [36]	-	-	28,472	0.8-0.2%	-	-
SlideShare-1M Araujo et al. [3]	-	-	49,801	0.8-0.2%	-	-

**tissue, and Breast Histopathology.** Figure 7 provides the LLM prompt to retrieve the top three sub-pathology classification based on a given text.

System Prompt: You are a histopathology text classifier

User Prompt: Imagine you are a text classifier. Classify the given text into one of the following surgical pathology types namely: Bone, Cardiac, Cytopathology, Dermatopathology, Endocrine, Gastrointestinal, Genitourinary, Gynecologic, Head and Neck, Hematopathology, Neuropathology, Ophthalmic, Pediatric, Pulmonary, Renal, Soft tissue, Breast pathology. Output only the top 3 pathology types in an ordered python list

Few-shot examples: "Radicular cyst arises within the periodontal ligament space, particularly the periapex from the epithelial cell of malassez. These radicular cysts are caused by inflammation following the death of the pulp extending into the periapical radix. Radicular cysts caused by inflammation are always associated with a non vital tooth."  
 ["'Soft tissue', 'Dermatopathology', 'Hematopathology'"]

**INPUT:**  
 "There is a lesion with slight thickening of the muscularis mucosa and submucosa. There is a subtle change in the lamina propria that doesn't look quite like normal stromal cells. Description of slight thickening of the muscularis mucosa and submucosa with subtle changes in the lamina propria. Highlighted field shows the changes more dramatically. Abnormal cells in the lamina propria that appear pink and spindly."  
**OUTPUT:** ["'Gastrointestinal', 'Soft tissue', 'Hematopathology'"]

Figure 7: Prompting LLM with few-shot examples to extract the top three sub-pathology classification of a given text.

**Pre-processing multi-plane figures.** Many figures in academic papers are multi-plane, which means a number of sub-figures (Charts, graphs, histopathology and non-histopathology sub-figures) are placed next to each other to make a larger figure. We extracted individual images from multi-plane



figures to create multiple instance bags. To locate boundaries and white gaps between sub-figures, we utilized Sobel filters. Binary thresholding was then used to find the contours surrounding the sub-figures. We employ image size and image ratio thresholds to remove undesirable sub-figures and our histopathology classifier to maintain just histopathology sub-figures. We supply the histological sub-figures individually for this type of figure by appending "\_[0-9]+" to the end of the multi-plane figure id. If a figure is divided into more than 5 sub-figures, we preserve the original image to ensure that the resolution of these sub-figures remains reasonable. Figure 8 shows an overview of this pre-processing step in different scenarios of successful and unsuccessful crops.

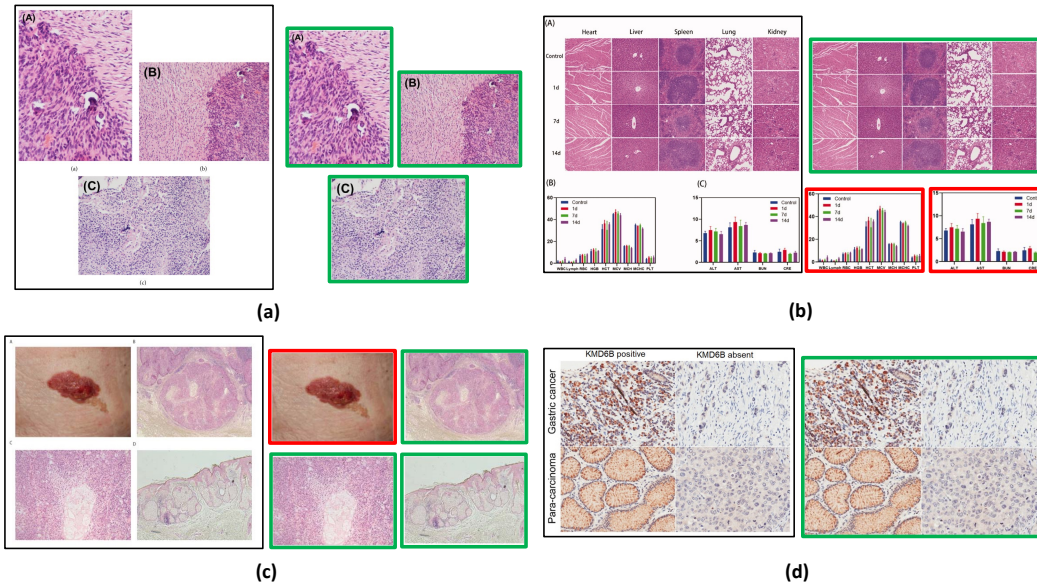


Figure 8: (a), (b), and (c) successfully cropped sub-figures where histopathology images (green box) are kept and non-histopathology (red box) images are removed. (b) histopathology crops are kept as not separated because the individual crops don't meet the size threshold so the original figure is kept. (d) Unsuccessful crop due to minimal gap between sub-figures. Original image is stored.

## A.5 Privacy preserving steps

In order to ensure privacy while handling the dataset, several steps were taken to protect sensitive information. These steps include:

- **Reduction of Signal to Noise using a LLM:** To protect the privacy of the dataset, a LLM was utilized to reduce the signal-to-noise ratio. By applying the LLM, irrelevant or sensitive information was masked or removed.
- **Exclusion of Videos Not Fully in Narrative Style:** Videos that did not adhere to a fully narrative style were intentionally left out of the dataset. This step was taken to minimize the risk of including any potentially private or sensitive content that could compromise individuals' privacy.
- **Release of Video IDs and Reconstruction Code:** Instead of directly releasing the complete dataset, only video IDs from YouTube were made public. Additionally, the code is provided to enable researchers to recreate the dataset.
- **Collection from Diverse Channels:** Data collection was performed from a wide range of sources, including both large and small channels. This approach aimed to decrease the risk of overfitting to specific channel types, thereby mitigating privacy concerns associated with potential biases linked to particular channels.

## B Exploratory analysis of the collected data

In this section, we provide the statistics of the QUILT dataset. Figure 4 illustrates the distribution of data across 18 sub-pathology types, offering a comprehensive analysis of the dataset’s text distribution. Moreover, for additional statistical details regarding QUILT, please refer to Table 4, which presents supplementary information on various aspects of the dataset.

Table 4: Additional QUILT statistics

Property	Average value
Medical text per image	1.74
ROI text per chunk	2.30
Medical text per chunk	1.93
Words per medical text	22.92
Words per ROI text	8.75
Images per chunk	2.49
Image-text pair per chunk	2.36
UMLS entity per medical text	4.36
UMLS entity per ROI text	1.61

## C Pretraining and downstream evaluation details

### C.1 External Evaluation Datasets

**PatchCamelyon** Veeling et al. [37] contains 327,680 color images (96×96px) from histopathology scans of lymph node sections. The images are assigned a binary label indicating whether they contain metastatic tissue or not. **NCT-CRC-HE-100K** Kather et al. [20] consists of 100,000 non-overlapping image patches from hematoxylin and eosin (H&E) stained histological images (224x224px) of human colorectal cancer and is categorized into cancer and normal tissue. **SICAPv2** Silva-Rodríguez et al. [35] contains 182 prostate histology WSIs with 10,340 patches (512 x 512px) and both annotations of global Gleason scores and patch-level Gleason grades. Images are labeled as Non cancerous, Grade 3, Grade 4, and Grade 5. **DatabioX** [6] consists of 922 Invasive Ductal Carcinoma cases of breast cancer. This data set has been collected from pathological biopsy samples of 150 patients which are labeled as Grade I, II and III. Each pathological sample in has four levels of magnification: 4x, 10x, 20x and 40x. **BACH** [4] consists of 400 WSIs of breast tissue which are labeled as normal, benign, in-situ and invasive carcinoma. **Osteo** [5] is a set of 1,144 patches (1024 x 1024px) taken from 40 WSIs representing the heterogeneity of osteosarcoma. Images are labeled as Viable tumor (VT), Non-tumor (NT) and Necrotic tumor (NEC). **RenalCell** [8] contains 27,849 images of clear-cell renal cell carcinoma H&E-stained (300 x 300px) annotated into five tissue texture types. **SkinCancer** [22] consists of 36,890 patches (395 x 395px) from WSIs skin biopsies from patients with Basal cell carcinoma (BCC), squamous cell carcinoma (SqCC), naevi and melanoma. Images were annotated for 16 categories: chondral tissue, dermis, elastosis, epidermis, hair follicle, skeletal muscle, necrosis, nerves, sebaceous glands, subcutis, eccrine glands, vessels, BCC, SqCC, naevi and melanoma. **MHIST** [38] contains 3,152 patches (224 x 224px) from 328 Formalin Fixed Paraffin-Embedded WSIs of colorectal polyps. These images are labeled as hyperplastic polyps (HPs) or sessile serrated adenomas (SSAs). **LC25000** [7] which we divide into **LC25000 (Lung)** with 15,000 and **LC25000 (Colon)** with 10,000 color images (768×768px). The lung subset is labeled as lung adenocarcinomas, lung squamous cell carcinomas, and benign lung tissues and the colon subset is labeled as colon adenocarcinomas and benign colonic tissues. Table 5 summarizes these datasets.

Table 5: Downstream tasks and datasets. Note that SkinTumor dataset is a subset of SkinCancer. [ $\mu\text{m}$  per pixel - MPP]

	Task	Sub-Pathology	Dataset	Classes	Magnification	Size (Train/Val/Test)	Image-size
Classification	Lymph-node metastasis detection	Breast	PatchCamelyon [37]	2	1 MPP	(0.75/0.125/0.125) * 327,680	96 x 96
	Tissue Phenotyping	Colorectal	NCT-CRC-HE-100K [20]	8	0.5 MPP	89,434/ - /6333	224 x 224
	Gleason scoring	Prostate	SICAPv2 [35]	4	1 MPP	- / - /10,340	512 x 512
	Bloom Richardson grading	Breast	Databiox [6]	3	[2,1,0.5,0.25] MPP	- / - /922	(2100 × 1574), (1276 × 956)
	Tissue classification (normal, benign, in-situ and invasive carcinoma)	Breast	BACH [4]	4	0.5 MPP	- / - / 400	2048 x 1536
	Osteosarcoma classification (non-tumor, necrotic tumor, and viable tumor)	Bone	Osteo [5]	3	1 MPP	- / - / 1,144	1024 x 1024
	clear-cell renal cell carcinoma tissue phenotyping (renal cancer, normal, stromal, other textures)	Renal	RenalCell [8]	5	[0.5, 0.25] MPP	- / - / 27,849	300 x 300
	Classification of skin neoplasms and various anatomical compartments	Skin	SkinCancer [22]	16	.5 MPP	28039/-/8851 <sup>imb</sup>	395 x 395
	Colorectal Polyp Classification	Colorectal	MHIST [38]	2	1 MPP	- / - / 3,152	224 x 224
	Lung adenocarcinoma classification (normal, adenocarcinoma and SCC)	Lung	LC25000 (LUNG) [7]	3	- MPP	- / - / 15,000	768 x 768
Colon adenocarcinoma classification (normal and colon adenocarcinoma)	Colon	LC25000 (Colon) [7]	2	- MPP	- / - / 10,000	768 x 768	
Retrieval	histopathology image-text retrieval	-	Quilt-1M	1.02M	-	13,559	-
	histopathology image-text retrieval	-	ARCH [12]	-	-	7500	-

## C.2 QUILTNET Implementation

All model implementations in this study are built upon the open source repository OpenCLIP [16], which enables large-scale training with contrastive image-text supervision. The experiments were conducted using PyTorch and utilized up to 4 NVIDIA A40 GPUs. The hyperparameters for finetuning and training from scratch are provided in Table 6. During the training process, gradient checkpointing and automatic mixed precision (AMP) techniques were employed, with a datatype of `bfloat16`.

All models were trained with image size of 224, except for the finetuned ViT-B-32 models, where the images were first resized to 512 before randomly cropping them to the desired size of 224. In the case of ViT-B-32 finetuning, the data was kept stretched, meaning it maintained a one-to-one mapping between the image and text. However, for all other models, the data was unstretched. This means that for those models, sampling from medical texts occurred with a probability of  $p = \text{sample prob}$ , or sampling from ROI texts. Within the medical or ROI texts, sampling was done uniformly. For all finetuned GPT/77 models we use the OpenAI CLIP [30] pretrained network as initialization and for ViT-32 maintain the use of QuickGeLU<sup>8</sup>. We perform hyperparameter tuning for all linear

<sup>8</sup><https://github.com/openai/CLIP/blob/main/clip/model.py>

probing results, exploring different values for learning rate, epochs, and weight decay. This process helped optimize the performance of the models during the linear probing stage.

Table 6: Training hyperparameters for QUILTNET

Hyperparameter	Finetuning	Training
batch size (per gpu)	256/1024	1024
peak learning rate	1e-5	5.0e-4
learning rate schedule	constant	cosine decay
epochs	15	40
warmup (in steps)	200	2000
random seed	0	0
image mean	(0.48145466, 0.4578275, 0.40821073)	same
image std	(0.26862954, 0.26130258, 0.27577711)	same
augmentation	Resize; RandomCrop (0.8, 1.0)	RandomResizedCrop (0.8, 1.0)
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.98$	same
weight decay	0.1	0.2
eps	1.0e-6	same
optimizer	AdamW [27]	same
<i>sample prob</i>	0.85	same

Table 7: Zero-shot image classification. accuracy (%). \* denotes models trained from scratch. SkinTumor is the Neoplastic Subset of SkinCancer. Also note that PMB refers to PubmedBert, a BERT model of 256 context length pre-trained on PMC-15M. We swapped our model’s text encoder from GPT2 to PubmedBert to assess performance differences

Dataset	ViT-B/32				ViT-B/16			
	CLIP	PLIP	QUILTNET		CLIP	BiomedCLIP	QUILTNET	
	GPT/77	GPT/77	GPT/77	(GPT/77)*	GPT/77	PMB/256	GPT/77	PMB/256
SkinCancer	5.40	36.65	<b>45.38</b>	8.93	5.40	24.75	23.41	<b>28.93</b>
SkinTumor	10.35	56.36	<b>58.29</b>	36.26	13.85	37.0	<b>51.47</b>	51.20
NCT-CRC	26.4	54.02	<b>59.56</b>	17.35	20.09	51.71	28.68	<b>59.20</b>
PatchCamelyon	61.88	58.61	<b>64.6</b>	49.92	50.45	53.25	<b>67.91</b>	53.52
MHIST	52.92	57.52	<b>62.54</b>	44.52	52.3	40.23	44.32	<b>52.71</b>
LC25000(LUNG)	61.36	78.77	<b>80.16</b>	67.71	50.29	72.44	50.71	<b>81.87</b>
LC25000(COLON)	62.5	77.79	<b>93.28</b>	72.08	78.56	<b>90.57</b>	62.26	87.1
SICAPv2	39.40	<b>44.53</b>	39.49	25.07	27.38	<b>45.81</b>	25.54	45.1
BACH	26.0	<b>43.0</b>	41.25	33.75	27.25	54.75	40.75	<b>62.0</b>
Databiox	37.53	39.48	<b>42.52</b>	32.32	<b>33.51</b>	31.24	33.19	29.93
Osteo	19.49	54.02	<b>64.16</b>	27.88	16.08	50.79	38.37	<b>59.79</b>
RenalCell	20.3	50.7	<b>52.57</b>	16.35	28.80	47.08	28.32	<b>50.72</b>

Table 8: Classes for each dataset on zero-shot image classification. Note that we used the same prompt templates for each dataset. The templates used are: ["a histopathology slide showing {c}", "histopathology image of {c}", "pathology tissue showing {c}", "presence of {c} tissue on image"]

Dataset	Classes
SkinCancer	'Necrosis', 'Skeletal muscle', 'Eccrine sweat glands', 'Vessels', 'Elastosis', 'Chondral tissue', 'Hair follicle', 'Epidermis', 'Nerves', 'Subcutis', 'Dermis', 'Sebaceous glands', 'Squamous-cell carcinoma', 'Melanoma in-situ', 'Basal-cell carcinoma', 'Naevus'
PatchCamelyon	'Lymph node', 'Lymph node containing metastatic tumor tissue'
NCK-CRC	'Adipose', 'Debris', 'Lymphocytes', 'Mucus', 'Smooth muscle', 'Normal colon mucosa', 'Cancer-associated stroma', 'Colorectal adenocarcinoma epithelium'
MHIST	'Hyperplastic polyp', 'Sessile serrated adenoma'
LC25000Lung	'Lung adenocarcinoma', 'Benign lung', 'Lung squamous cell carcinoma'
LC25000Colon	'Colon adenocarcinoma', 'Benign colonic tissue'
BACH	'Breast non-malignant benign tissue', 'Breast malignant in-situ carcinoma', 'Breast malignant invasive carcinoma', 'Breast normal breast tissue'
SICAPv2	'Benign glands', 'Atrophic dense glands', 'Cribriform ill-formed fused papillary patterns', 'Isolated nest cells without lumen rosetting patterns'
Databiox	'Well differentiated bloom richardson grade one', 'Moderately differentiated bloom richardson grade two', 'Poorly differentiated grade three'
RenalCell	'Red blood cells', 'Renal cancer', 'Normal tissue', 'Torn adipose necrotic tissue', 'Muscle fibrous stroma blood vessels'
Osteo	'Normal non-tumor', 'Necrotic', 'Tumor'
SkinTumor	'Squamous-cell carcinoma', 'Melanoma in-situ', 'Basal-cell carcinoma', 'Naevus'

## D Exploration of trained model representations

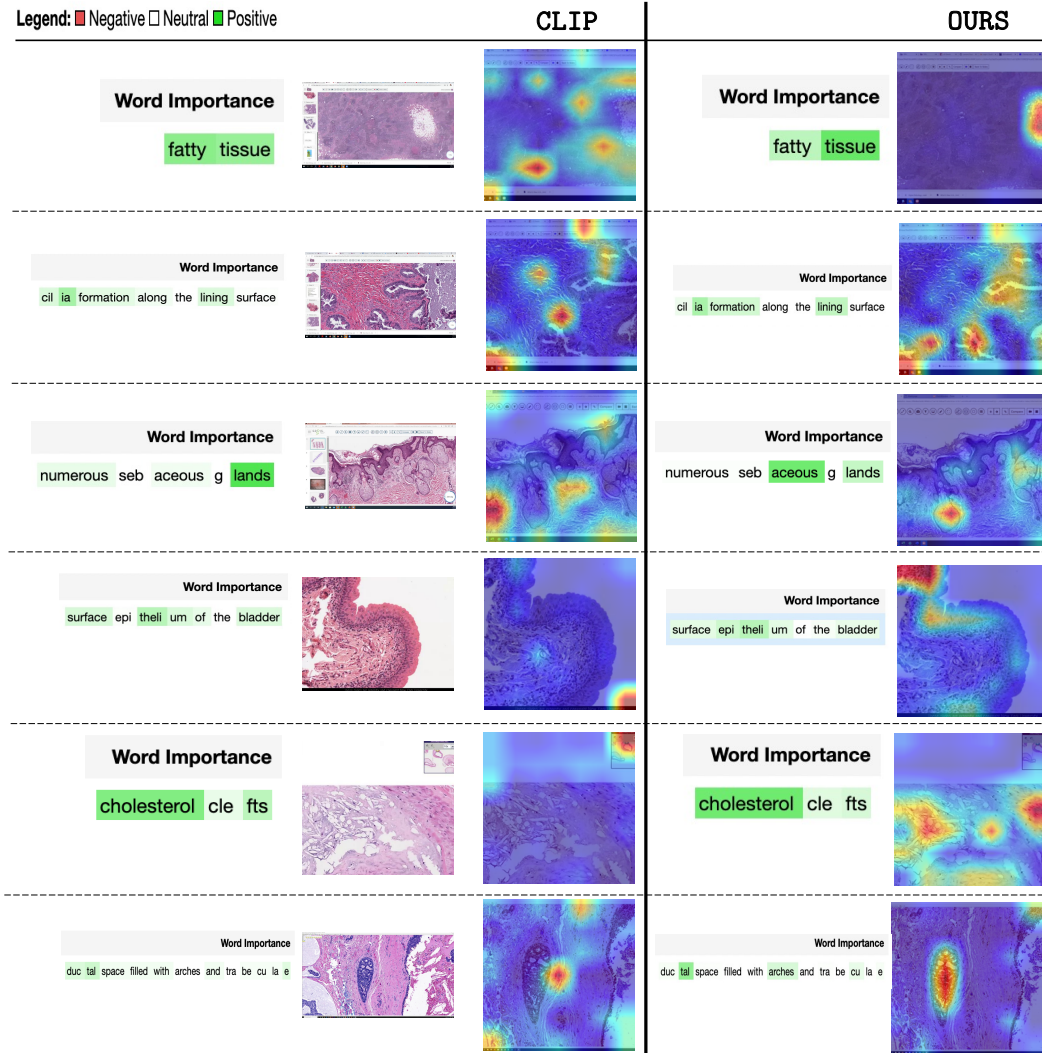
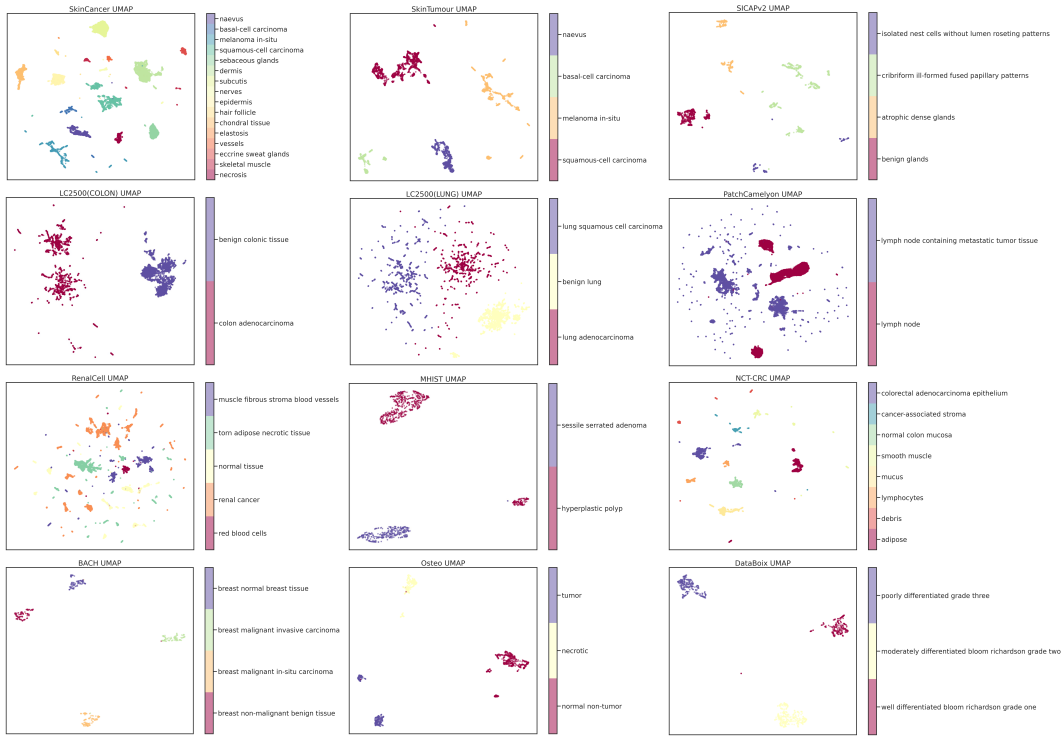


Figure 9: Comparison of the attention maps generated by QUILTNET and CLIP. The corresponding words are highlighted based on their importance. Attention masks were generated using GradCAM [34].

Table 9: UMAP visualization of image embeddings generated by QUILTNET from the different datasets listed in Table 5.



## E Datasheet for QUILT

In this section, we present a DataSheet [13] for QUILT, synthesizing many of the other analyses we performed in this paper.

### 1. Motivation For Datasheet Creation

- **Why was the dataset created?** To train histopathology multi-modal models, on in-domain data, useful for diagnostically relevant downstream tasks.
- **Has the dataset been used already?** Yes.
- **What (other) tasks could the dataset be used for?** Could be used as training data for representation learning, and also for supervised learning on metadata
- **Who funded dataset creation?** This work was funded by the Office of the Assistant Secretary of Defense328 for Health Affairs through the Melanoma Research Program under Awards No. W81XWH-20-1-0797329 and W81XWH-20-1-0798.

### 2. Data composition

- **What are the instances?** The instances that we consider in this work are histopathology images derived from educational videos, paired with aligned text, derived from ASR and denoise using an LLM.
- **How many instances are there?** We include greater than 1 million image-text pairs, from videos and additionally from less noisy sources like PubMed articles.
- **What data does each instance consist of?** Each instance consists of an image, a descriptive text for the image as a whole and for its regions of interest, an estimated microscope magnification of the image, medical UMLS entities in the text, and the subpathology type. Each instance is representative of a video chunk based on where histopathology content is stable.
- **Is there a label or target associated with each instance?** We use the raw ASR and LLM denoised captions as labels in this work as well as auxiliary information which includes magnification, UMLS entities and pathology type.

- **Is any information missing from individual instances?** Yes, for instances in the dataset that are not from QUILT (i.e. videos), e.g. from PubMed Article datapoints, the additional auxiliary information is not included.
- **Are relationships between individual instances made explicit?** Not applicable – we do not study relationships between disparate videos (even from the same narrator) nor the relationship between chunks in the same video.
- **Does the dataset contain all possible instances or is it a sample?** Contains all instances our curation pipeline collected, as the list of videos is not exhaustive of what is available online, there is a high probability more instances can be collected in the future.
- **Are there recommended data splits (e.g., training, development/validation, testing)?** There are no recommended data splits, as this data was curated mainly for pretraining rather than evaluation.
- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** Yes. Despite our numerous attempts to reduce noise using various models, algorithms and human knowledge databases, ASR is often noisy, and there are many errors that we cannot fix.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** The dataset is self-contained. However, we plan to only release the video URLs and some paired non-pixel data points, rather than the videos themselves, so as to protect user privacy (allowing users to delete videos).

### 3. Collection Process

- **What mechanisms or procedures were used to collect the data?** We leveraged the YouTube API and the youtube-dl library.
- **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data?** The data was directly observable (public) (from YouTube).
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** We used a probabilistic strategy with many algorithms and heuristics, more details are in Appendix A.1.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** Data collection was primarily done by the first authors of this paper.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** The data was collected from January 2023 to May 2023, even though the YouTube videos are often much older.

### 4. Data Preprocessing

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Yes, we discuss this in Section ?? and in Appendix A.1: of note, we use a large language model, UMLS database and a set of algorithms to ‘denoise’ ASR transcripts, an ensemble of histopathology classifiers to inform relevant segments of the video, and extract the representative image(s) for each video segment.
- **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the ‘raw’ data.** The raw data was saved, but at this time we do not plan to release it directly due to copyright and privacy concerns.
- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** Yes, software for downloading and processing the data is available on GitHub through our website.



- **Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?**

Yes, the dataset does allow for the study of our goal, as it covers various histopathology sub-domains and provides crucial data points and metadata for pretraining. Some of its limitations we are aware of involve various biases on YouTube, as well as various inaccuracies of the models (e.g ASR model) within the curation pipeline, which we discuss in Appendix A.1 and A.3.

#### 5. Dataset Distribution

- **How will the dataset be distributed?** We distribute all the derived data (denoised frames, captions, magnifications etc), as well as links to the YouTube videos that we used under the MIT license and strictly for research purposes.
- **When will the dataset be released/first distributed? What license (if any) is it distributed under?** The data has been released, under the permissible MIT license for research-based use only.
- **Are there any copyrights on the data?** We believe our use is ‘fair use,’ however, due to an abundance of caution, we will not be releasing any of the videos themselves.
- **Are there any fees or access restrictions?** No.

#### 6. Dataset Maintenance

- **Who is supporting/hosting/maintaining the dataset?** The first authors of this paper.
- **Will the dataset be updated? If so, how often and by whom?** We do not plan to update it at this time.
- **Is there a repository to link to any/all papers/systems that use this dataset?** Not right now, but we encourage anyone who uses the dataset to cite our paper so it can be easily found.
- **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?** Not at this time.

#### 7. Legal and Ethical Considerations

- **Were any ethical review processes conducted (e.g., by an institutional review board)?** No official processes were done, as our research is not on human subjects, however, because the dataset is in the medical domain we had significant internal discussions and deliberations when choosing the scraping strategy.
- **Does the dataset contain data that might be considered confidential?** No, we only use public videos.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why** No – because many of these videos are medical and educational in nature, we have not seen any instance of offensive or abusive content.
- **Does the dataset relate to people?** Yes, it relates sometimes to deidentified patients, typically studied by pathologists.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** Not explicitly (e.g. through labels)
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** Yes, some of our data includes content from known pathologists, albeit niche, they sometimes include their faces in the corner of the video. All of the videos that we use are of publicly available data, following the Terms of Service that users agreed to when uploading to YouTube.

## References

- [1] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.

- [2] M. Amith, L. Cui, K. Roberts, H. Xu, and C. Tao. Ontology of consumer health vocabulary: providing a formal and interoperable semantic resource for linking lay language and medical terminology. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1177–1178. IEEE, 2019.
- [3] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod. Large-scale query-by-image video retrieval using bloom filters. *arXiv preprint arXiv:1604.07939*, 2016.
- [4] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.
- [5] H. B. Arunachalam, R. Mishra, O. Daescu, K. Cederberg, D. Rakheja, A. Sengupta, D. Leonard, R. Hallac, and P. Leavey. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706, 2019.
- [6] H. Bolhasani, E. Amjadi, M. Tabatabaeian, and S. J. Jassbi. A histopathological image dataset for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked*, 19:100341, 2020.
- [7] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides. Lung and colon cancer histopathological image dataset (1c25000). *arXiv preprint arXiv:1912.12142*, 2019.
- [8] O. Brummer, P. Polonen, S. Mustjoki, and O. Bruck. Integrative analysis of histological textures and lymphocyte infiltration in renal cell carcinoma using deep learning. *bioRxiv*, pages 2022–08, 2022.
- [9] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.
- [10] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier. An open-source speaker gender detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.
- [11] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright. Overview and utilization of the nci thesaurus. *Comparative and functional genomics*, 5(8):648–654, 2004.
- [12] J. Gamper and N. Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16549–16559, 2021.
- [13] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [14] Z. Huang, F. Bianchi, M. Yuksekogonul, T. Montine, and J. Zou. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, pages 2023–03, 2023.
- [15] E. Hussain, L. B. Mahanta, H. Borah, and C. R. Das. Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data in brief*, 30:105589, 2020.
- [16] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>.
- [17] K. Jobin, A. Mondal, and C. Jawahar. Docfigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE, 2019.
- [18] S. Jupp, J. Malone, T. Burdett, J.-K. Hériché, E. Williams, J. Ellenberg, H. Parkinson, and G. Rustici. The cellular microscopy phenotype ontology. *Journal of biomedical semantics*, 7: 1–8, 2016.

- [19] Z. Karishma. Scientific document figure extraction, clustering and classification. 2021.
- [20] J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo*10, 5281, 2018.
- [21] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
- [22] K. Kriegsmann, F. Lobers, C. Zgorzelski, J. Kriegsmann, C. Janßen, R. R. Meliß, T. Muley, U. Sack, G. Steinbuss, and M. Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12, 2022.
- [23] J. Liu, Q. Wang, H. Fan, S. Wang, W. Li, Y. Tang, D. Wang, M. Zhou, and L. Chen. Automatic label correction for the accurate edge detection of overlapping cervical cells. *arXiv preprint arXiv:2010.01919*, 2020.
- [24] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1815–1824, 2022.
- [25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [26] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.
- [27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [28] D. Morris, E. Müller-Budack, and R. Ewerth. Slideimages: a dataset for educational image classification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 289–296. Springer, 2020.
- [29] N. F. Noy, M. A. Musen, J. L. Mejino Jr, and C. Rosse. Pushing the envelope: challenges in a frame-based representation of human anatomy. *Data & Knowledge Engineering*, 48(3): 335–359, 2004.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.
- [32] P. N. Schofield, J. P. Sundberg, B. A. Sundberg, C. McKerlie, and G. V. Gkoutos. The mouse pathology ontology, mpath; structure and applications. *Journal of biomedical semantics*, 4(1): 1–8, 2013.
- [33] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- [34] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. arxiv 2016. *arXiv preprint arXiv:1610.02391*.
- [35] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195:105637, 2020.

- [36] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [37] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.
- [38] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, N. Tomita, L. Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.
- [39] P. Weitz, M. Valkonen, L. Solorzano, C. Carr, K. Kartasalo, C. Boissin, S. Koivukoski, A. Kuusela, D. Rasic, Y. Feng, et al. Acrobat—a multi-stain breast cancer histological whole-slide-image data set from routine diagnostics for computational pathology. *arXiv preprint arXiv:2211.13621*, 2022.
- [40] P. S. Wright, K. A. Briggs, R. Thomas, G. F. Smith, G. Maglennon, P. Mikulskis, M. Chapman, N. Greene, B. U. Phillips, and A. Bender. Statistical analysis of preclinical inter-species concordance of histopathological findings in the etox database. *Regulatory Toxicology and Pharmacology*, 138:105308, 2023.
- [41] W. Wu, S. Mehta, S. Nofallah, S. Knezevich, C. J. May, O. H. Chang, J. G. Elmore, and L. G. Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9: 163526–163541, 2021.
- [42] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.