



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


Abstract

1 Recent accelerations in multi-modal applications have been made possible with the
2 plethora of image and text data available online. However, the scarcity of analogous
3 data in the medical field, specifically in histopathology, has slowed comparable
4 progress. To enable similar representation learning for histopathology, we turn
5 to YouTube, an untapped resource of videos, offering 1,087 hours of valuable
6 educational histopathology videos from expert clinicians. From YouTube, we curate
7 QUILT: a large-scale vision-language dataset consisting of 802,148 image and
8 text pairs. QUILT was automatically curated using a mixture of models, including
9 large language models, handcrafted algorithms, human knowledge databases, and
10 automatic speech recognition. In comparison, the most comprehensive datasets
11 curated for histopathology amass only around 200K samples. We combine QUILT
12 with datasets from other sources, including Twitter, research papers, and the internet
13 in general, to create an even larger dataset: QUILT-1M, with 1M paired image-
14 text samples, marking it as the largest vision-language histopathology dataset to
15 date. We demonstrate the value of QUILT-1M by fine-tuning a pre-trained CLIP
16 model. Our model outperforms state-of-the-art models on both zero-shot and
17 linear probing tasks for classifying new histopathology images across 13 diverse
18 patch-level datasets of 8 different sub-pathologies and cross-modal retrieval tasks².

19 1 Introduction

20 Whole-slide histopathology images are dense in information, and even individual image patches can
21 hold unique, complex patterns critical for tissue characterization. Summarizing this information
22 into a single label is an oversimplification that fails to capture the complexity of the field, which
23 covers thousands of evolving disease sub-types [59]. This highlights the need for more expressive,
24 dense, interconnected representations beyond the reach of a singular categorical label. As such,
25 natural language descriptions can provide this comprehensive signal, linking diverse features of
26 histopathology sub-patch structures [20, 25].

27 If there were a large-scale vision-language dataset for histopathology, researchers would be able to
28 leverage the significant advancements in self-supervised vision and language pre-training to develop

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

²The data and code will be available at Quilt-1M

29 effective histopathology models [49]. Unfortunately, there is a significant scarcity of comprehensive
 30 datasets for histopathology. Notable open-source contributions have been made with datasets like
 31 ARCH [20] and OpenPath [25]. Yet, these sources are still somewhat limited due to their size, as the
 32 former has only $\approx 8K$ samples and the latter (the largest histopathology vision-language dataset to
 33 date) has about 200K samples. Although recent efforts (e.g. PMC-15M [71]) curated 15M image-text
 34 pairs across a variety of different biomedical domains from Pubmed [51], whether their samples are
 35 specific to histopathology remains ambiguous; worse, their dataset is not openly available.

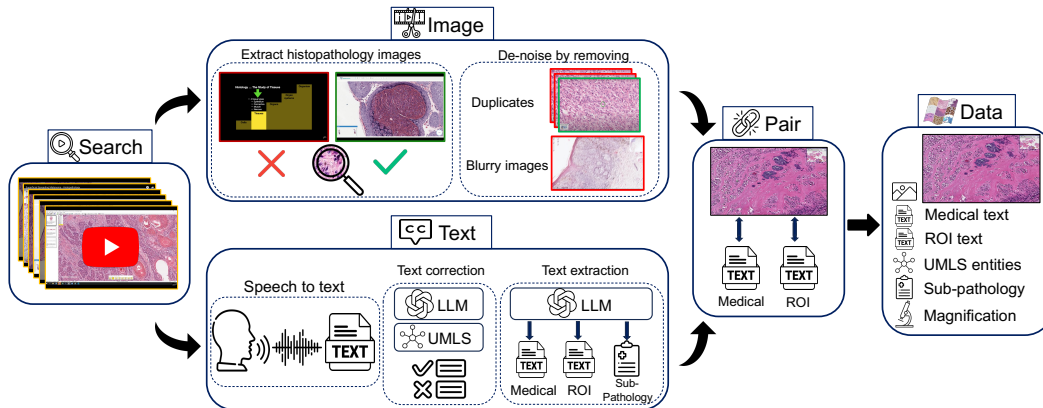


Figure 1: **Overview of QUILT curation pipeline.** We identify relevant histopathology YouTube videos in **Search**. For **Image** extraction, we find and de-noise histopathology frames using trained models. In **Text** section, we rely on a conventional Automatic Speech Recognition (ASR) model and leverage Unified Medical Language System (UMLS) and large language models (LLMs) for post-processing and ASR error correction. Relevant sub-pathology, medical and region-of-interest (ROI) text are extracted using an LLM. Finally, domain-specific algorithms are used to **Pair** images and text, eliminating duplicates to yield QUILT, a richly annotated image-text dataset for histopathology.

36 To address the need for a large-scale vision-language dataset in histopathology, we introduce QUILT:
 37 containing 437, 878 images aligned with 802, 148 text pairs across multiple microscopic magnifi-
 38 cation scales covering from 10x to 40x. We draw on the insight that publicly available educational
 39 YouTube histopathology content represents an untapped potential. We curate QUILT using 1, 087
 40 hours of valuable educational histopathology videos from expert pathologists on YouTube. To extract
 41 aligned image and text pairs from the videos, we utilize a mixture of models: large language models
 42 (GPT-3.5), handcrafted algorithms, human knowledge databases, and automatic speech recognition.
 43 QUILT does not overlap with any current open-access histopathology data sources. This allows
 44 us to merge our dataset with other open-source datasets available. Therefore, to create an even
 45 larger and more diverse dataset, we combine QUILT with data from other sources, such as Twitter,
 46 research papers, and the Internet, resulting in QUILT-1M. The larger QUILT-1M contains one million
 47 image-text pairs, making it the largest public vision-language histopathology dataset to date.

48 Using QUILT and QUILT-1M, we finetune vision-language models using a contrastive objective
 49 between the two modalities. We extensively evaluate it on 13 external histopathology datasets taken
 50 across different sub-pathologies. We report zero-shot classification, linear probe, and image-to-text
 51 and text-to-image retrieval tasks. Against multiple recently proposed baselines (CLIP [49], PLIP [25],
 52 and BiomedCLIP [71]), models trained with QUILT-1M outperform all others. Our ablations identify
 53 the importance of QUILT.

54 QUILT offers three significant advantages: First, QUILT does not overlap with existing data sources;
 55 it ensures a unique contribution to the pool of available histopathology knowledge. Second, its
 56 rich textual descriptions extracted from experts narrating within educational videos provide more
 57 expressive, dense interconnected information. Last, the presence of multiple sentences per image
 58 fosters diverse perspectives and a comprehensive understanding of each histopathological image. We
 59 hope that both computer scientists and histopathologists will benefit from QUILT’s potential.

60 2 Related work

61 We built upon a growing literature applying self-supervised learning and other machine learning
62 methods to medical image understanding.

63 **Machine learning for histopathology.** Early representation learning work in computational pathol-
64 ogy primarily relied on weakly-supervised learning, with each whole-slide image (WSI) receiving
65 a single label. The limited nature (single label to many patches) has produced sub-optimal models
66 [12, 27] at the patch level. Lately, a self-supervised learning approach, which learns useful representa-
67 tions from unlabeled data, has shown some success [27, 13, 12]. Most of this work has been unimodal.
68 They use image augmentations similar to those used for natural images [14], mostly differing by way
69 of consciously injecting domain knowledge. For example, they leverage the compositional nature of
70 H&E stain information of whole-slice images [27], or inject hierarchical morphological information
71 at different magnifications [13], or combine with other modalities like genomic features [12] or with
72 descriptive text [20]. When text data is used, the objectives similarly use augmentations seen in natural
73 language [53]. By contrast, we explore self-supervised mechanisms that learn better histopathology
74 information representations that go beyond a single label, aided by language descriptions.

75 **Medical vision-language datasets.** Learning vision-language representations demands a large
76 dataset of images aligned with descriptive text, a resource that is notably lacking in histopathology.
77 The MIMIC-CXR-JPG v2.0.0 dataset [30], for example, consists of de-identified hospital-sourced
78 chest radiographs and reports. For histopathology, The Cancer Genome Atlas³ provides de-identified
79 PDF-reports for a limited number of WSIs. Despite this resource, the enormous size of this data
80 (reaching up to 120,000² pixels) makes processing challenging, limiting its use to a small number of
81 focused studies [42]. A majority of medical vision-language datasets are concentrated in the radiology
82 sub-domain, due to the relatively straightforward process of collecting validated multimodal data [30].
83 Many models are trained on a subset of PubMed [51] or comparable radiology datasets [72, 24, 18, 46].
84 PMC-15M [71], a recent subset of PubMed not specific to histopathology, was used to train multiple
85 models. While the models themselves are public, PMC-15M is not, making it hard to determine what
86 portion of it is histopathology-relevant.

87 **Vision-language pairs on histopathology.** One of the first histopathology vision-language datasets,
88 ARCH, contains only 7,614 accessible image-text pairs [20, 23]. Later on, [25] released OpenPath,
89 a dataset of 200K image-text pairs extracted from Twitter. This was the largest histopathology dataset
90 until QUILT-1M.

91 **Video data for self-supervision.** Numerous recent studies have started to tap into video data.
92 For instance, millions of publicly accessible YouTube videos were used to train a vision-language
93 model [69, 70]. Similarly, a causal video model was trained by using sequential gaming videos [6].
94 Localized narratives [62, 47] provide another example of dense, interconnected supervision for a
95 single image. Despite the potential of video content, video often yields noisier datasets compared
96 to static sources. Recently, the enhanced capabilities of automatic speech recognition models
97 streamlined the curation of large-scale cleaner datasets from videos [69, 6, 71]. Furthermore, the
98 growing versatility of large language models has shown promise as data annotators, information
99 extractors [35, 63, 15, 22], text correctors [67], and as tools for medical information extraction and
100 reasoning [1, 60].

101 3 Curating QUILT: Overview

102 Creating a vision-language dataset from videos is a significant undertaking, as not all videos are
103 suitable for our pipeline. Many either lack voiced audio, are not in English, fail to contain medically
104 relevant content, or have insufficient medical relevance—for example, videos that present static
105 images of histopathology content on a slide deck, or those that briefly cover histopathology images
106 in pursuit of a different objective. Conventional automatic speech recognition (ASR) systems also
107 struggle with the specialized requirements of histopathology transcription, necessitating a non-trivial

³<https://www.cancer.gov/tcga>

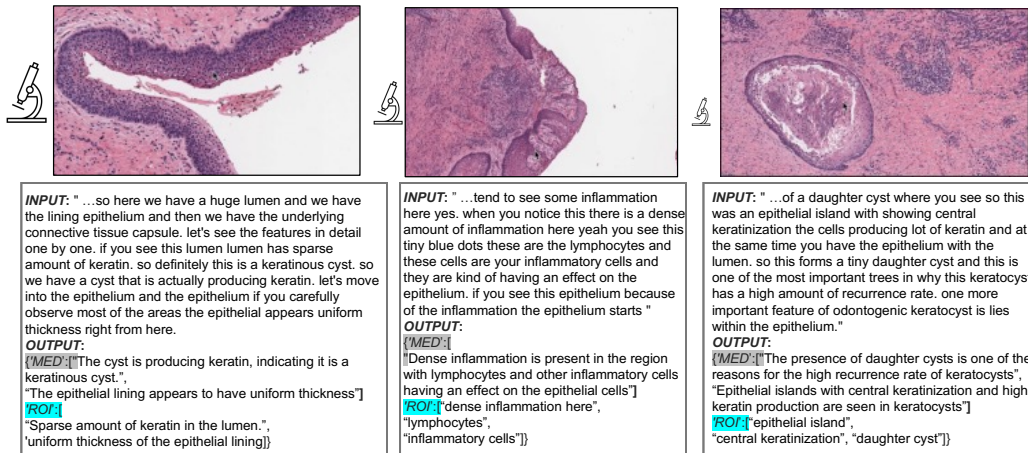


Figure 2: **QUILT examples.** **Input** is the corrected ASR caption for the representative image. **Output** are the medical and ROI extracted text(s) paired with the image (see Section 3.1). In histopathology, understanding tissue characteristics often involves views from varying magnification levels. Thus, in QUILT we estimate an image’s magnification (indicated by the relative size of the microscope icon).

108 solution. The de-noising of text and image modalities adds further complexity as the videos are
 109 typically conversational and, therefore, inherently noisy. Instructors pan and zoom at varying speeds,
 110 recording a mix of relevant and irrelevant histopathological visual content in their videos. As such,
 111 trivially extracting frames at static intervals fails to capture the data appropriately. To collect QUILT
 112 we trained models and handcrafted algorithms that leverage the nuances in the instructors’ textual
 113 and visual behavior, ensuring accurate collection and alignment of both modalities.

114 **3.1 QUILT: Collecting medical image and text pairs from YouTube**

115 Our proposed dataset curation pipeline involves (1) gathering channel and video data covering the
 116 histopathology domain, (2) filtering videos based on a certain "narrative style", (3) extracting and
 117 denoising image and text modalities from videos using various models, tools, and algorithms, (4)
 118 postprocessing denoised text by LLMs to extract medical text and finally, (5) splitting and aligning
 119 all modalities for curating the final vision-language pre-training (VLP) data. See Figure 1 (and A.1 in
 120 the Appendix) for a detailed overview of the pipeline.

121 **Collecting representative channels and videos.** Our pipeline begins by searching for relevant
 122 channels and video ids on YouTube, focusing on the domain of histopathology. Using keywords
 123 spanning 18 sub-pathology fields (see section A.4 in the Appendix), we search among channels before
 124 searching for videos to expedite discovery, considering that video searches are time-consuming and
 125 the APIs pose limitations on numerous requests [69]. Channels with subscriber count $\geq 300K$ are
 126 excluded to avoid large general science channels, as educational histopathology channels often have
 127 fewer subscribers. We then download low-resolution versions of all identified videos, with the lowest
 128 resolution at 320p.

129 **Filtering for narrative-style medical videos.** For each video within each channel, we exclude videos
 130 that are shorter than 1 minute, non-voiced, or have non-English audio. For videos meeting these
 131 heuristics, two decisions are made:

- 132 (A) Do they have the required medical content, i.e., histopathology image-text pairs?
- 133 (B) If so, are they in narrative style – videos wherein the presenter(s) spend a significant time
 134 panning and zooming on the WSI, while providing vocal descriptions of image content?

135 For **(A)** we automatically identify the relevant videos by extracting keyframes from a video. These
136 keyframes are automatically extracted using FFmpeg⁴, marking the beginning or end of a scene
137 (frames containing significant visual changes). The software requires a threshold that determines
138 the minimum amount of visual change required to trigger a keyframe. Through experimentation,
139 we set different thresholds for various video durations, with smaller thresholds for longer videos.
140 Next, we train and use an ensemble of three histopathology image classifiers to identify videos with
141 histopathology images (See section A.3 in the Appendix).

142 For **(B)**, in which we identify narrative-style videos, we randomly select keyframes predicted to be
143 histopathology. For each such selected frame, we extract the next three histopathology key-frames
144 and compute the cosine similarity between the selected frame and each of the subsequent three
145 frames. If all three have similarity scores \geq a preset threshold of 0.9, we count it as a narrative
146 streak. A video is identified as narrative style if at least 10% of the selected frames exhibit a narrative
147 streak. Consequently, we download all narrative-style videos at high-resolution. Narrative-style
148 videos typically cover WSIs at various magnifications, hence, we train a tissue-image-magnification
149 classifier to predict the following three scales: $\{(1 - 10)x, (> 10 - 20)x, (> 20)x\}$. This provides
150 relevant metadata for downstream objectives.

151 **Text Extraction using ASR and text denoising.** The high costs associated with private medical
152 ASR APIs⁵ necessitated the use of a more conventional ASR model: Whisper [50]. As anticipated,
153 this model often misinterprets medical terms, thus requiring the use of post-processing algorithms to
154 minimize its error rates.

155 We propose a four-step text de-noising and quality control pipeline: **i)** We utilize the Rake key-
156 word extraction algorithm to extract keywords or key-phrases up to four words and refine them by
157 eliminating stopwords [52]. **ii)** We then cross-check each refined entry against UMLS [7] using
158 the SciSpacy entity linking package [44]. If an entry is not found within UMLS, we check for
159 misspelled words within the entry using a spell-checking algorithm⁶, instantiated with a specialized
160 list of histopathology terms curated from various histopathology ontology labels and definitions. **iii)**
161 With this probable list of misspelled keywords, we *condition* and prompt the LLM with examples
162 to correct the misspelled entry within its context (sentence), and secondly, we task the LLM with
163 identifying additional *unconditioned* errors/misspelled entries. For both, we leverage a set of manually
164 curated examples to prompt the LLM in-context. For more examples and failure cases, see Table 11
165 and Figure 9 in the Appendix. **iv)** Finally, to de-noise the text, we resolve the output mapping of
166 incorrect \mapsto correct entries by verifying the corrected words against UMLS and our curated list of
167 histopathology words/phrases. Entries that pass this double-validation process are used to replace
168 the initial noisy transcription. Leveraging domain-specific databases to extract the text and filter out
169 noise allows us to bypass the correction of repetition errors and filler words, such as 'ah', 'uhm', 'the',
170 *etc.* in tandem, using LLMs allows us to concentrate on correcting medically-relevant misspelled
171 words, rather than correcting non-medically-relevant terms.

172 From the ASR-corrected text, we extract *medical text* which describes the image(s) as a whole. Also,
173 when the speaker describes/gestures at visual regions-of-interest through statements like "look here
174 ...", we extract the text entity being described as *ROI text*. To filter relevant medical text and ROI text
175 from the ASR-corrected text, we utilize LLMs (see Figure 9 in Appendix), a decision rooted in a
176 few compelling reasons: 1) Curating pre-training datasets at a scale that can tolerate higher levels of
177 noise, LLMs are more cost-effective than expert-human (medical) labor. 2) The task does not require
178 LLMs to generate new information but instead they discriminate useful versus irrelevant signals,
179 serving to improve the signal-to-noise ratio of the data. To extract relevant text, we prompt LLMs
180 to filter out all non-medically relevant text, providing context as necessary. See Figure 2 for some
181 example image-text pairs. Lastly, we instruct the LLMs to refrain from introducing any new words
182 beyond the corrected noisy text and set the model's temperature to zero. Finally, we use LLMs to
183 categorize our videos into one of the 18 identified sub-pathology classes. Similar to the previous

⁴<https://ffmpeg.org/>

⁵nuance.com/en-au/healthcare/provider-solutions/speech-recognition/dragon-medical-one.html

⁶<https://github.com/barrust/pyspellchecker>

184 tasks, this categorization is done by conditioning with a few examples and prompting the LLM to
185 predict the top three possible classes given the text. More details, prompts, and additional examples
186 are presented in Figure 12 within the Appendix.

187 **Image frame extraction and denoising.** For each video, we employ a similar method to that
188 described in **Filtering for narrative-style medical videos** subsection to extract histopathology
189 key-frames; our method leverages these frames’ times t as beacons to break the entire video into
190 time-intervals called *chunks* from which to extract representative image(s). Next, we extract the
191 median image (pixel-space) of stable (static) frames in each chunk if they exists, else we de-duplicate
192 the histopathology keyframes (beacons of the chunk). In essence, we use the extracted histopathology
193 scene frames as guides for data collection, exploiting the human tendency in educational videos to
194 pause narration during explanation, and we extract the relevant frame(s).

195 **Aligning both modalities.** For each narrative-style video, we perform the following steps to
196 align image and text modalities: First, we compute histopathology time chunks denoted as
197 $[(t_1, t_2), (t_3, t_4), \dots (t_{n-1}, t_n)]$ from keyframes after discriminating histopathology frames using the
198 histopathology ensemble classifier – (*scene_chunks*). Each *scene_chunk* is padded with *pad_time* to
199 its left and right; see Figure 8 and Table 9 in the Appendix for more details.

- 200 1. **Text:** we use the ASR output to extract the words spoken during each chunk in *scene_chunks*.
201 Using the method described in **Text Extraction using ASR and text denoising** subsection,
202 we extract the Medical and ROI caption for this chunk.
- 203 2. **Image:** we extract representative image(s) for every chunk/time-interval in *scene_chunks* as
204 described in **Filtering for narrative-style medical videos** subsection above.

205 Finally, each chunk in *scene_chunks* is mapped to texts (both medical and ROI captions) and images.
206 Next we map each medical image to one or more medical text. Using the time interval in which
207 the image occurs, we extract its raw text from ASR and then correct and extract keywords using
208 the Rake method, which we refer to as *raw_keywords*. We extract keywords from each medical
209 text returned using the LLM, and we refer to these as *keywords*. Finally, if the *raw_keywords* occur
210 before or slightly after a selected representative image, and overlap with the *keywords* in one of the
211 Medical/ROI texts for that chunk, we map the image to the medical/ROI text. Example. *keywords:*
212 *psammoma bodies*, match with *raw_keyword: psammoma bodies* within the ASR-corrected text
213 ‘*Meningiomas typically have a meningothelial pattern with lobular-like arrangements and psammoma*
214 *bodies.*’ Refer to Figure 7 and Figure 15 in the Appendix for a detailed explanation of the method and
215 examples of aligned image and text.

216 3.2 QUILT-1M: Combining QUILT with other histopathology data sources

217 To create QUILT-1M, we expanded QUILT by adding other disparate histopathology image-text
218 open-access sources: LAION, Twitter, and PubMed.

219 **PubMed Open Access Articles.** We searched the PubMed open-access from 2010-2022, extracting
220 59,371 histopathology image-text pairs, using our histopathology classifier and multi-plane figure
221 cropping algorithm. The images are categorized into (1) images that are fully histopathology, (2)
222 multi-plane images that contain histopathology sub-figures, and (3) histopathology sub-figures
223 cropped from (1) and (2). See Figure 16, and section A.2.1 in the Appendix.

224 **Histopathology Image Retrieval from LAION.** The Large-scale Artificial Intelligence Open Net-
225 work (LAION-5B) [55] curated over 5 billion pairs of images and text from across the Internet,
226 including a substantial volume of histopathology-related data. We tapped into this resource by
227 retrieving 22,682 image and text pairs. See section A.2.2 in the Appendix.

228 **Twitter Data from OpenPath.** We utilized a list of tweets curated by Huang et al. [25], which totaled
229 up to 55,000 unique tweets and made up 133, 511 unique image-text pairs. This exhibits a one-to-
230 many relationship where many images were matched with multiple captions; this differentiated our
231 work from the OpenPath approach. To maintain comparability, we followed their text pre-processing
232 pipeline [25]. See section A.2.3 in the Appendix.

233 **3.3 Quality**

234 To evaluate our pipeline’s performance, we assess several aspects. First, we calculate the precision of
 235 our LLM’s corrections by dividing the number of *conditioned* misspelled errors replaced (i.e., passed
 236 the UMLS check) by the total number of *conditioned* misspelled words found, yielding an average of
 237 57.9%. We also determined the *unconditioned* precision of the LLM, similar to the previous step, and
 238 found it to be 13.8%. Therefore, we replace our detected incorrect words with the LLM’s correction
 239 57.9% of the time, and 13.8% of the time we replace the LLM’s detected errors with its correction
 240 (see Table 11 in the Appendix). To estimate the ASR model’s transcription performance, we compute
 241 the total number of errors replaced (both conditioned and unconditioned) and divide it by the total
 242 number of words in each video, resulting in an average ASR error rate of 0.79%. To assess the LLM’s
 243 sub-pathology classification, we manually annotated top-k ($k = 1, 2, 3$) sub-pathology types for 100
 244 random videos from our dataset. The LLM’s accuracy for top-3, top-2, and top-1 was 94.9%, 91.9%,
 245 and 86.8%, respectively. Also note that, by prompting the LLM to extract only medically relevant
 246 text, we further eliminate identifiable information, such as clinic addresses, from our dataset.

247 **3.4 Final dataset statistics**

248 We collected QUILT, from 4504 narrative videos spanning over 1087 hours with over 437K unique
 249 images with 802K associated text pairs. The mean length of the text captions is 22.76 words, and 8.68
 250 words for ROI text, with an average of 1.74 medical sentences per image (max=5.33, min=1.0). Our
 251 dataset spans a total of 1.469M UMLS entities from those mentioned in the text (with 28.5K unique).
 252 The images span varying microscopic magnification scales (0-10x, 10-20x, 20-40x), obtaining (280K,
 253 75K, 107K) images from each scale respectively with an average height and width of 882 x 1468
 254 pixels, as we leverage the max image resolution of videos. Figure 14 (a, c) in the Appendix plots our
 255 dataset’s diversity across multiple histopathology sub-domains. This plot shows that the captions
 256 cover histopathology-relevant medical subtypes: findings, concepts, organs, neoplastic processes,
 257 cells, diseases, and a mix of laboratory and diagnostic procedures. Overall, across all 127 UMLS
 258 semantic types, our entities cover 76.2% of medically-related semantic types (e.g., findings, disease,
 259 or syndrome) and 23.75% non-medical (e.g., geographic area, governmental or regulatory activity).

260 **4 QUILTNET: Experiments training with QUILT-1M**

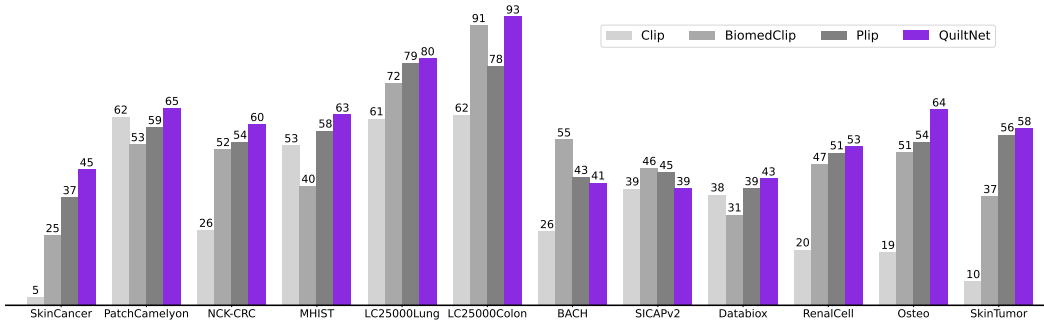


Figure 3: QUILTNET, outperforms out-of-domain CLIP baseline and state-of-the-art histopathology models across 12 zero-shot tasks, covering 8 different sub-pathologies (accuracy percentage provided).

We use the Contrastive Language-Image Pre-training (CLIP) objective [49] to pretrain QUILTNET using QUILT-1M. CLIP takes a batch of N (image, text) pairs and optimizes a contrastive objective to create a joint embedding space. The optimization process involves concurrent training of both image and text encoders to increase the cosine similarity of embeddings from aligned pairs, while decreasing it for unaligned pairs. The objective is minimized via the InfoNCE loss, expressed as:

$$\mathcal{L} = -\frac{1}{2N} \left(\sum_{i=1}^N \log \frac{e^{\cos(I_i, T_i)}}{\sum_{j=1}^N e^{\cos(I_i, T_j)}} + \sum_{i=1}^N \log \frac{e^{\cos(I_i, T_i)}}{\sum_{j=1}^N e^{\cos(I_j, T_i)}} \right)$$

261 where I_i and T_i are the embeddings for the aligned i -th image and text, respectively. For the image
 262 encoder, we use both ViT-B/32 and ViT-B/16 architectures [16]. For the text encoder, we use GPT-
 263 2 [48] with a context length of 77, and PubmedBert [71]. We train QUILTNET by finetuning an
 264 OpenAI pre-trained CLIP model [49] on QUILT-1M to enhance its performance in histopathology.
 265 Once finetuned, we conduct experiments on two types of downstream tasks: image classification (zero-
 266 shot and linear probing) and cross-modal retrieval (zero-shot). We also compare the performance of
 267 fine-tuning a pre-trained CLIP model versus training it from scratch.

268 **Downstream histopathology datasets.** We evaluate the utility of QUILTNET on 13 downstream
 269 datasets: **PatchCamelyon** [61] contains histopathology scans of lymph node sections labeled for
 270 metastatic tissue presence as a binary label. **NCT-CRC-HE-100K** [33] consists of colorectal cancer
 271 images and is categorized into cancer and normal tissue. For **SICAPv2** [57] the images are labeled
 272 as non-cancerous, Grade 3-5. **Databiox** [8] consists of invasive ductal carcinoma cases of Grades
 273 I-III. **BACH** [4] consists of breast tissues labeled as normal, benign, in-situ, and invasive carcinoma.
 274 **Osteo** [5] is a set of tissue patches representing the heterogeneity of osteosarcoma. **RenalCell** [10]
 275 contains tissue images of clear-cell renal cell carcinoma annotated into five tissue texture types.
 276 **SkinCancer** [36] consists of tissue patches from skin biopsies of 12 anatomical compartments
 277 and 4 neoplasms that make up the **SkinTumor** Subset. **MHIST** [64] contains tissue patches from
 278 Formalin-Fixed Paraffin-Embedded WSIs of colorectal polyps. **LC25000** [9], which we divide into
 279 **LC25000 (Lung)** and **LC25000 (Colon)**, contains tissue of lung and colon adenocarcinomas. For
 280 more details on the datasets refer to C.1 and Table 15 in the Appendix.

Table 1: **Linear probing.** Classification results, denoted as accuracy % (standard deviation). Came-
 lyon denotes the PatchCamelyon dataset. Supervised results are from each dataset’s SOTA models.

Dataset	%shot	ViT-B/32			ViT-B/16			
		CLIP GPT/77	PLIP GPT/77	QUILTNET GPT/77	CLIP GPT/77	QUILTNET GPT/77	BiomedCLIP PMB/256	QUILTNET PMB/256
NCT-CRC [33] (94.0)	1	91.0 (0.10)	93.75 (0.09)	94.64 (0.22)	90.96 (0.10)	93.36 (0.23)	92.14 (0.12)	93.55 (0.25)
	10	92.02 (1.30)	93.83 (0.06)	95.30 (0.03)	92.58 (0.12)	93.85 (0.04)	92.90 (0.07)	93.72 (0.08)
	100	91.83 (0.01)	94.16 (0.01)	95.22 (0.01)	92.26 (0.09)	93.76 (0.02)	92.97 (0.05)	93.60 (0.01)
Camelyon [61] (97.5)	1	80.38 (0.16)	87.26 (0.23)	87.62 (0.35)	80.28 (0.20)	84.78 (0.14)	83.63 (0.44)	83.48 (0.18)
	10	82.67 (0.19)	87.48 (0.08)	87.55 (0.03)	82.20 (0.04)	86.77 (0.09)	84.18 (0.15)	84.42 (0.10)
	100	82.80 (0.01)	87.34 (0.01)	87.48 (0.01)	82.55 (0.02)	86.81 (0.04)	84.23 (0.01)	84.44 (0.02)
SkinCancer [36] (93.3)	1	84.27 (0.22)	91.07 (0.25)	90.93 (0.25)	85.62 (0.16)	88.29 (0.15)	87.53 (0.21)	88.06 (0.20)
	10	89.0 (0.02)	93.39 (0.05)	92.99 (0.02)	90.28 (0.01)	91.20 (0.0)	89.23 (0.03)	90.03 (0.02)
	100	89.02 (0.02)	93.29 (0.01)	93.03 (0.02)	90.29 (0.03)	91.20 (0.0)	89.16 (0.02)	89.91 (0.01)
SICAPv2 [57] (67.0)	1	52.45 (2.41)	65.76 (2.65)	69.92 (1.02)	56.01 (0.66)	66.86 (1.16)	69.43 (1.03)	68.49 (1.06)
	10	62.24 (0.65)	69.23 (0.43)	74.14 (0.38)	63.70 (0.69)	72.37 (0.65)	71.61 (0.31)	72.48 (0.42)
	100	65.75 (0.16)	73.0 (0.14)	75.48 (0.12)	68.74 (0.10)	74.14 (0.16)	74.57 (0.04)	74.60 (0.17)

281 **Results using zero-shot learning.** Given the vast diversity of cancer sub-types in histopathology,
 282 it is critical that a model maintains comprehensive understanding without requiring specific data
 283 for retraining. Thus, we evaluate our model’s zero-shot performance against three state-of-the-art
 284 models: CLIP, BiomedCLIP, and PLIP. Our model demonstrates superior performance, as illustrated
 285 in Figure 3, where it outperforms the other models in all but two datasets, in which BiomedCLIP
 286 performs marginally better. See Table 17 for UMap visualizations and Figure 17 for cross-modal
 287 attention visualization comparison in the Appendix. The prompts used for these evaluations are
 288 presented in Table 16 in the Appendix. To ensure a fair comparison with BiomedCLIP, which uses a
 289 ViT-B/16 and PMB/256 (pre-trained with [71]), we trained three different variants of our model. For
 290 detailed insights into the results, please refer to Table 14 in the Appendix.

291 **Results using linear probing.** We assess the few-shot and full-shot performance of our model
 292 by conducting linear probing with 1%, 10%, and 100% of the training data, sampled with three

293 different seeds; we report the average accuracy and their standard deviation in Table 1. We deploy
 294 our evaluation across four distinct datasets, specifically those with dedicated training and testing sets
 295 among our external datasets. Remarkably, our model, utilizing the ViT-B/32 architecture with GPT/77,
 296 outperforms its counterparts, BiomedCLIP, PLIP, and CLIP, in most datasets. On the NCT-CRC and
 297 SICAPv2 datasets, our model surpasses even the fully supervised performance using only 1% of the
 298 labels. Also, note that for some results 10% does better than 100%; this is because we are sampling
 299 from each class equally, and thus the 10% subset contains a more balanced training set than 100%,
 300 for datasets that are very imbalanced, resulting in sub-optimal performance at 100%.

301 **Results using cross-modal retrieval.** In our study, we evaluate cross-modal retrieval efficacy by
 302 examining both zero-shot text-to-image and image-to-text retrieval capabilities. We accomplish this by
 303 identifying the nearest neighbors for each modality and then determining whether the corresponding
 304 pair is within the top N nearest neighbors, where $N \in \{1, 50, 200\}$. Our experiments are conducted
 305 on two datasets: our holdout dataset from QUILT-1M and the ARCH dataset. Results are in Table 2.

Table 2: Cross-modal retrieval results on the QUILT-1M holdout set and ARCH dataset. In each cell, the results are displayed in the format (%/%), with QUILT-1M holdout results on the left and ARCH results on the right. The best-performing results are highlighted in bold text.

model	config	Text-to-Image (%)			Image-to-Text (%)		
		R@1	R@50	R@200	R@1	R@50	R@200
CLIP	ViT-B/32 GPT/77	0.49/0.07	4.73/2.42	10.15/7.21	0.39/0.05	3.99/2.52	8.80/7.22
PLIP	ViT-B/32 GPT/77	1.05/0.56	10.79/13.10	21.80/29.85	0.87/0.74	11.04/13.75	21.63/29.46
QUILTNET	ViT-B/32 GPT/77	1.17/1.41	16.31/19.87	31.99/39.13	1.24/1.35	14.89/19.20	28.97/38.57
CLIP	ViT-B/16 GPT/77	0.83/0.09	5.63/2.73	11.26/8.72	0.66/0.13	5.02/3.09	10.82/9.04
QUILTNET	ViT-B/16 GPT/77	2.42/1.29	22.38/20.30	41.05/40.89	2.00/1.01	21.66/16.18	39.29/34.15
BiomedCLIP	ViT-B/16(224) PMB/256	4.34/ 8.89	14.99/53.24	25.62/71.43	3.88/ 9.97	13.93/52.13	23.53/68.47
QUILTNET	ViT-B/16(224) PMB/256	6.20/8.77	30.28/55.14	50.60/77.64	6.27/9.85	31.06/53.06	50.86/73.43

306 5 Discussion

307 **Limitations.** Despite the promising results, QUILT was curated using several handcrafted algorithms
 308 and LLMs. Such curation methods, while effective, introduce their own biases and errors. For
 309 instance, our histopathology classifier had occasional false positives ($\approx 5\%$) confirmed by human
 310 evaluation. Occasionally, ASR can misinterpret a medical term and transcribe it as a different existing
 311 term, such as transcribing 'serous carcinoma' as 'serious carcinoma'. Unfortunately, such errors
 312 are not rectifiable using our current pipeline (see Table 11 in the Appendix). While not directly a
 313 limitation of our dataset, training a CLIP model trained from scratch underperformed compared to
 314 fine-tuning a pre-trained CLIP (see Table 14 in the Appendix). This suggests that a million image-text
 315 pairs may still not be sufficient. Future works may explore other self-supervised objectives.

316 **Data Collection and Societal Biases** Aligning in strategies with [69], we release QUILT derived from
 317 public videos, taking structured steps to limit privacy and consent harms (see A.5 in the Appendix).
 318 Complying with YouTube's privacy policy, we only provide video IDs, allowing users to opt-out
 319 of our dataset. Researchers can employ our pipeline to create QUILT. Regarding societal biases,
 320 a significant portion of our narrators originate from western institutions, a situation that is further
 321 amplified by our focus on English-only videos. Consequently, QUILTNET may exhibit inherent
 322 biases, potentially performing better on data associated with these demographics, while possibly
 323 underperforming when applied to other cultural or linguistic groups.

324 **Conclusion.** We introduced QUILT-1M, the largest open-sourced histopathology dataset to date.
 325 Empirical results validate that pre-training using QUILT is valuable, outperforming larger state-of-
 326 the-art models like BiomedCLIP across various sub-pathology types and tasks including zero-shot,
 327 few-shot, full-shot, and cross-modal retrieval. We established a new state-of-the-art in zero-shot,
 328 linear probing, and cross-modal retrieval tasks in the field of Histopathology.

329 Acknowledgments

330 Research reported in this study was supported by the National Cancer Institute under Awards No. R01
331 CA15130, R01 CA225585, and R01 CA201376 and the Office of the Assistant Secretary of Defense
332 for Health Affairs through the Melanoma Research Program under Awards No. W81XWH-20-1-0797
333 and W81XWH-20-1-0798. Opinions, conclusions, and recommendations are those of the authors.

334 References

- 335 [1] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag. Large language models are
336 zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.
- 337 [2] M. Amith, L. Cui, K. Roberts, H. Xu, and C. Tao. Ontology of consumer health vocabulary:
338 providing a formal and interoperable semantic resource for linking lay language and medical
339 terminology. In *2019 IEEE International Conference on Bioinformatics and Biomedicine*
340 (*BIBM*), pages 1177–1178. IEEE, 2019.
- 341 [3] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod. Large-scale query-by-image video
342 retrieval using bloom filters. *arXiv preprint arXiv:1604.07939*, 2016.
- 343 [4] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami,
344 M. Prastawa, M. Chan, M. Donovan, et al. Bach: Grand challenge on breast cancer histology
345 images. *Medical image analysis*, 56:122–139, 2019.
- 346 [5] H. B. Arunachalam, R. Mishra, O. Daescu, K. Cederberg, D. Rakheja, A. Sengupta, D. Leonard,
347 R. Hallac, and P. Leavey. Viable and necrotic tumor assessment from whole slide images of
348 osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706,
349 2019.
- 350 [6] B. Baker, I. Akkaya, P. Zhokov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro,
351 and J. Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos.
352 *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.
- 353 [7] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminol-
354 ogy. *Nucleic Acids Res.*, 32(Database-Issue):267–270, 2004. URL <http://dblp.uni-trier.de/db/journals/nar/nar32.html#Bodenreider04>.
- 355 [8] H. Bolhasani, E. Amjadi, M. Tabatabaeian, and S. J. Jassbi. A histopathological image dataset
356 for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked*, 19:100341,
357 2020.
- 358 [9] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mas-
359 torides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint*
360 *arXiv:1912.12142*, 2019.
- 361 [10] O. Brummer, P. Polonen, S. Mustjoki, and O. Bruck. Integrative analysis of histological textures
362 and lymphocyte infiltration in renal cell carcinoma using deep learning. *bioRxiv*, pages 2022–08,
363 2022.
- 364 [11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging
365 properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international*
366 *conference on computer vision*, pages 9650–9660, 2021.
- 367 [12] R. J. Chen, M. Y. Lu, W.-H. Weng, T. Y. Chen, D. F. Williamson, T. Manz, M. Shady, and
368 F. Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole
369 slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*,
370 pages 4015–4025, 2021.
- 371

- 372 [13] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood. Scaling
373 vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings*
374 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155,
375 2022.
- 376 [14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning
377 of visual representations. In *International conference on machine learning*, pages 1597–1607.
378 PMLR, 2020.
- 379 [15] B. Ding, C. Qin, L. Liu, L. Bing, S. Joty, and B. Li. Is gpt-3 a good data annotator? *arXiv*
380 *preprint arXiv:2212.10450*, 2022.
- 381 [16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani,
382 M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for
383 image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- 384 [17] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier. An open-source speaker gender
385 detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing*
386 *(ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.
- 387 [18] S. Eslami, G. de Melo, and C. Meinel. Does clip benefit visual question answering in the
388 medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*,
389 2021.
- 390 [19] G. Fragoso, S. de Coronado, M. Haber, F. Hartel, and L. Wright. Overview and utilization of
391 the nci thesaurus. *Comparative and functional genomics*, 5(8):648–654, 2004.
- 392 [20] J. Gamper and N. Rajpoot. Multiple instance captioning: Learning representations from
393 histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer*
394 *Vision and Pattern Recognition*, pages 16549–16559, 2021.
- 395 [21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford.
396 Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- 397 [22] F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd-workers for text-annotation
398 tasks. *arXiv preprint arXiv:2303.15056*, 2023.
- 399 [23] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual
400 question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- 401 [24] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local
402 representation learning framework for label-efficient medical image recognition. In *Proceedings*
403 *of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.
- 404 [25] Z. Huang, F. Bianchi, M. Yuksekgonul, T. Montine, and J. Zou. Leveraging medical twitter to
405 build a visual–language foundation model for pathology ai. *bioRxiv*, pages 2023–03, 2023.
- 406 [26] E. Hussain, L. B. Mahanta, H. Borah, and C. R. Das. Liquid based-cytology pap smear dataset
407 for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data in brief*,
408 30:105589, 2020.
- 409 [27] W. O. Ikezogwo, M. S. Seyfioglu, and L. Shapiro. Multi-modal masked autoencoders learn
410 compositional histopathological representations. *arXiv preprint arXiv:2209.01534*, 2022.
- 411 [28] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar,
412 H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL
413 <https://doi.org/10.5281/zenodo.5143773>.

- 414 [29] K. Jobin, A. Mondal, and C. Jawahar. Docfigure: A dataset for scientific document figure
415 classification. In *2019 International Conference on Document Analysis and Recognition*
416 *Workshops (ICDARW)*, volume 1, pages 74–79. IEEE, 2019.
- 417 [30] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G.
418 Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of
419 labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
- 420 [31] S. Jupp, J. Malone, T. Burdett, J.-K. Heriche, E. Williams, J. Ellenberg, H. Parkinson, and
421 G. Rustici. The cellular microscopy phenotype ontology. *Journal of biomedical semantics*, 7:
422 1–8, 2016.
- 423 [32] Z. Karishma. Scientific document figure extraction, clustering and classification. 2021.
- 424 [33] J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer
425 and healthy tissue. *Zenodo*10, 5281, 2018.
- 426 [34] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth
427 a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The*
428 *Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.
- 429 [35] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot
430 reasoners. *arXiv preprint arXiv:2205.11916*, 2022.
- 431 [36] K. Kriegsmann, F. Lobers, C. Zgorzelski, J. Kriegsmann, C. Janßen, R. R. Meliß, T. Muley,
432 U. Sack, G. Steinbuss, and M. Kriegsmann. Deep learning for the detection of anatomical tissue
433 structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in*
434 *Oncology*, 12, 2022.
- 435 [37] J. Liu, Q. Wang, H. Fan, S. Wang, W. Li, Y. Tang, D. Wang, M. Zhou, and L. Chen. Automatic
436 label correction for the accurate edge detection of overlapping cervical cells. *arXiv preprint*
437 *arXiv:2010.01919*, 2020.
- 438 [38] S. Liu, C. Zhu, F. Xu, X. Jia, Z. Shi, and M. Jin. Bci: Breast cancer immunohistochemical
439 image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on*
440 *Computer Vision and Pattern Recognition*, pages 1815–1824, 2022.
- 441 [39] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings*
442 *of International Conference on Computer Vision (ICCV)*, December 2015.
- 443 [40] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In
444 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages
445 11976–11986, 2022.
- 446 [41] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint*
447 *arXiv:1711.05101*, 2017.
- 448 [42] N. Marini, S. Marchesin, S. Otálora, M. Wodzinski, A. Caputo, M. Van Rijthoven, W. As-
449 wolinskiy, J.-M. Bokhorst, D. Podareanu, E. Petters, et al. Unleashing the potential of digital
450 pathology data by training computer-aided diagnosis models without human annotations. *NPJ*
451 *digital medicine*, 5(1):102, 2022.
- 452 [43] D. Morris, E. Müller-Budack, and R. Ewerth. Slideimages: a dataset for educational image
453 classification. In *Advances in Information Retrieval: 42nd European Conference on IR Research,*
454 *ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 289–296.
455 Springer, 2020.

- 456 [44] M. Neumann, D. King, I. Beltagy, and W. Ammar. ScispaCy: Fast and Robust Models for
457 Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop
458 and Shared Task*, pages 319–327, Florence, Italy, Aug. 2019. Association for Computational
459 Linguistics. doi: 10.18653/v1/W19-5034. URL [https://www.aclweb.org/anthology/
460 W19-5034](https://www.aclweb.org/anthology/W19-5034).
- 461 [45] N. F. Noy, M. A. Musen, J. L. Mejino Jr, and C. Rosse. Pushing the envelope: challenges
462 in a frame-based representation of human anatomy. *Data & Knowledge Engineering*, 48(3):
463 335–359, 2004.
- 464 [46] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology objects in context
465 (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting
466 and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint Interna-
467 tional Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in
468 Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages
469 180–189. Springer, 2018.
- 470 [47] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari. Connecting vision and
471 language with localized narratives. In *ECCV*, 2020.
- 472 [48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are
473 unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 474 [49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell,
475 P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision.
476 In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- 477 [50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech
478 recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.
- 479 [51] R. J. Roberts. Pubmed central: The genbank of the published literature, 2001.
- 480 [52] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual
481 documents. *Text mining: applications and theory*, pages 1–20, 2010.
- 482 [53] T. Santos, A. Tariq, S. Das, K. Vayalpati, G. H. Smith, H. Trivedi, and I. Banerjee. Pathologybert-
483 pre-trained vs. a new transformer language model for pathology domain. *arXiv preprint
484 arXiv:2205.06885*, 2022.
- 485 [54] P. N. Schofield, J. P. Sundberg, B. A. Sundberg, C. McKerlie, and G. V. Gkoutos. The mouse
486 pathology ontology, mpath; structure and applications. *Journal of biomedical semantics*, 4(1):
487 1–8, 2013.
- 488 [55] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes,
489 A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next
490 generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- 491 [56] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual
492 explanations from deep networks via gradient-based localization. arxiv 2016. *arXiv preprint
493 arXiv:1610.02391*.
- 494 [57] J. Silva-Rodríguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo. Going deeper through
495 the gleason scoring scale: An automatic end-to-end system for histology prostate grading and
496 cribriform pattern detection. *Computer methods and programs in biomedicine*, 195:105637,
497 2020.
- 498 [58] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach.
499 Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer
500 vision and pattern recognition*, pages 8317–8326, 2019.

- 501 [59] H. Singh and M. L. Graber. Improving diagnosis in health care—the next imperative for patient
502 safety. *The New England journal of medicine*, 373(26):2493–2495, 2015.
- 503 [60] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani,
504 H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *arXiv*
505 *preprint arXiv:2212.13138*, 2022.
- 506 [61] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for
507 digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI*
508 *2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings,*
509 *Part II 11*, pages 210–218. Springer, 2018.
- 510 [62] P. Voigtlaender, S. Changpinyo, J. Pont-Tuset, R. Soricut, and V. Ferrari. Connecting Vision
511 and Language with Video Localized Narratives. In *IEEE/CVF Conference on Computer Vision*
512 *and Pattern Recognition*, 2023.
- 513 [63] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng. Want to reduce labeling cost? gpt-3 can help.
514 *arXiv preprint arXiv:2108.13487*, 2021.
- 515 [64] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, N. Tomita,
516 L. Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in*
517 *Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021,*
518 *Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.
- 519 [65] P. Weitz, M. Valkonen, L. Solorzano, C. Carr, K. Kartasalo, C. Boissin, S. Koivukoski,
520 A. Kuusela, D. Rasic, Y. Feng, et al. Acrobat—a multi-stain breast cancer histological whole-
521 slide-image data set from routine diagnostics for computational pathology. *arXiv preprint*
522 *arXiv:2211.13621*, 2022.
- 523 [66] P. S. Wright, K. A. Briggs, R. Thomas, G. F. Smith, G. Maglennon, P. Mikulskis, M. Chapman,
524 N. Greene, B. U. Phillips, and A. Bender. Statistical analysis of preclinical inter-species
525 concordance of histopathological findings in the etox database. *Regulatory Toxicology and*
526 *Pharmacology*, 138:105308, 2023.
- 527 [67] H. Wu, W. Wang, Y. Wan, W. Jiao, and M. Lyu. Chatgpt or grammarly? evaluating chatgpt on
528 grammatical error correction benchmark. *arXiv preprint arXiv:2303.13648*, 2023.
- 529 [68] W. Wu, S. Mehta, S. Nofallah, S. Knezevich, C. J. May, O. H. Chang, J. G. Elmore, and
530 L. G. Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:
531 163526–163541, 2021.
- 532 [69] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi. Merlot:
533 Multimodal neural script knowledge models. *Advances in Neural Information Processing*
534 *Systems*, 34:23634–23651, 2021.
- 535 [70] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and
536 Y. Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In
537 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages
538 16375–16387, 2022.
- 539 [71] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong,
540 et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv*
541 *preprint arXiv:2303.00915*, 2023.
- 542 [72] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of
543 medical visual representations from paired images and text. In *Machine Learning for Healthcare*
544 *Conference*, pages 2–25. PMLR, 2022.
- 545 [73] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image
546 database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,
547 2017.

548 **Checklist**

- 549 1. For all authors...
- 550 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's
551 contributions and scope? [Yes]
- 552 (b) Did you describe the limitations of your work? [Yes]
- 553 (c) Did you discuss any potential negative societal impacts of your work? [Yes]
- 554 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
555 them? [Yes]
- 556 2. If you are including theoretical results...
- 557 (a) Did you state the full set of assumptions of all theoretical results? [N/A]
- 558 (b) Did you include complete proofs of all theoretical results? [N/A]
- 559 3. If you ran experiments...
- 560 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
561 mental results (either in the supplemental material or as a URL)? [Yes]
- 562 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
563 were chosen)? [Yes]
- 564 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
565 ments multiple times)? [Yes]
- 566 (d) Did you include the total amount of compute and the type of resources used (e.g., type
567 of GPUs, internal cluster, or cloud provider)? [Yes]
- 568 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 569 (a) If your work uses existing assets, did you cite the creators? [Yes]
- 570 (b) Did you mention the license of the assets? [Yes]
- 571 (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
- 572 (d) Did you discuss whether and how consent was obtained from people whose data you're
573 using/curating? [Yes]
- 574 (e) Did you discuss whether the data you are using/curating contains personally identifiable
575 information or offensive content? [Yes]
- 576 5. If you used crowdsourcing or conducted research with human subjects...
- 577 (a) Did you include the full text of instructions given to participants and screenshots, if
578 applicable? [N/A]
- 579 (b) Did you describe any potential participant risks, with links to Institutional Review
580 Board (IRB) approvals, if applicable? [N/A]
- 581 (c) Did you include the estimated hourly wage paid to participants and the total amount
582 spent on participant compensation? [N/A]

613 To identify videos containing medical content, we employ a keyframe extraction process with a
 614 specific threshold to determine the minimum visual change required to trigger keyframes. For a new
 615 video, the thresholds for keyframe extraction are determined by linearly interpolating between the
 616 lowest threshold, 0.008 (5-minute video) and the highest 0.25 (200-minute video). Following the
 617 keyframe extraction process, we utilize a histopathology image classifier to identify histopathology
 618 content within the extracted keyframes. See A.3 for more details. To identify narrative-style videos,
 619 we randomly select a $\min(\text{num_of_histo_scene_frames}, 20)$ keyframes from a video and utilize a
 620 pre-trained CLIP⁸ (ViT-B-32) model to embed and compute a cosine similarity on the next three
 621 keyframes. If all three have similarity scores \geq a threshold of 0.9, we count the video as a narrative
 622 streak.

623 **Text extraction using ASR and text denoising.** Another challenge involves automatic speech
 624 recognition (ASR), as YouTube captions are often inadequate for medical vocabulary. To address this
 625 issue, we employed the Large-V2 open-source Whisper model [50] for speech-to-text conversion.
 626 However, general-purpose ASR models like Whisper can misinterpret medical terms, particularly
 627 when the speaker’s voice is choppy or accented. There are no straightforward trivial solutions due
 628 to: **1)** the absence of openly available medical ASR models or data for fine-tuning in the medical
 629 domain; **2)** the inadequacy of medical named entity recognition models in detecting transcription
 630 errors, because these models are typically trained on correctly spelled words; **3)** the ineffectiveness
 631 of methods like semantically searching over a medical glossary, such as UMLS, which only prove
 632 effective when the erroneous text has significant similarity to the correct terms; and **4)** the inability of
 633 simpler methods like finding the longest common substring, which might work in finding a match in
 634 the glossary/ontology for replacement, but cannot identify the wrong words/phrases in the first place.
 635 To rectify ASR errors, we employed UMLS (a knowledge database) and a LLM (GPT-3.5). This,
 636 however, introduces a new challenge of identifying incorrectly transcribed words and determining
 637 which words were mistakenly "corrected" and correctly formatted by the LLM after error correction
 638 and resolving unintended parsing errors [1]. See Figure 9 for LLM prompt examples for ASR
 639 correction and medical and ROI text extraction from the corrected ASR text. Refer to Table 11 for
 640 error examples of ASR correction using the LLM.

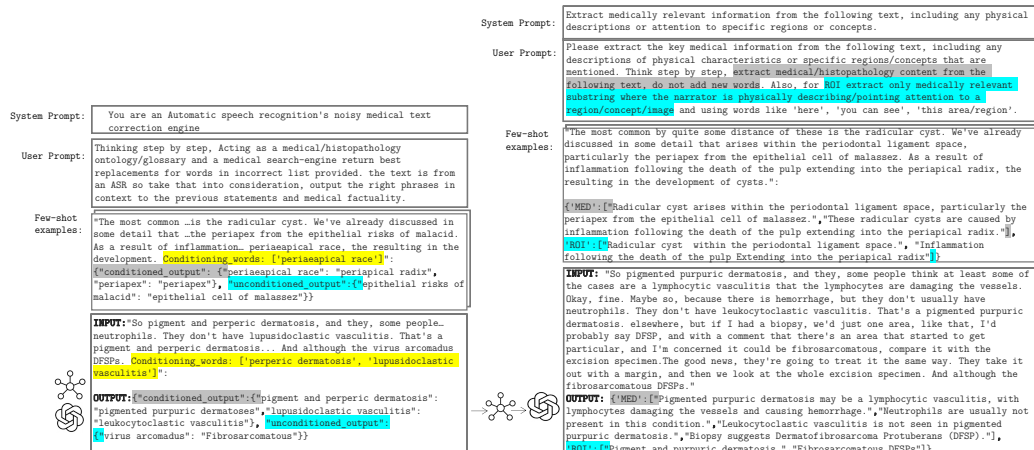


Figure 9: Prompting the LLM by providing few-shot examples to perform the following tasks: (Left) correcting noisy ASR text within its context. We highlight the probable list of misspelled keywords in yellow and their corrections by the LLM in gray. Additional missed errors/misspelled entries identified by the LLM are highlighted in blue. (Right) extracting medical (MED) and ROI text from a given text. We highlight the definition of medical and ROI text in blue and gray respectively.

641 **Image frame extraction and denoising.** The image processing aspect of this task adds to its complexity,
 642 as it requires static frame detection, quality control for frames, and histology magnification

⁸<https://huggingface.co/sentence-transformers/clip-ViT-B-32>

Table 11: 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 sarcoidal to me. The reason I say that is they, there's a kind of positive 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 differential hypersensium nitose , but I would be more worried about sarcoidosis.	differential hypersensium nitose : hypersensitivity pneumonitis, cranialomas : granulomas	positive : paucity
LLM	high-larbidia-stinal lymphadenocathy lymphing-giatic pattern distribution	returns hilar lymphadenopathy instead of a more appropriate hilar mediastinal lymphadenopathy	returns lymphatic pattern distribution instead of a more appropriate lymphangitic pattern distribution
Incomplete UMLS checker	... picnotic	-	LLM correctly returns pyknotic however, UMLS(2020) does not have the word <i>pyknotic</i> if fails to pass the UMLS check.

643 classification. Each model utilized it these steps introduces its own biases and errors. We extract
 644 time-intervals (*chunks*) from each video from which we extract representative image(s). For each
 645 of the extracted *chunks* (t_n, t_{n+1}), the static chunk detection algorithm 1 is used to extract sub-
 646 time-intervals with static frames within the chunk. If found, we save the median (in pixel space to
 647 prevent blurry outputs) of the stable frames, else (i.e no stable duration of frames) we leverage the
 648 structural similarity index (SSIM) method on histopathology key-frames to find the most dissimilar
 649 histopathology image to make up the representative images for the chunk, essentially de-duplicating
 650 the frames. Figure 13 demonstrates this process.



Figure 13: 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.

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

```

651 **Aligning both modalities.** The alignment of the images with their corresponding text requires the
652 implementation of unique algorithms. These algorithms are designed to reduce duplicate content
653 and ensure accurate mappings between image and text. See Figures 7 and 8 and Table 9 for a
654 demonstration of image-text alignment process. See Figure 15 for sample images and their
655 corresponding medical and ROI texts and the sub-pathology classification provided by the LLM.

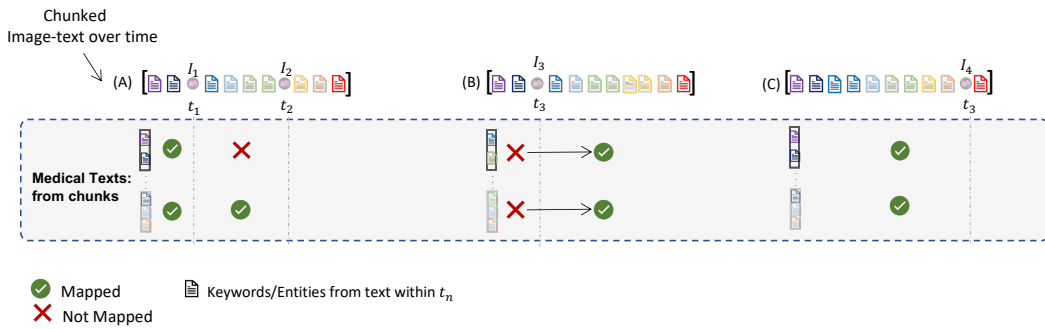


Figure 7: **Overview of use of timing and keywords for Alignment** Images within a video chunk, i.e. {A, B, C}, I_n at t_n are aligned with medical texts extracted within the same chunk. The *raw_keywords* within each example chunk is colour coded to illustrate matches with *keywords* extracted from 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.

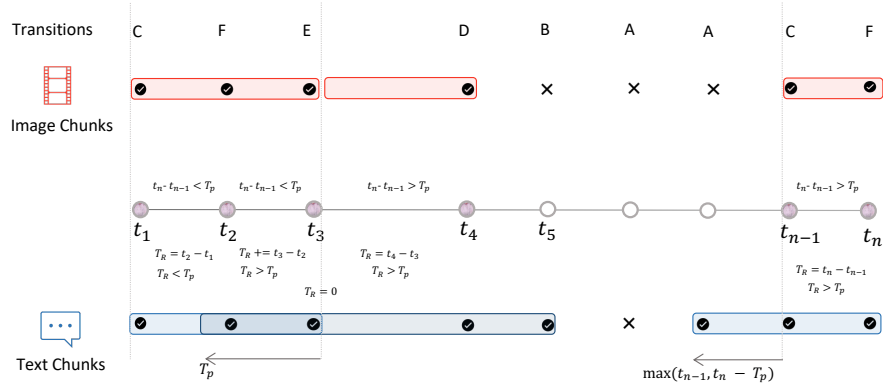


Figure 8: **Video Chunking algorithm illustrate.** With each transition tag explained in Table 9, we leverage predicted histopathology frames at times t_1, \dots, t_n to segment videos into chunks. Chunks at are minimum are T_P in duration, this value is estimated based on the word-per-second of the 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 9: **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

656 A.2 Other data sources

657 A.2.1 PubMed Open Access Articles

658 We searched the PubMed open-access from 2010 – 2022 with keywords (pathology, histopathology,
659 whole-slide image, H&E, and 148 keywords from a histopathology glossary⁹). We utilized Entrez¹⁰
660 to retrieved the top 10,000 most relevant articles for each keyword. This query yielded 109,518 unique
661 articles with PMIDs. We extracted 162,307 images and their corresponding captions. Using our
662 histopathology classifier and cropping multi-plane figures as described in A.4, we extracted 59,371
663 histopathology image and caption pairs with an average caption length of 54.02 tokens. Figure 16
664 demonstrates the pipeline of collecting data from PubMed.

⁹<https://lab-ally.com/histopathology-resources/histopathology-glossary>

¹⁰<http://www.ncbi.nlm.nih.gov/Entrez>

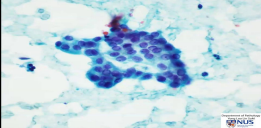
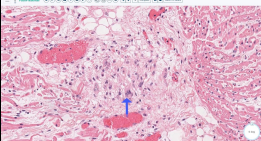
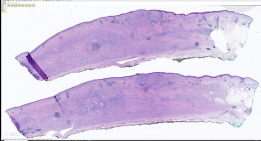
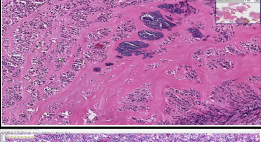
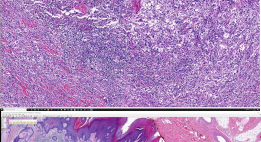
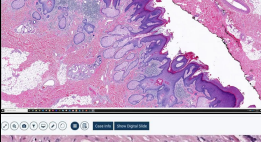
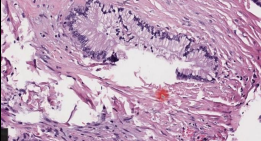
Image	Medical TEXT	ROI Text	Sub-pathology Classification
	['There are clusters of cells with micro-follicular formations.', 'Nuclear pseudo-inclusions, oval nuclei, nuclear grooves, and small nucleoli are present in some cells.']	['clusters of cells', 'micro-follicular formations', 'nuclear pseudo-inclusions', '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 15: 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.

665 A.2.2 Histopathology Image Retrieval from LAION

666 The Large-scale Artificial Intelligence Open Network (LAION-5B) [55] curated over 5 billion pairs
667 of images and text from across the Internet, including a substantial volume of histopathology-related
668 data. We tapped into this resource by retrieving the 3000 most similar LAION samples for each of the
669 1,000 pairs of images and text sampled from PubMed and QUILT, using a CLIP model pre-trained
670 on the LAION data. The retrieval process utilized both image and text embeddings, with cosine
671 similarity serving as the distance metric. Subsequently, we eliminated the duplicate images and
672 removed all non-English pairs from the remaining pairs using LangDetect¹¹. Consequently, the
673 process yielded 22,682 image and text pairs.

¹¹<https://github.com/fedlopez77/langdetect>

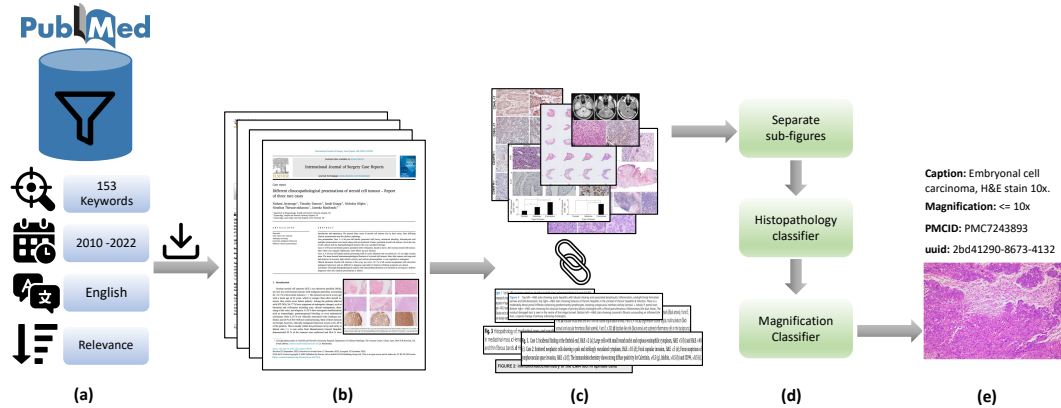


Figure 16: (a) Search PubMed open access database, filter based on keywords, date, language and sort by relevance. (b) Download paper and media for each search result. (c) Extract and pair figures and captions. (d) Separate multi-plane figures, find histopathology images and their magnification. (e) Final result.

674 A.2.3 Twitter Data from OpenPath

675 We utilized a list of tweets curated by Huang et al. [25] which totaled up to 55,000 unique tweets
 676 and 133,511 unique image-text pairs. This exhibits a one-to-many relationship that leans towards the
 677 image side, differentiating our work from the OpenPath approach, where we had one image matching
 678 with multiple captions (as in the case of MS-COCO captions). In order to maintain comparability
 679 with OpenPath, we followed their text pre-processing pipeline given in [25].

680 A.3 Histopathology and Magnification classifier

681 We use an ensemble of three histopathology image classifiers. To ensure robustness, our ensemble
 682 approach consists of two small Conv-NeXt models [40] and one linear classifier fine-tuned with DINO
 683 features [11]. This combination is necessary due to the homogenous appearance of histopathology
 684 images and the risk of false positives from similar pinkish-purple images. One Conv-NeXt model is
 685 trained in detecting non-H&E Immunohistochemistry (IHC) stained tissue images, while the other
 686 models are trained to handle all IHC stains and tissue types. The training data includes eight sub-
 687 groups of the TCGA WSI dataset and a mix of general-domain images, PowerPoint (slide) images,
 688 and scientific figure datasets. See Table 7 for details of these datasets.

689 For the magnification classifier, we finetune a pretrained ConvNeXt-Tiny model [40], with standard
 690 preset hyperparameters for a few epochs and select the best performing model on the validation set.
 691 To generate a training set for the magnification model, TCGA subsets were segmented into patches
 692 using a method similar to [68]. These patches were generated at various magnifications, which were
 693 then categorized into three labels: 0: {1.25x, 2.5x, 5x, 10x}, 1: {20x}, 2: {40x}. The TCGA subsets
 694 were chosen to ensure a diverse representation of tissue morphologies and cancer types, thereby
 695 ensuring robust and comprehensive model training. The model was also trained on cytopathology
 696 microscopy images and various IHC stains beyond H&E to enhance the model’s generalizability
 697 across different conditions. Only the ACROBAT and TCGA datasubsets are preprocessed to divide
 698 the WSIs into patches at various scales.

699 A.4 Support Models, Ontology Databases and Algorithms

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

Table 7: Datasets used to train the histopathology image classifier. [μm per pixel - MPP]

Data Source	Subset	#WSI	#pathces	Train-Test	Magnification	Image-size
TCGA (H&E Stain)	GBM	19			89,022 - 40x	384 x 384
	LUSC	20				
	LIHC	20	169,431	84715-16943	57,671 - 20x	
	SARC	23				
	KIRC	16			16,660 - 10x	
	KICH	4			4,748 - 5x	
	BRCA	17			1,465 - 2.5x	
SKCM	19			466 - 1.25x		
ACROBAT Weitz et al. [65]	H&E KI67	99	50589	28105-22484	(10x, 5x, 2.5x)	384 x 384
	ER , PGR, HER2					
BCI Liu et al. [38]	-	-	4,870		20x (0.46 MPP)	1024 x 1024
CESD Liu et al. [37]	-	-	686		100x/400x	2048 x 1536
Smear Hussain et al. [26]	-	-	963		400x	2048 x 1536
Celeb Liu et al. [39]	-	-	202,599	8,103-1,944	-	-
Places Zhou et al. [73]	-	-	36,550	2,109-1,372	-	-
AI2D Kembhavi et al. [34]	-	-	4,903	0.7-0.3%	-	-
DocFig Jobin et al. [29]	-	-	33,004	0.8-0.2%	-	-
SciFig-pilot Karishma [32]	-	-	1,671	0.8-0.2%	-	-
SlideImages Morris et al. [43]	-	-	8,217	0.8-0.2%	-	-
TextVQA Singh et al. [58]	-	-	28,472	0.8-0.2%	-	-
SlideShare-1M Araujo et al. [3]	-	-	49,801	0.8-0.2%	-	-

702 **Ontology databases.** We employ various ontologies, both specific to histopathology and general
703 ones. Among them are OCHV [2], FMA [45], BCGO¹², NCIT [19], MPATH [54], HPATH [66], and
704 CMPO [31]. These ontologies serve a dual purpose. First, we used histopathology-specific ontologies
705 (HPATH, MPATH, BCGO, and CMPO) to provide words/phrases to condition the LLM, enabling it
706 to identify incorrect words. Second, all ontologies, in conjunction with UMLS, are used to obtain
707 terms or phrases for validating the output of the LLM.

708 **Sub-pathology types.** The list of all 18 sub-pathology types used to prompt LLM on the text
709 classification task are: *Bone, Cardiac, Cyto, Dermato, Endocrine, Gastrointestinal, Genitourinary,*
710 *Gynecologic, Head and Neck, Hemato, Neuro, Ophthalmic, Pediatric, Pulmonary, Renal, Soft*
711 *tissue, and Breast Histopathology.* Figure 12 provides the LLM prompt to retrieve the top three
712 sub-pathology classification based on a given text.

713 **Pre-processing multi-plane figures.** Many figures in academic papers are multi-plane, which means
714 a number of sub-figures (Charts, graphs, histopathology and non-histopathology sub-figures) are
715 placed next to each other to make a larger figure. We extracted individual images from multi-plane
716 figures to create multiple instance bags. To locate boundaries and white gaps between sub-figures,
717 we utilized Sobel filters. Binary thresholding was then used to find the contours surrounding the
718 sub-figures. We employ image size and image ratio thresholds to remove undesirable sub-figures and
719 our histopathology classifier to maintain just histopathology sub-figures. We supply the histological
720 sub-figures individually for this type of figure by appending "_[0-9]+" to the end of the multi-plane
721 figure id. If a figure is divided into more than 5 sub-figures, we preserve the original image to ensure
722 that the resolution of these sub-figures remains reasonable. Figure 11 shows an overview of this
723 pre-processing step in different scenarios of successful and unsuccessful crops.

724 A.5 Privacy preserving steps

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

- 727 • **Reduction of Signal to Noise using a LLM:** To protect the privacy of the dataset, a LLM
728 was utilized to reduce the signal-to-noise ratio. By applying the LLM, irrelevant or sensitive
729 information was masked or removed.

¹²<https://bioportal.bioontology.org/ontologies/BCGO>

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

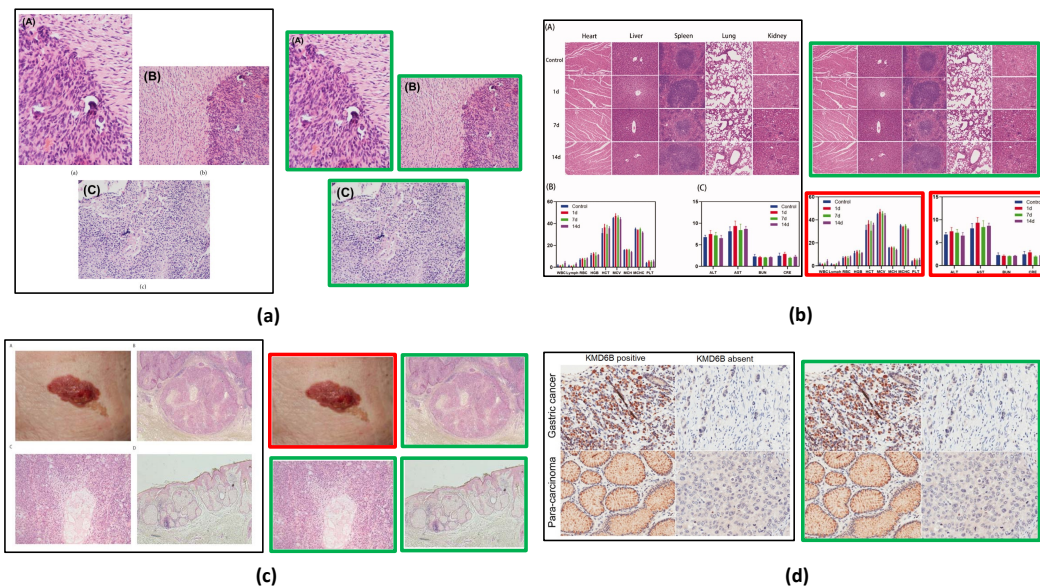


Figure 11: (a), (b), and (c) successfully cropped sub-figures where histopathology images (green box) are kept and non-histopathology (red box) images are removed. (b) histopathology crops are kept as not separated because the individual crops don't meet the size threshold so the original figure is kept. (d) Unsuccessful crop due to minimal gap between sub-figures. Original image is stored.

- 730 • Exclusion of Videos Not Fully in Narrative Style: Videos that did not adhere to a fully
731 narrative style were intentionally left out of the dataset. This step was taken to minimize
732 the risk of including any potentially private or sensitive content that could compromise
733 individuals' privacy.
- 734 • Release of Video IDs and Reconstruction Code: Instead of directly releasing the complete
735 dataset, only video IDs from YouTube were made public. Additionally, the code is provided
736 to enable researchers to recreate the dataset.
- 737 • Collection from Diverse Channels: Data collection was performed from a wide range of
738 sources, including both large and small channels. This approach aimed to decrease the risk
739 of overfitting to specific channel types, thereby mitigating privacy concerns associated with
740 potential biases linked to particular channels.

741 B Exploratory analysis of the collected data

742 In this section, we provide the statistics of the QUILT dataset. Figure 14 illustrates the distribution of
743 data across 18 sub-pathology types, offering a comprehensive analysis of the dataset's text distribution.
744 Moreover, for additional statistical details regarding QUILT, please refer to Table 10, which presents
745 supplementary information on various aspects of the dataset.

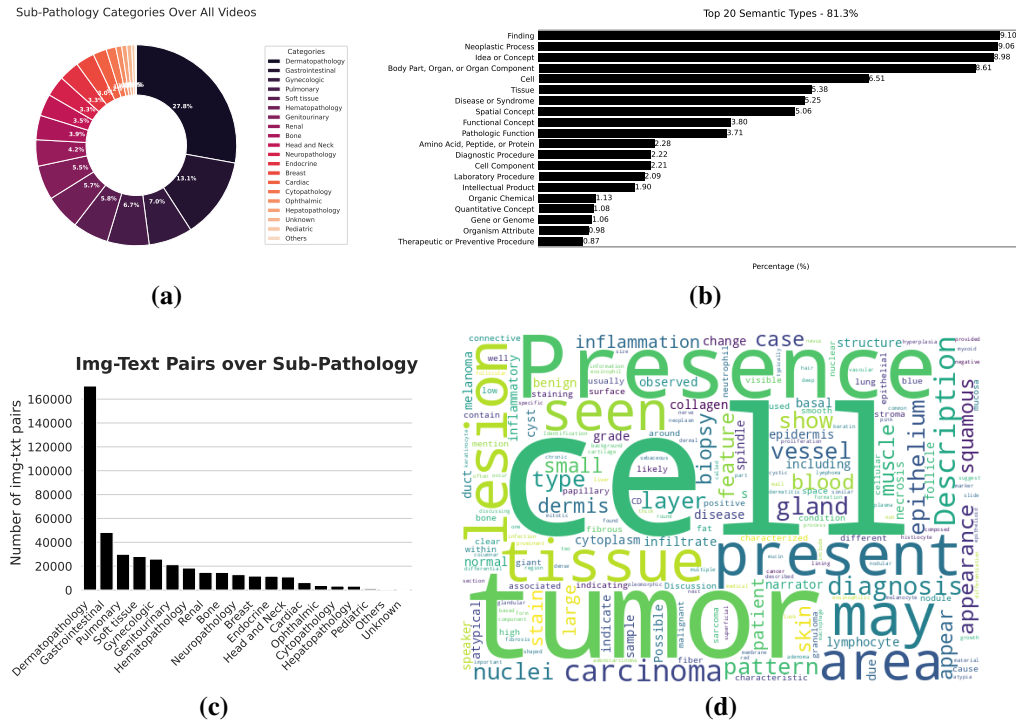


Figure 14: (a) Distribution of all videos over sub-pathology types. (b) Distribution of our entities across top 20 UMLS semantic types. (c) Number of image-text pairs within each sub-pathology type. (d) word cloud of all the text in QUILT.

746 C Pretraining and downstream evaluation details

747 C.1 External Evaluation Datasets

748 **PatchCamelyon** Veeling et al. [61] contains 327,680 color images (96×96px) from histopathology
749 scans of lymph node sections. The images are assigned a binary label indicating whether they contain

Table 10: 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

750 metastatic tissue or not. **NCT-CRC-HE-100K** Kather et al. [33] consists of 100,000 non-overlapping
751 image patches from hematoxylin and eosin (H&E) stained histological images (224x224px) of human
752 colorectal cancer and is categorized into cancer and normal tissue. **SICAPv2** Silva-Rodríguez et al.
753 [57] contains 182 prostate histology WSIs with 10,340 patches (512 x 512px) and both annotations of
754 global Gleason scores and patch-level Gleason grades. Images are labeled as Non cancerous, Grade
755 3, Grade 4, and Grade 5. **Databiox** [8] consists of 922 Invasive Ductal Carcinoma cases of breast
756 cancer. This data set has been collected from pathological biopsy samples of 150 patients which are
757 labeled as Grade I, II and III. Each pathological sample in has four levels of magnification: 4x, 10x,
758 20x and 40x. **BACH** [4] consists of 400 WSIs of breast tissue which are labeled as normal, benign,
759 in-situ and invasive carcinoma. **Osteo** [5] is a set of 1,144 patches (1024 x 1024px) taken from 40
760 WSIs representing the heterogeneity of osteosarcoma. Images are labeled as Viable tumor (VT),
761 Non-tumor (NT) and Necrotic tumor (NEC). **RenalCell** [10] contains 27,849 images of clear-cell
762 renal cell carcinoma H&E-stained (300 x 300px) annotated into five tissue texture types. **SkinCancer**
763 [36] consists of 36,890 patches (395 x 395px) from WSIs skin biopsies from patients with Basal cell
764 carcinoma (BCC), squamous cell carcinoma (SqCC), naevi and melanoma. Images were annotated
765 for 16 categories: chondral tissue, dermis, elastosis, epidermis, hair follicle, skeletal muscle, necrosis,
766 nerves, sebaceous glands, subcutis, eccrine glands, vessels, BCC, SqCC, naevi and melanoma.
767 **MHIST** [64] contains 3,152 patches (224 x 224px) from 328 Formalin Fixed Paraffin-Embedded
768 WSIs of colorectal polyps. These images are labeled as hyperplastic polyps (HPs) or sessile serrated
769 adenomas (SSAs). **LC25000** [9] which we divide into **LC25000 (Lung)** with 15,000 and **LC25000**
770 **(Colon)** with 10,000 color images (768x768px). The lung subset is labeled as lung adenocarcinomas,
771 lung squamous cell carcinomas, and benign lung tissues and the colon subset is labeled as colon
772 adenocarcinomas and benign colonic tissues. Table 15 summarizes these datasets.

Table 15: Downstream tasks and datasets. Note that SkinTumor dataset is a subset of SkinCancer. [μm per pixel - MPP]

	Task	Sub-Pathology	Dataset	Classes	Magnification	Size (Train/Val/Test)	Image-size
Classification	Lymph-node metastasis detection	Breast	PatchCamelyon [61]	2	1 MPP	(0.75/0.125/0.125) * 327,680	96 x 96
	Tissue Phenotyping	Colorectal	NCT-CRC-HE-100K [33]	8	0.5 MPP	89,434/ - /6333	224 x 224
	Gleason scoring	Prostate	SICAPv2 [57]	4	1 MPP	- / - /10,340	512 x 512
	Bloom Richardson grading	Breast	Databiox [8]	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 [10]	5	[0.5, 0.25] MPP	- / - / 27,849	300 x 300
	Classification of skin neoplasms and various anatomical compartments	Skin	SkinCancer [36]	16	.5 MPP	28039/-/8851 ^{imb}	395 x 395
	Colorectal Polyp Classification	Colorectal	MHIST [64]	2	1 MPP	- / - / 3,152	224 x 224
	Lung adenocarcinoma classification (normal, adenocarcinoma and SCC)	Lung	LC25000 (LUNG) [9]	3	- MPP	- / - / 15,000	768 x 768
Colon adenocarcinoma classification (normal and colon adenocarcinoma)	Colon	LC25000 (Colon) [9]	2	- MPP	- / - / 10,000	768 x 768	
Retrieval	histopathology image-text retrieval	-	Quilt-1M	1.02M	-	13,559	-
	histopathology image-text retrieval	-	ARCH [20]	-	-	7500	-

773 C.2 QUILTNET Implementation

774 All model implementations in this study are built upon the open source repository OpenCLIP
775 [28], which enables large-scale training with contrastive image-text supervision. The experiments
776 were conducted using PyTorch and utilized up to 4 NVIDIA A40 GPUs. The hyperparameters for
777 finetuning and training from scratch are provided in Table 12. During the training process, gradient
778 checkpointing and automatic mixed precision (AMP) techniques were employed, with a datatype of
779 bfloat16.

780 All models were trained with image size of 224, except for the finetuned ViT-B-32 models, where
781 the images were first resized to 512 before randomly cropping them to the desired size of 224. In
782 the case of ViT-B-32 finetuning, the data was kept stretched, meaning it maintained a one-to-one
783 mapping between the image and text. However, for all other models, the data was unstretched. This
784 means that for those models, sampling from medical texts occurred with a probability of $p = \text{sample}$
785 prob , or sampling from ROI texts. Within the medical or ROI texts, sampling was done uniformly.
786 For all finetuned GPT/77 models we use the OpenAI CLIP [49] pretrained network as initialization

787 and for ViT-32 maintain the use of QuickGeLU¹³. We perform hyperparameter tuning for all linear
 788 probing results, exploring different values for learning rate, epochs, and weight decay. This process
 789 helped optimize the performance of the models during the linear probing stage.

Table 12: 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 [41]	same
<i>sample prob</i>	0.85	same

Table 14: 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	45.38	8.93	5.40	24.75	23.41	28.93
SkinTumor	10.35	56.36	58.29	36.26	13.85	37.0	51.47	51.20
NCT-CRC	26.4	54.02	59.56	17.35	20.09	51.71	28.68	59.20
PatchCamelyon	61.88	58.61	64.6	49.92	50.45	53.25	67.91	53.52
MHIST	52.92	57.52	62.54	44.52	52.3	40.23	44.32	52.71
LC25000(LUNG)	61.36	78.77	80.16	67.71	50.29	72.44	50.71	81.87
LC25000(COLON)	62.5	77.79	93.28	72.08	78.56	90.57	62.26	87.1
SICAPv2	39.40	44.53	39.49	25.07	27.38	45.81	25.54	45.1
BACH	26.0	43.0	41.25	33.75	27.25	54.75	40.75	62.0
Databiox	37.53	39.48	42.52	32.32	33.51	31.24	33.19	29.93
Osteo	19.49	54.02	64.16	27.88	16.08	50.79	38.37	59.79
RenalCell	20.3	50.7	52.57	16.35	28.80	47.08	28.32	50.72

¹³<https://github.com/openai/CLIP/blob/main/clip/model.py>

Table 16: 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'

790 **D Exploration of trained model representations**

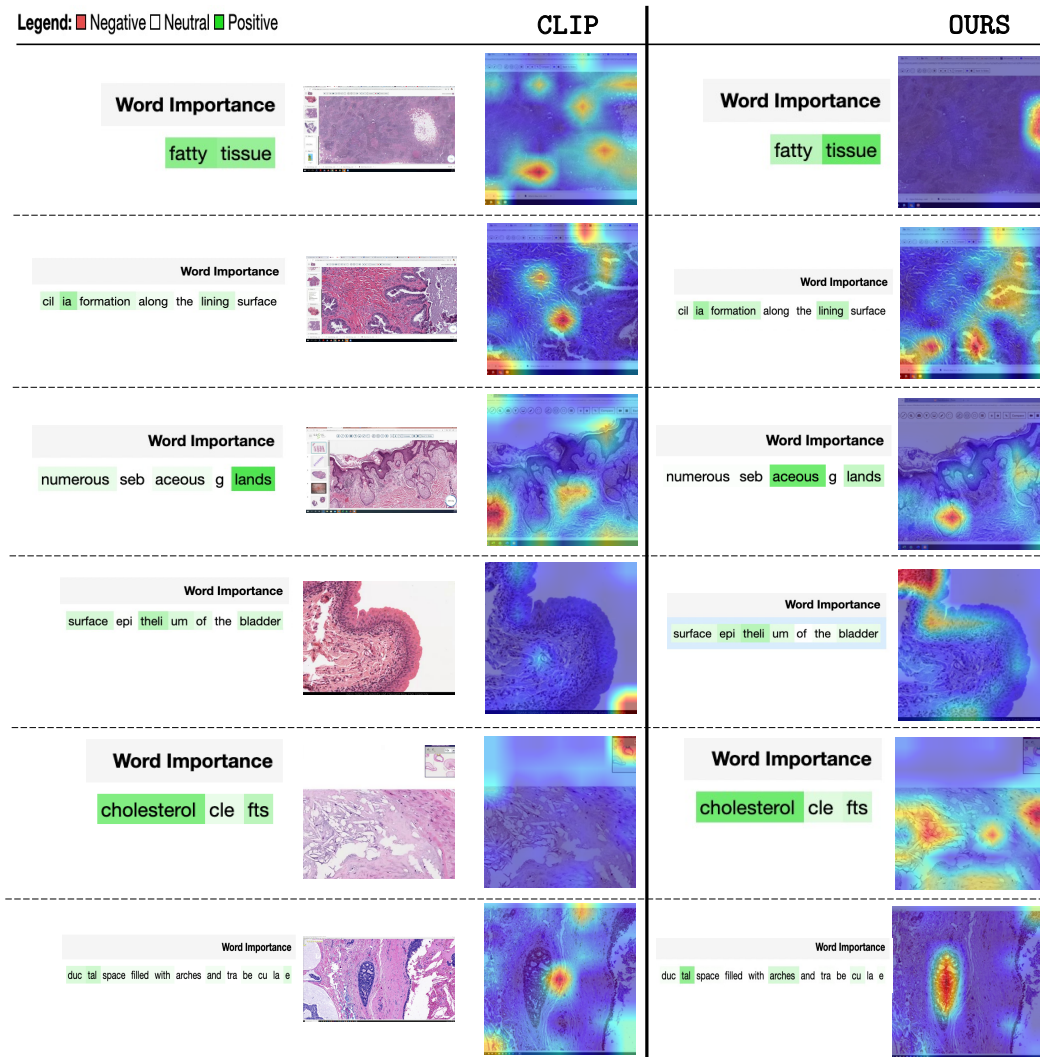
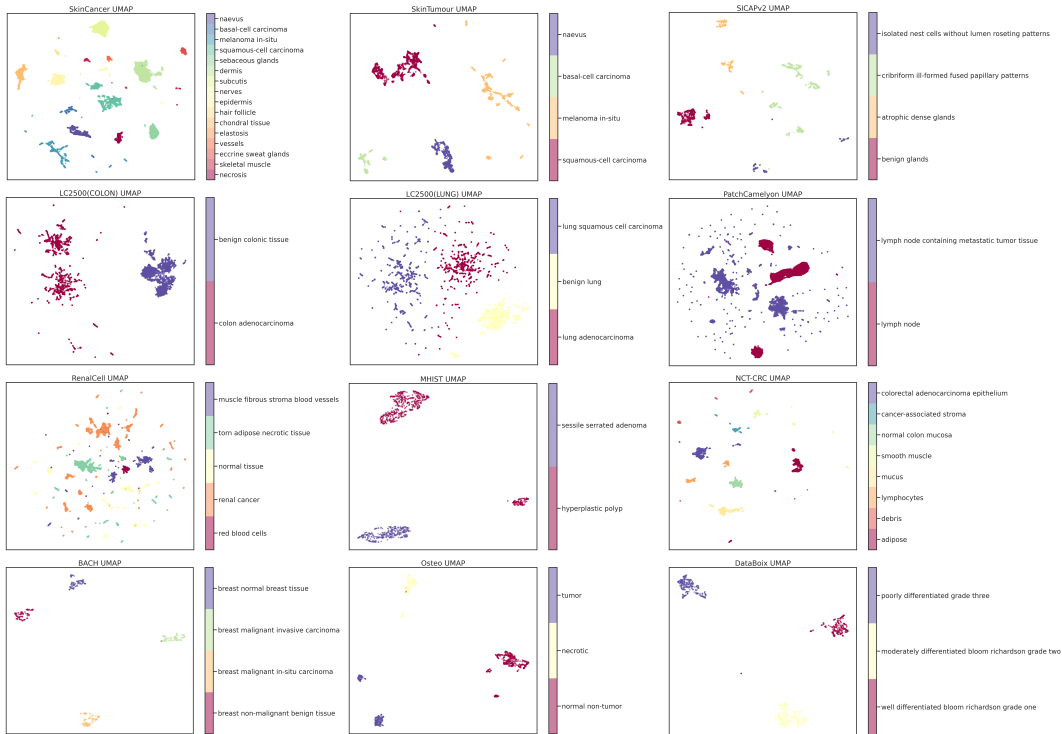


Figure 17: Comparison of the attention maps generated by QUILTNET and CLIP. The corresponding words are highlighted based on their importance. Attention masks were generated using GradCAM [56].

Table 17: UMAP visualization of image embeddings generated by QUILTNET from the different datasets listed in Table 15.



791 E Datasheet for QUILT

792 In this section, we present a DataSheet [21] for QUILT, synthesizing many of the other analyses we
 793 performed in this paper.

794 1. Motivation For Datasheet Creation

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

803 2. Data composition

- 804 • **What are the instances?** The instances that we consider in this work are histopathology
 805 images derived from educational videos, paired with aligned text, derived from ASR
 806 and denoise using an LLM.
- 807 • **How many instances are there?** We include greater than 1 million image-text pairs,
 808 from videos and additionally from less noisy sources like PubMed articles.
- 809 • **What data does each instance consist of?** Each instance consists of an image, a
 810 descriptive text for the image as a whole and for its regions of interest, an estimated
 811 microscope magnification of the image, medical UMLS entities in the text, and the
 812 subpathology type. Each instance is representative of a video chunk based on where
 813 histopathology content is stable.

- 814 • **Is there a label or target associated with each instance?** We use the raw ASR and
815 LLM denoised captions as labels in this work as well as auxiliary information which
816 includes magnification, UMLS entities and pathology type.
- 817 • **Is any information missing from individual instances?** Yes, for instances in the
818 dataset that are not from QUILT (i.e videos), e.g. from PubMed Article datapoints, the
819 additional auxiliary information is not included.
- 820 • **Are relationships between individual instances made explicit?** Not applicable – we
821 do not study relationships between disparate videos (even from the same narrator) nor
822 the relationship between chunks in the same video.
- 823 • **Does the dataset contain all possible instances or is it a sample?** Contains all
824 instances our curation pipeline collected, as the list of videos is not exhaustive of what
825 is available online, there is a high probability more instances can be collected in the
826 future.
- 827 • **Are there recommended data splits (e.g., training, development/validation, test-
828 ing)?** There are no recommended data splits, as this data was curated mainly for
829 pretraining rather than evaluation.
- 830 • **Are there any errors, sources of noise, or redundancies in the dataset? If so, please
831 provide a description.** Yes. Despite our numerous attempts to reduce noise using
832 various models, algorithms and human knowledge databases, ASR is often noisy, and
833 there are many errors that we cannot fix.
- 834 • **Is the dataset self-contained, or does it link to or otherwise rely on external
835 resources (e.g., websites, tweets, other datasets)?** The dataset is self-contained.
836 However, we plan to only release the video URLs and some paired non-pixel data
837 points, rather than the videos themselves, so as to protect user privacy (allowing users
838 to delete videos).

839 3. Collection Process

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

857 4. Data Preprocessing

- 858 • **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or
859 bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal
860 of instances, processing of missing values)?** Yes, we discuss this in Section 3.1 and
861 in Appendix A.1: of note, we use a large language model, UMLS database and a set of
862 algorithms to ‘denoise’ ASR transcripts, an ensemble of histopathology classifiers to
863 inform relevant segments of the video, and extract the representative image(s) for each
864 video segment.

- 865 • **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data**
866 **(e.g., to support unanticipated future uses)? If so, please provide a link or other**
867 **access point to the ‘raw’ data.** The raw data was saved, but at this time we do not
868 plan to release it directly due to copyright and privacy concerns.
- 869 • **Is the software used to preprocess/clean/label the instances available? If so, please**
870 **provide a link or other access point.** We will make our code public to support future
871 research.
- 872 • **Does this dataset collection/processing procedure achieve the motivation for cre-**
873 **ating the dataset stated in the first section of this datasheet? If not, what are the**
874 **limitations?**
875 Yes, the dataset does allow for the study of our goal, as it covers various histopathology
876 sub-domains and provides crucial data points and metadata for pretraining. Some of
877 its limitations we are aware of involve various biases on YouTube, as well as various
878 inaccuracies of the models (e.g ASR model) within the curation pipeline, which we
879 discuss in Appendix A.1 and A.3.

880 5. Dataset Distribution

- 881 • **How will the dataset be distributed?** At this time, we plan to distribute all the derived
882 data (captions, magnifications etc), as well as links to the YouTube videos that we used.
883 We will do this on our website under the MIT license.
- 884 • **When will the dataset be released/first distributed? What license (if any) is it**
885 **distributed under?** We will release it as soon as possible, using a permissible license
886 for research-based use.
- 887 • **Are there any copyrights on the data?** We believe our use is ‘fair use,’ however, due
888 to an abundance of caution, we will not be releasing any of the videos themselves.
- 889 • **Are there any fees or access restrictions?** No.

890 6. Dataset Maintenance

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

899 7. Legal and Ethical Considerations

- 900 • **Were any ethical review processes conducted (e.g., by an institutional review**
901 **board)?** No official processes were done, as our research is not on human subjects,
902 however, because the dataset is in the medical domain we had significant internal
903 discussions and deliberations when choosing the scraping strategy.
- 904 • **Does the dataset contain data that might be considered confidential?** No, we only
905 use public videos.
- 906 • **Does the dataset contain data that, if viewed directly, might be offensive, insulting,**
907 **threatening, or might otherwise cause anxiety? If so, please describe why** No –
908 because many of these videos are medical and educational in nature, we have not seen
909 any instance of offensive or abusive content.
- 910 • **Does the dataset relate to people?** Yes, it relates sometimes to deidentified patients,
911 typically studied by pathologists.
- 912 • **Does the dataset identify any subpopulations (e.g., by age, gender)?** Not explicitly
913 (e.g. through labels)

914
915
916
917
918
919

- **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.