A MEDICALNARRATIVES: Video Curation

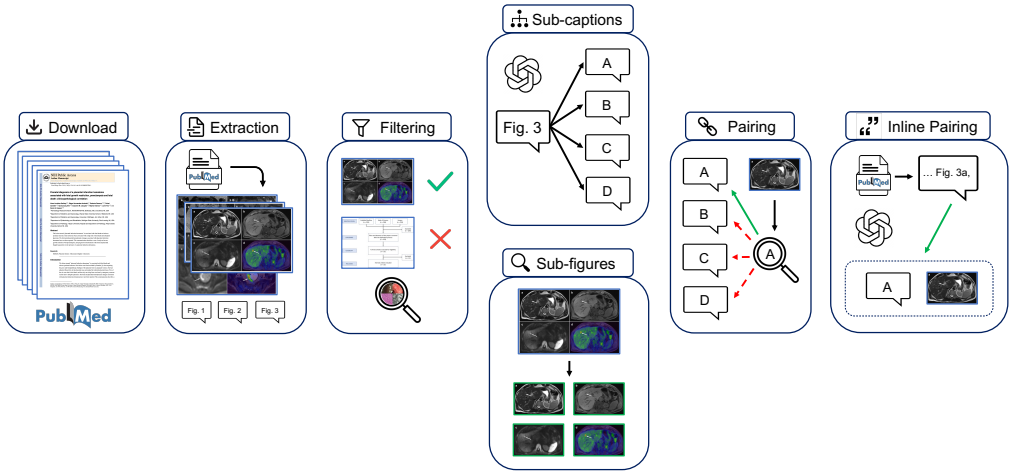


Figure 6: The data curation pipeline for the PubMed subset of the MEDICALNARRATIVES dataset. **Download**: downloading PMC-OA. **Extraction**: extracting figures, captions, and inline references. **Filtering**: filtering for medical images. **Sub-captions**: splitting compound figure captions into sub-captions. **Sub-figures**: detecting and cropping sub-figures from compound figures. **Pairing**: matching sub-figures and sub-captions. **Inline pairing**: matching inline mentions of figures with the most relevant sub-figure or sub-figures.

Distilling the volume of data YouTube offers into a grounded vision-language dataset that captures the all available signal of medical pedagogy video data is a significant task. Each step in the data curation process presents unique challenges when scaling to handle multiple medical domains.

With MEDICALNARRATIVES, we collect vision-language datasets grounded in time with language-correlated traces across twelve medical domains with the first three domains defined to be *static* where representative samples are usually static images: (1) computed tomography (CT), (2) magnetic resonance imaging (MRI), and (3) xray, and *non-static* domain with representative samples exhibiting significant visual change: (4) ultrasound, (5) surgery, (6) endoscopy, (7) dentistry, (8) dermatology, (9) mammography, (10) ophthalmology, (11) histopathology and (12) general medical illustrations. When processing these subsets, our approach differs to accommodate the nuances of the video data. Our data curation pipeline can be split into these high-level tasks of:

(A) Searching for representative videos in each medical domain.

(B) Filtering videos for narrative style.

(C) Extracting image, text, and cursor traces from selected videos.

(D) Denoising and de-duplicating the collected raw data.

(E) Aligning image, text, and localization traces.

(F) Collecting metadata useful for varying downstream tasks (e.g. subdomains) and interleaving the dataset.

In the following sections, we present a detailed overview of the major steps in curating MEDICAL-NARRATIVES starting with search. We also present examples of curated narrative samples in Figures 12, 13, 10, and 11.

A.1 Domain-Specific Search

We first identify medical channels and videos for each domain on YouTube, using keywords from online medical glossaries specific to each imaging modality or medical domain. To increase the percentage of narrative or educational style videos, a list of priority keywords: "educational", "interpretation", "case study", and similar phrases, are appended to search keywords. We limit

our search to channels with <1M subscribers since some channels focus on multiple domains (e.g. radiology channels span CT, MRI, X-ray) and therefore might have a large subscriber base, and channels with >1M subscribers often contain non-imaging videos.

We observe during channel search that searching YouTube for channels by keyword tends to produce irrelevant results, hence, we adopt a video-first search strategy: since video titles are more informative than channel titles, we first find relevant videos, then evaluate the channel of the relevant video for more hits. Each video result is downloaded in low resolution for further analysis. To limit searching irrelevant channels we implement early stopping, wherein, if the first 10 videos of a channel fail the medical filtering step, the channel is skipped, allowing us to keep compute cost low while increasing our pool of visited channels.

A.2 Medical Filtering

Each potential pedagogy video is evaluated by the following heuristics:

1. The video duration is longer than 1 minute and shorter than 2 hours as videos outside this range usually contain little medical imaging information.
2. Video contains speech. We check this either through the video's transcript from the YouTube API, or if not present using the inaSpeechSegmenter [2] tool on the first minute of audio.
3. The number of medical scene frames exceeds the empirically determined threshold unique to each medical domain. This heuristic filters for narrative-style videos (See Section A.3).

To expand on the third heuristic, we extract the key-frames of a video for classification; for static domains, we utilize FFmpeg [3] to detect scenes and extract key-frames (frames with significant visual changes from previous frames). We experiment with scene detection thresholds to determine the optimal threshold per domain across various video durations. For non-static domains, we leverage adaptive content scene detection in [4] to avoid capturing frames that are visually different but still part of the same shot (which are characteristic of non-static domains). Camera movements are common in domains such as surgery or endoscopy, and nearly duplicate frames that would be generated by thresholding on video content are instead merged when using PySceneDetect's adaptive detection algorithm. We specifically tune the adaptive detection for each domain by experimentally determining parameters for the algorithm on sample videos from each domain.

We then classify the key-frames of a video using pre-trained classifiers per domain (see Section D.4). Using the percentage of key-frames predicted to be medical images, videos are differentiated into three categories: positive videos, near-positive videos, and negative videos. For example, a video with 50% of key-frames predicted to be MRI images is a candidate for further processing, while a video with only 2% of key-frames is not. Positive videos contain sufficient medical content for the given domain, while near-positive videos may or may not contain sufficient medical content. The thresholds defining positive/near-positive/negative are unique to each domain. We then manually examine a subset of the near-positive category, and determine a more fine-grained percentage threshold to extract more positive videos out of the pool of near-positive videos. See Table 4 for the final percentage thresholds used.

Domain	**Threshold (%)**
CT & X-ray	10
MRI	5
Dermatology & Dentistry	30
Endoscopy & Surgery	50
Ultrasound	40
Ophthalmology	35
Mammography	25
General medical illus.	20

Table 4: Final percentage thresholds used during video key-frame classification.

[2] https://github.com/ina-foss/inaSpeechSegmenter

[3] https://github.com/FFmpeg/FFmpeg

[4] https://github.com/Breakthrough/PySceneDetect

A.3 Narrative Filtering

We define narrative-style videos as pedagogy videos where the narrator focuses on describing or analyzing medical images onscreen. To filter for these videos, we first check the first minute of each medical video for speech using inaSpeechSegmenter to ascertain the presence of a narrator.

For static domains like X-ray, we define a narrative streak as any partition of the video where frames sampled close (w.r.t. time) together are similar using cosine similarity, indicating the narrator is spending a lot of time analyzing that frame. Specifically, we randomly sample a fixed number of clips across each video, sampling three consecutive frames from each clip and checking for similarity. If all three have similarity scores $\geq$ a preset threshold of 0.9, we count it as a narrative streak. A video is tagged as narrative if a domain-specific preset percentage (%) of the selected frames exhibit a narrative streak. This simple filtering algorithm helps us reduce the number of videos we process from 748k to 74k videos.

For non-static domains like surgery or ultrasound, consecutive key-frames often exhibit significant change so we instead look for persistent narration around key-frames classified as medical. For example, for ultrasound clips, we extract the times for each consecutive positive key-frame accumulating a sequence of start and end times. Within these time intervals, we determine whether speech exists either through the video's YouTube API transcript or by extracting the audio during the selected time interval and using inaSpeechSegmenter to determine if the segment contains any speech. A video is considered narrative if more than half the key-frames have text for more than a domain-specific number of seconds.

A.4 Text Extraction using ASR and Text Denoising.

In line with Quilt-1M [53] we leverage an open-source ASR model - Whisper [99] to transcribe all speech from the selected videos and make sure to account for transcription errors using a similar methodology of finding these types of errors and correcting with a language model. We use the whisper-large-v2 model in the stable-ts library for word-level and sentence-level transcription. As anticipated, this model often misinterprets medical terms, thus requiring the use of post-processing algorithms to minimize its error rates. For this, we adopt a similar methodology proposed in Quilt-1M [53] to identify, correct, and verify these errors, please see section A.1 in Quilt-1M [53] supplementary for more details.

From the ASR-corrected text, we extract *medical text* which describes the image(s) as a whole. Also, when the speaker describes/gestures at visual regions-of-interest through statements like *"look here ..."*, we extract the text entity being described as *ROI text* in line with Quilt-1M [53].

To extract relevant text, we prompt LLMs to filter out all non-medically relevant text, providing context as necessary, while conditioning the LLMs to refrain from introducing new words beyond the corrected noisy text and set the model's temperature to zero. Lastly, the LLM is used to categorize our videos into subdomains by conditioning with a few examples and prompting with the corrected video transcript as input (see Figure 23 for prompt and sample input/output).

A.5 Aligning modalities

Videos: We modify Quilt-LLaVA [113] pipeline. To align image, text, and trace modalities we compute time chunks for each video denoted as $[(t_1, t_2), (t_3, t_4), \cdots (t_{n-1}, t_n)]$ from key-frames after discriminating for medical frames using the pretrained classifiers – (*scene_chunks*). Each *scene_chunk* is padded with *pad_time* to its left and right. We use the methods described above to extract the medical/ROI captions as well as the representative image(s) for every chunk/time-interval in *scene_chunks* Finally, each chunk in *scene_chunks* is mapped to text (both medical and ROI captions), traces, and images. Next, we map each image to one or more text (with traces). Using the images' time interval, we extract *raw_keywords* using the Rake method from the transcript. We extract *keywords* from each medical text returned using the LLM. Finally, if the *raw_keywords* occur before or slightly after a selected representative image, and overlap with the *keywords* in one of the Medical/ROI texts for that chunk, we map the image to the medical/ROI text. Traces are encoded as the cartesian position of the cursor relative to the image size, we use (x_j^t, y_j^t), where $x \in [0, W]$ and $y \in [0, H]$, with W and H representing the image width and height, respectively, t spans from 0 up to the total duration of the jth stable chunk.

Articles: The majority of our curated PubMed data uses alphabetic labels in compound figures to denote sub-figures, which increases the complexity of pairing individual sub-figures from compound figures to sub-captions. Our solution leverages an optical character recognition (OCR) model [5] on each sub-figure to detect the sub-figure labels, which we then match to the extracted sub-caption labels. We impose a 95% confidence threshold on predicted text to isolate the sub-figure label, as text detected at lower confidence is often non-label text present in the figure (e.g. axis titles, graphs). We then match and pair the detected sub-figure label with the sub-caption label. Despite the generality of this approach, we identified a few failure cases and proposed an error-handling solution in section B.5 in the Appendix.

Hyperparameter	Training
Batch size (per GPU)	256
Epochs	20
Peak learning rate	1e-5
Learning rate schedule	cosine decay
Warmup (in steps)	2000
Augmentation	Resize; RandomCrop (0.8, 1.0)
Optimizer momentum	β_1, β_2 = 0.9, 0.98
Weight decay	0.2
Optimizer	AdamW

Table 5: Training hyperparameters for GENMEDCLIP

Hyperparameter	ResNet50	ViT-Small
Batch size (per GPU)	256	32
Epochs	10	100
Peak learning rate	1e-2	1e-3
Learning rate schedule	-	cosine annealing
Augmentation	RandomResizedCrop (224), Resize	RandomResizedCrop (384, 0.98, 1.0), RandomHorizontalFlip
Optimizer momentum	0.9	0.9
Weight decay	1e-4	-
Optimizer	SGD	SGD

Table 6: Training hyperparameters for domain classifiers.

Datasets	**prompt**
Pcam [121], Nck [62], Lc2500 [18], Mhist [125] Bach [10], Skin [69], Osteo [12]	["a histopathology slide showing {c}.", "histopathology image of {c}.", "pathology tissue showing {c}.", "presence of {c} tissue on image"]
Tcga_til [109], DDI [30], Isic [27], Dental [100] Gastrovision [55], G1020 [13], Octdl [70], VinDrM [88] VinDrXR [87], Dresden [21] , Radimagenet [80]	["{c} presented in image", "evidence of {c} in image", "an image showing {c}"]

Table 7: Zero-shot classification templates used to evaluate GENMEDCLIP's zero-shot capacity across all multiple dataset that constitute the medical benchmark.

B MEDICALNARRATIVES: Article Curation

To curate the PubMed subset of MEDICALNARRATIVES, we download the PubMed Central Open Access Subset (PMC-OA) [86], containing 5.47 million articles and filter article figures for the same 12 domains as the YouTube subset of MEDICALNARRATIVES. Our data curation pipeline for PubMed is as follows:

(A) Downloading PMC-OA and extracting each article's XML and images.

(B) Parsing each XML to extract figure captions and inline mentions of figures.

(C) Filtering for figures with medical imaging with pretrained classifiers.

[5] https://github.com/JaidedAI/EasyOCR

Models	Models	Xray (mAP)	CT		MRI							Mammo (mAP)	US	Optha		Derm		Endo	Surg	Dental	Histo								Overall
	vindrXR	lung	abd	af	brain	hip	knee	abd	shdr	spine	vindrM	rad	g1020	octdl	ddi	isic	gastro	dresden	dental	til	pcam	lc_lung	nck	skin	skin_tumor	lc_colon	mhist	bach	osteo	
CLIP-ViT-B-32 [98]	-	6.95	2.74	4.14	2.75	0.77	2.67	1.86	2.23	3.13	12.29	10.64	8.45	69.61	10.22	41.31	21.76	4.94	10.66	-	21.33	61.81	61.55	29.20	4.47	9.84	65.57	50.67	25.25	58.85	21.63
CLIP-ViT-B-16 [98]	-	6.55	1.61	3.76	3.59	6.55	3.00	1.13	1.49	4.29	14.07	10.65	5.41	31.47	6.98	60.67	13.23	1.77	11.38	2.51	20.32	51.80	47.06	21.41	5.55	13.22	79.56	52.61	23.75	43.54	18.93
PMC-CLIP [73]	A	6.80	10.89	5.73	1.19	8.31	3.24	7.90	0.11	2.00	2.55	11.86	3.72	37.06	3.78	52.44	3.77	0.86	8.15	27.85	68.92	47.06	32.66	14.37	3.86	29.66	49.96	47.39	19.50	46.20	19.23
PUBMEDCLIP [35]	A	7.40	6.06	2.04	1.08	8.95	1.38	4.07	0.54	8.50	31.21	10.45	15.94	68.33	7.56	35.52	18.19	2.05	9.72	12.19	24.55	50.38	33.33	26.45	8.07	23.01	63.66	62.74	15.25	43.63	20.77
BIOMEDCLIP [134]	A	10.29	6.80	2.53	3.11	12.31	2.98	4.88	0.97	6.09	11.35	10.57	58.10	29.12	18.60	51.22	20.70	3.49	16.37	16.83	37.03	71.71	71.34	49.17	24.83	40.39	84.98	38.59	44.25	46.75	27.43
GENMEDCLIP-32	V+A	10.23	14.25	1.59	2.25	23.40	3.68	9.45	1.18	2.36	24.00	10.30	44.98	66.67	20.35	62.96	19.78	2.10	15.04	26.89	23.14	70.56	81.11	48.05	28.45	49.67	93.24	55.37	37.25	55.82	31.18
GENMEDCLIP-PMB	V+A	9.90	10.36	2.30	3.95	8.80	1.80	4.33	0.98	8.38	25.04	12.06	52.98	29.22	24.56	57.01	37.63	2.08	19.30	16.83	49.63	71.90	82.05	51.05	39.32	48.34	71.68	61.82	52.00	42.44	30.96
GENMEDCLIP	V+A	9.66	27.35	1.38	2.75	7.52	2.61	9.93	2.80	3.10	22.59	11.12	63.48	33.53	21.22	72.26	37.10	2.38	18.97	16.44	20.34	65.90	72.37	52.16	42.37	59.87	94.16	60.59	52.50	41.43	**31.99**
Data Split Ablation																															
GENMEDCLIP *	A	6.52	0.84	1.19	4.03	2.25	3.12	7.51	2.44	3.42	24.25	11.91	1.83	29.80	17.83	41.01	6.94	5.16	9.25	11.22	79.29	71.57	52.23	30.24	3.62	20.27	49.52	37.67	26.50	42.90	20.84
GENMEDCLIP *	V+A	8.51	27.17	1.38	1.26	19.87	1.90	10.76	2.16	3.36	8.34	10.87	55.01	34.12	30.47	73.02	42.00	1.61	13.23	-	23.62	66.94	82.01	43.74	41.46	58.10	93.66	63.15	52.25	39.78	32.49

Table 8: **Expanded Zeroshot Classification Results** shows that our model GENMEDCLIP outperforms all other baselines including the out-of-domain CLIP and biomedical vision-language models BIOMEDCLIP and PUBMEDCLIP across the constructed medical benchmark. The benchmark covers all 11 medical domains represented, excluding the non-medical domain of medical illustrations. The metric for X-ray and Mammography is mean average precision while the rest is accuracy.

Domain	CT	MRI	X-ray
Image-text-ROI pairs	79562	82760	78983
Image-text-ROI-text pairs	127533	112940	135242
Avg. ROI Text/Image ROI	3.29	2.9	3.82
Num. ROI Text/Video	98547.0	86798.0	85684.0
Avg. Words/ROI Text	10.66	9.61	12.33
Avg. ROI UMLS/Text	1.47	1.48	1.47
Avg. ROI/Image	1.6	1.38	1.74
Avg. ROI Text/Chunk	2.61	2.33	2.54
Unique ROI BBox	45680	35102	49157
Unique ROI Traces	11429184	4797419	10661187
Avg. ROI Chunk Duration	12.85	6.19	14.42
Avg. BBox Height	319.31	204.05	312.08
Avg. Bbox Width	538.48	281.52	506.97

Table 9: Characterization of MEDICALNARRATIVES ***image-text-trace*** subset, categorized by individual medical domains. The table provides detailed statistics for each medical modality, including the number of unique images, total dataset duration, ASR error rate, and average image resolution. Note: "ROI" in the table is shorthand for traces.

(D) Splitting compounded figures and captions using fine-tuned object detection models and a language model.

(E) Pairing correctly split sub-captions and sub-figures together using a combination of optical character recognition (OCR), bounding box heuristics, and error correction.

(F) Matching inline mentions of figures with sub-figure/sub-caption pairs using a language model.

In the following sections, we will discuss the MEDICALNARRATIVES PubMed data pipeline.

B.1 Caption Extraction

From each obtained PubMed article, we extract the XML and image files for figure processing. The figure captions are extracted from the paper XML, cleaned, and paired with the corresponding image file of the figure. Additionally, we find all inline mentions of the figure and save them to the figure-caption sample. This yields 23.6M figure-caption samples.

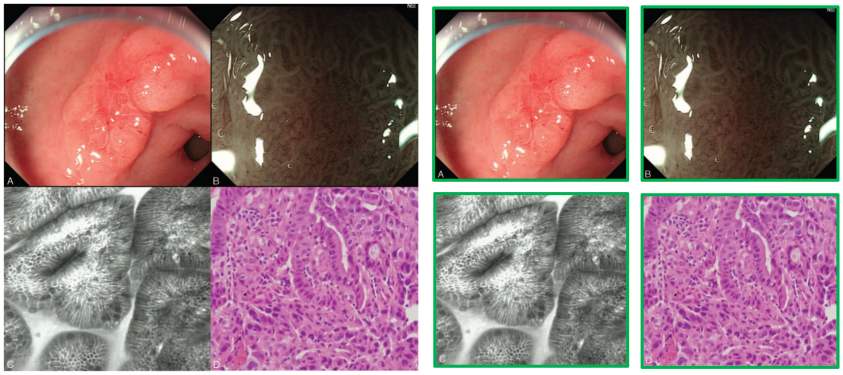
B.2 Medical Filtering

To determine whether a figure belongs to one of the twelve domains of MEDICALNARRATIVES, we train a ResNet-50 CNN for binary classification. We use the same training datasets (see Table 11) selected when curating MEDICALNARRATIVES YouTube data, with a binary medical/non-medical label as the target prediction. This filtering step reduces the number of potential figures to 1.03M figures. To determine the specific domain or domains of each figure, we re-use the medical domain classifiers from the medical filtering step.

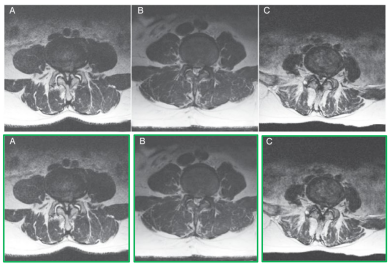
B.3 Sub-figure Detection

The majority of figures after medical filtering are compound figures, which compress detailed information into a single image and caption. Splitting these compound figures into sub-figures is a non-trivial task, since there is no uniform compound figure layout. In contrast to Quilt-1M's [53] image processing-based approach to splitting these figures, we opt for an object detection approach, which we empirically determined is capable of handling wider range of abstract layouts.

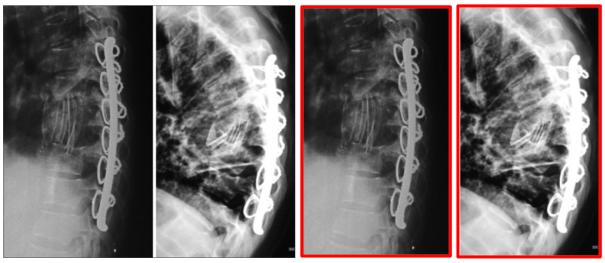
Specifically, we finetune a YOLO object detection model [58] to detect sub-figures within compound figures using two medical subfigure separation datasets: MedICaT's sub-figure annotations and ImageCLEF 2016's Figure Separation medical task [117, 37]. MedICaT contains 7507 sub-figure bounding box annotations from 2069 compound figures. ImageCLEF 2016 Figure Separation contains 6782 sub-figure bounding box annotations. We fine-tune a YOLOv8-Large [58] for 100 epochs using an 80/20 training/validation split on the subfigure separation data. One major advantage of the object detection approach is that our model can successfully detect sub-figures even when there is little to no gap/whitespace in-between sub-figures. We process each figure with the fine-tuned YOLOv8 sub-figure detection model, where each detected sub-figure is cropped, and up-scaled by a factor of 4. In the case of compounded figures with uncompounded caption i.e. all constituting images communicate a singular concept (see Figure 7) we pair the caption to the original compounded figure.



(a)



(b)



(c)

Figure 7: (a), (b) Compound figures successfully separated into sub-figures (green), which are then up-scaled and saved. (c) A figure that is incorrectly identified as a compound figure. Since the figure caption contains no sub-captions, the original figure will paired with the entire caption during sub-figure/sub-caption pairing.

B.4 Sub-caption Separation

A compound figure caption usually contains multiple sub-captions. A heuristics-based approach to splitting these compounded captions is difficult to design since figure sub-captions are labeled differently with article authors adhering to varying writing styles typically set by the publishing journal. We therefore opt for an LLM-based approach where we provide diverse examples of sub-caption separation, instructing the language model (GPT-3.5 Turbo) to follow the process below:

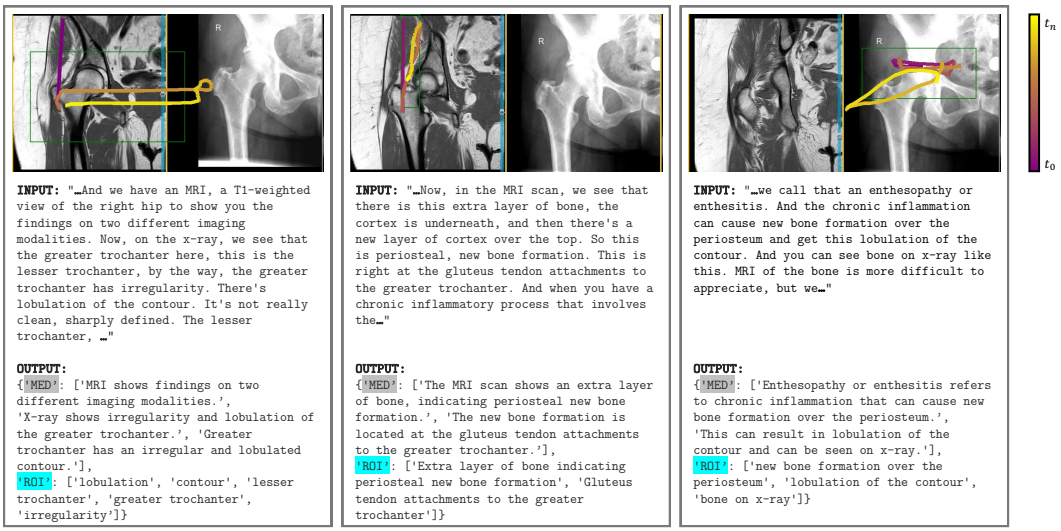


Figure 8: **MEDICALNARRATIVES**: Here we show 3 samples from the dataset, these samples come from a single video containing two medical modalities, MRI and X-ray scans, and can be concatenated into an interleaved sample with each sample showing the representative image captured, the raw input text grounded and aligned in-time with the spatial traces & bbox, and finally the denoised medical and ROI text describing the medical image removing all transcription errors and non-medical information.

1. Separate the figure caption into sub-captions based on the sub-figure labels present in the caption e.g. "(A)", "I)", "a.", "bottom left", etc.
2. Strip the sub-figure labels from each sub-caption text produced.
3. If any context in the caption pertains to the entire figure, add this context to each sub-caption. This step ensures that each individual sub-caption retains the entire context of the figure.
4. Return each sub-caption paired with its sub-figure label.

We condition the LLM with a few examples, including handling non-compound figures and captions that use spatial cues (e.g. left, center, right) to refer to sub-figures (see full prompt and sample input/output in Figure 24). We also process the sub-figure labels returned from the LLM, stripping parentheses and other extraneous characters to make sub-figure/sub-caption pairing easier.

B.5 Pairing Sub-figures to Sub-captions

Given the separated sub-captions and sub-figures for a compound figure, next we tackle the problem of pairing the correct sub-caption with the correct subfigure. The majority of our curated PubMed data uses alphabetic labels in compound figures to denote sub-figures. Our approach therefore leverages optical character recognition on each sub-figure to detect the sub-figure labels, which we then match to the sub-caption labels extracted during section B.4.

During the sub-figure detection step, we upscale the detected sub-figures by a factor of 4 to enlarge the sub-figure text label for OCR. We impose a 95% confidence threshold on predicted text during OCR to isolate the sub-figure label. Text detected at lower confidence is often other text in the figure (e.g. axis titles, graphs) being present. We then attempt to match the detected sub-figure label with the sub-caption label. If a match is found, we pair the selected sub-figure and sub-caption.

There are several types of cases where this approach requires error handling, e.g.:

1. In a single sub-figure, no labels are identified that exceed the 95% confidence threshold.
2. Sub-captions use spatial cues to identify sub-figures, e.g. "upper left", "center", "right".
3. If the number of detected sub-figures does not match the number of sub-captions: either some sub-figures or some sub-captions are unpaired.

In case 1, if the compound figure has exactly one sub-figure and one sub-caption left unpaired, we pair the two. Otherwise, we lower the confidence threshold to 80% and re-detect sub-figure labels, then re-match with sub-captions. Sub-figures that fall in this category tend to have their label close to the border of the cropped sub-figure, have small sub-figure text, or have backgrounds that resemble the font color of the label. For case 2, we use the bounding box coordinates of the detected sub-figures and the spatial cues provided in the caption to pair figures and captions. For example, a sub-caption with the label "upper left" will be paired with the sub-figure with the upper leftmost bounding box. Lastly, case 3 occurs when either sub-figure detection and/or sub-caption separation perform incorrectly. The majority of figures in this category occur when sub-figure detection identifies multiple sub-figures, but the figure caption contains no sub-captions. In this case, we pair the original figure and caption.

Domain	CT	MRI	Endo	Genmed	Surgery	Optha	Mammo	Derma	Ultrasound	X-ray	Dental
Unique images	47441	55784	43230	23985	75312	758	288	19639	69835	54732	15375
Image-text pairs	89036	97065	135108	54684	186807	681	42	27182	140251	101215	27391
Total Med UMLS	295064	262277	504753	224338	656811	3243	251	106291	542695	270191	57086
Avg. Med Text/Image	2.44	2.14	3.49	2.79	2.85	2.48	1.40	2.02	2.61	2.56	2.54
Num. Med Text/Video	80356	73048	134385	54635	185223	681	59	26904	139792	71778	14867
Avg. Words/Med Text	28.35	24.01	40.30	36.01	32.10	36.74	15.62	23.49	31.29	30.04	30.61
Avg. Med UMLS/Text	3.74	3.68	3.83	4.27	3.57	6.07	4.25	3.89	4.04	3.83	3.87
Total Chunks	41870	40133	42770	23947	74188	758	283	19432	69486	37048	7661
Avg. Chunk Duration	30.88	18.26	47.03	36.35	28.85	74.60	2.42	23.71	31.14	34.05	50.42
Avg. Med Text/Chunk	2.12	1.97	3.23	2.35	2.48	1.27	0.16	1.00	1.80	2.07	1.69
Avg. Images/Chunk	1.27	1.51	1.01	1.00	1.03	1.00	1.02	1.01	1.01	1.75	2.03
Avg. Image-Text/Chunk	2.55	2.66	3.25	2.35	2.52	1.27	0.12	1.01	1.81	3.36	3.18
Precision (Unconditioned)	0.16	0.15	0.18	0.17	0.18	0.20	0.33	0.23	0.16	0.19	0.20
Precision (Conditioned)	0.49	0.45	0.48	0.56	0.54	0.44	0.73	0.43	0.40	0.46	0.42
Clinical ASR Error Rate	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Total Duration (hrs)	327.0	416.0	428.0	281.0	389.0	15.0	0.0	187.0	1182.0	562.0	140.0
Avg. Duration (mins)	13.27	12.07	6.24	9.75	7.63	6.94	0.11	8.05	8.58	19.24	13.68
Total ASR len. (words)	2355609	3364120	2722335	2085480	2975575	89011	4399	1204875	9481084	4476475	1149722
Avg. ASR len. (words)	1592.70	1624.39	660.76	1204.09	972.73	659.34	879.80	863.09	1146.72	2550.70	1866.43

Table 10: Characterization of MEDICALNARRATIVES ***image-text*** subset, categorized by individual medical domains. The figure provides detailed statistics for each medical modality, including the number of unique images, total dataset duration, ASR error rate, and average image resolution.

B.6 Inline Figure Reference Pairing

In the final step of the pipeline, we pair the inline reference of a figure with the figure caption since inline references contain valuable context about the figure. However, an inline reference may refer to a sub-figure instead of the entire figure. We therefore utilize a language model to determine which sub-figure is most relevant to an inline reference. For each sample, we prompt an LLM (GPT-3.5 Turbo) with the list of sub-figure labels and a list of inline references and task the model with determining which sub-figure label best corresponds to the inline reference. In the case that the reference cites the entire figure instead of a sub-figure, we consider the inline reference relevant to all sub-figures. For each relevant sub-figure, we add the inline reference to its list of captions. See Figure 25 for the complete prompt and sample input/output.

C Characterizing MEDICALNARRATIVES

To create MEDICALNARRATIVES we combine medical narratives curated from videos with image-text pairs curated from PubMed, resulting in 4.7M total image-text samples of which 1M samples are localized narratives. Section 3.1 gives an overview characterization of the entire dataset, and Tables 10, 9 below provide additional specific characterization details split per domain. Note we omit characterization for Histopathology in the tables below as the details for the domain can be found in prior work.

D Training, Benchmark, and Evaluation

D.1 GENMEDCLIP Training

We leverage OpenCLIP [54] to train our models as it allows us to quickly import our datasets and adapt varying components of our model including the underlying image and text towers and the

training hyperparameters. Our experiments utilize Pytorch on 4 NVIDIA L40s GPUs, as well as gradient checkpointing, automatic mixed precision with bfloat16 to reduce memory usage. All other hyperparameters used are listed in Table 5. Our dataset is split into 16 tar files in the WebDataset [6] format for training.

D.2 Benchmarking on Downstream Medical Tasks

We evaluate the utility of GENMEDCLIP on a new medical imaging benchmark of all medical domains represented in our pre-training dataset MEDICALNARRATIVES, with some domains represented by $>= 1$ dataset/task for classification, totaling 29 downstream datasets and on a held-out set of 1000 unique images for the retrieval task downstream. For MRI we use the **RadImageNet** [80] MRI subsets tasks based on the anatomical region scanned in the image these include Ankle/foot with 25 classes, Brain with 10 classes, Knee with 18 classes, Abdomen/pelvis with 26 classes, Hip with 14 classes, Shoulder with 14 classes, Spine with 9 classes. To evaluate on CT domain we also use RadImageNet's [80] CT dataset which cover two (2) anatomical regions with Lung having 6 sub-classes and Abdomen/pelvis with 28 subclasses. For ultrasound, we evaluate on RadImageNet's [80] US dataset which covers a total of 15 classes across Thyroid and Abdomen/pelvis anatomical regions. For Xray, we evaluate on **VinDr-CXR** Chest Xrays [87] test set and report the mean average precision (mAP) across all 28 findings, similarly to evaluate on Mammography we use **VinDr-Mammo** [88] and report the mAP on all X findings, leveraging only the standard bilateral craniocaudal (CC) view of the test set. We evaluate on surgical organ classification using **Dresden** [21] which covers 8 abdominal organs; to evaluate for endoscopy domain we test on all procedures images in **GastroVison** [55] with 27 classes. For Dermatology we evaluate on the **Diverse Dermatology Images** (DDI) [30] binary (benign or malignant) dataset and Isic 2018 dataset [27]. For Dentistry we evaluate on **Dental orthopantomography** (OPG) [100] X-ray dataset with 6 classes. To evaluate the Ophthalmology domain we evaluate on **G1020** [13] a retinal fundus glaucoma dataset and on **Optical Coherence Tomography Dataset** (OCTDL) [70] with 6 disease classes. We evaluate the Histopathology domain on the following datasets: **PatchCamelyon** [121] for lymph node metastatic tissue binary prediction task, **NCT-CRC-HE-100K** [62] on 8 morphological classes, **BACH** [10] which consists of breast tissues with 4 classes including being and invasive carcinoma, **Osteo** [12] osteosarcoma dataset with 3 classes including necrotic tumor, **SkinCancer** [69] dataset of tissue patches from skin biopsies of 12 anatomical classes and 4 neoplasm categories that make up the SkinTumor Subset, we also evaluate on **MHIST** [125] dataset of colorectal polyps tissue, **LC25000** [18] dataset, which is split in-between LC25000 (Lung) and LC25000 (Colon), for lung and colon adenocarcinomas classification, and on TCGA-TIL [109] for tumor-infiltrating lymphocytes (TILs) binary classification, based on H&E images from 13 of The Cancer Genome Atlas (TCGA) tumor types.

D.3 Evaluation

To evaluate zero-shot classification capacity across all constituting datasets in our medical benchmark outlined in 4 we leverage simple prompts listed in Table 7, with the specific results shown in Table 8.

D.4 Search Classifiers

To classify images into domains, we train a ResNet50 for 10 epochs and a ViT-Small for 100 epochs using DINO on a binary classification task for each medical domain. Both types of models are trained on 4 NVIDIA A4000 GPUs. All hyperparameters are listed in Table 6. For each classifier, we use domain-specific datasets as positive samples and non-medical datasets as negative samples. For the binary medical/non-medical classifier used in Section B.2, we use all medical domain datasets as positive samples, and the same group of non-medical datasets as negative samples. See Table 11 for an overview of the datasets used to train these classifiers.

E MEDICALNARRATIVES Examples

Below, we show examples in the dataset across all 12 modalities and representative examples of the types of interleaved samples within the dataset.

[6] https://github.com/webdataset/webdataset

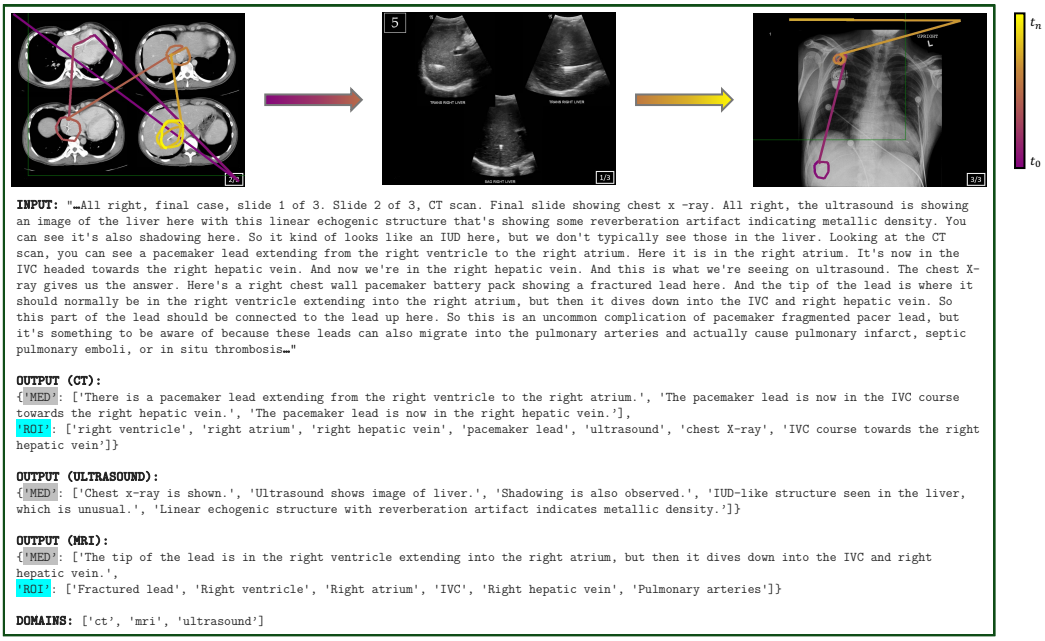


Figure 9: **Interleaved examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical text corresponding to different modalities. **Domains**: classification of the sample into domains.

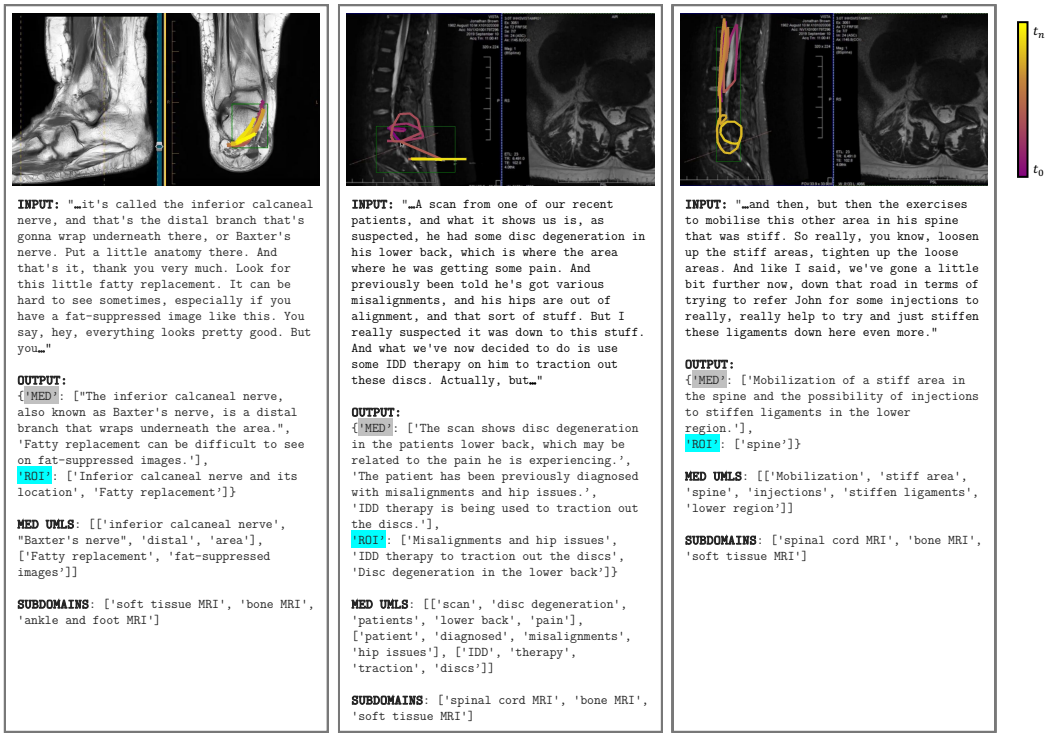


Figure 10: **MRI examples** with in the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical and ROI text. **Traces**: Cursor traces and bounding boxes aligned in-time with the raw text. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

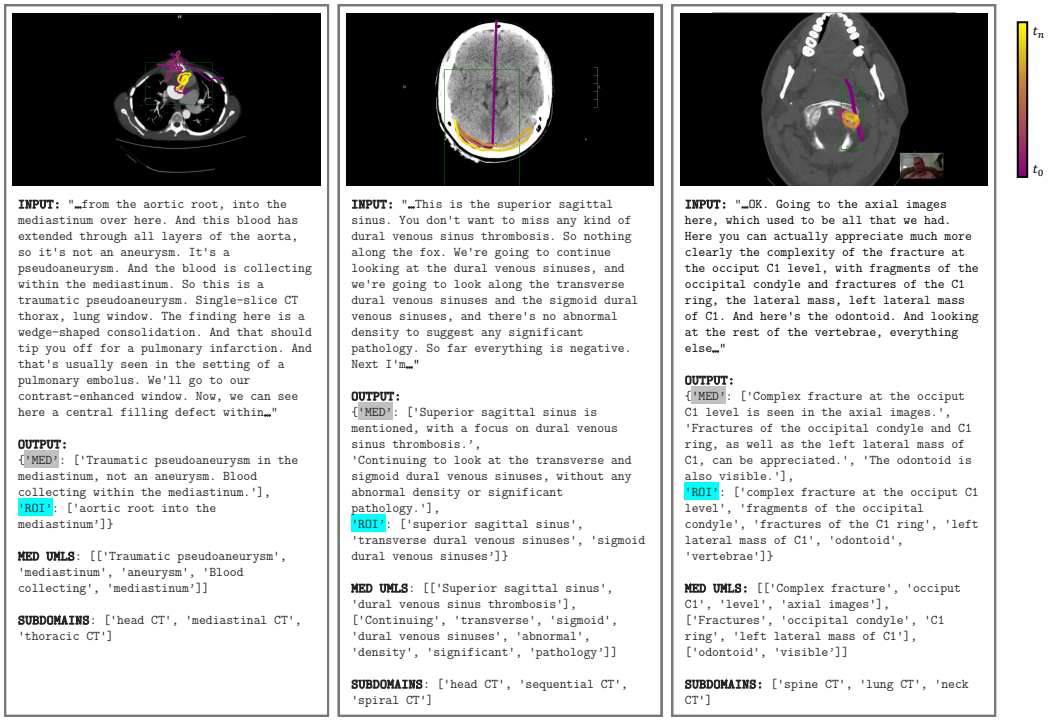


Figure 11: **CT examples** with in the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical and ROI text. **Traces**: Cursor traces and bounding boxes aligned in-time with the raw text. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

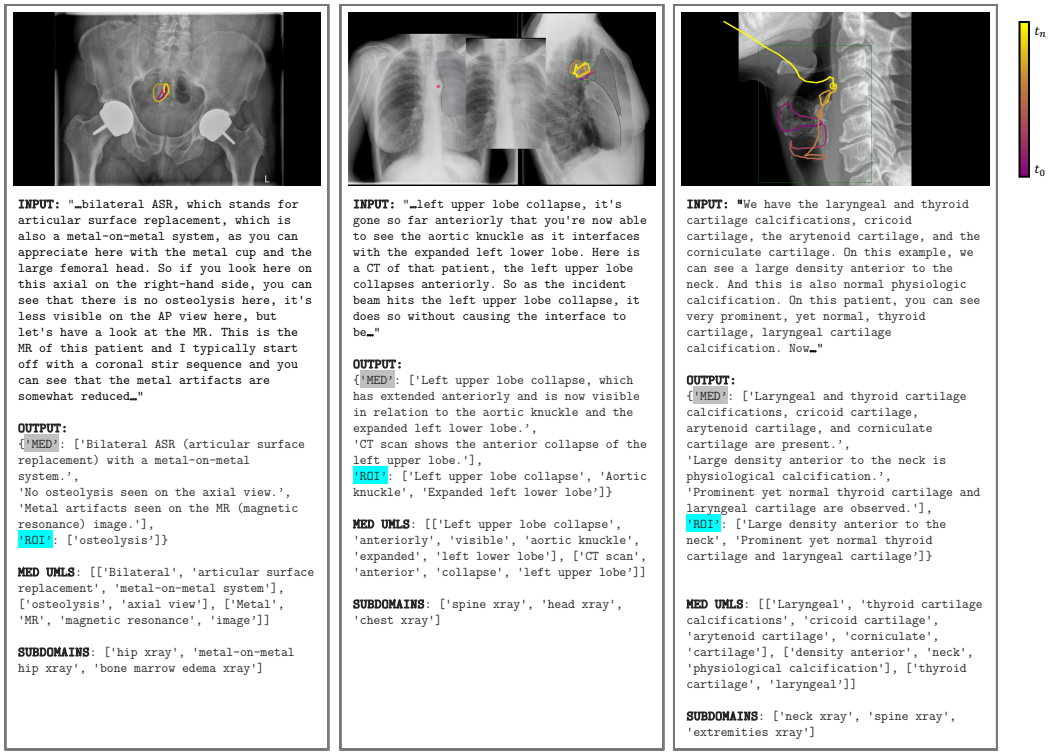


Figure 12: **X-ray examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical and ROI text. **Traces**: Cursor traces and bounding boxes aligned in-time with the raw text. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

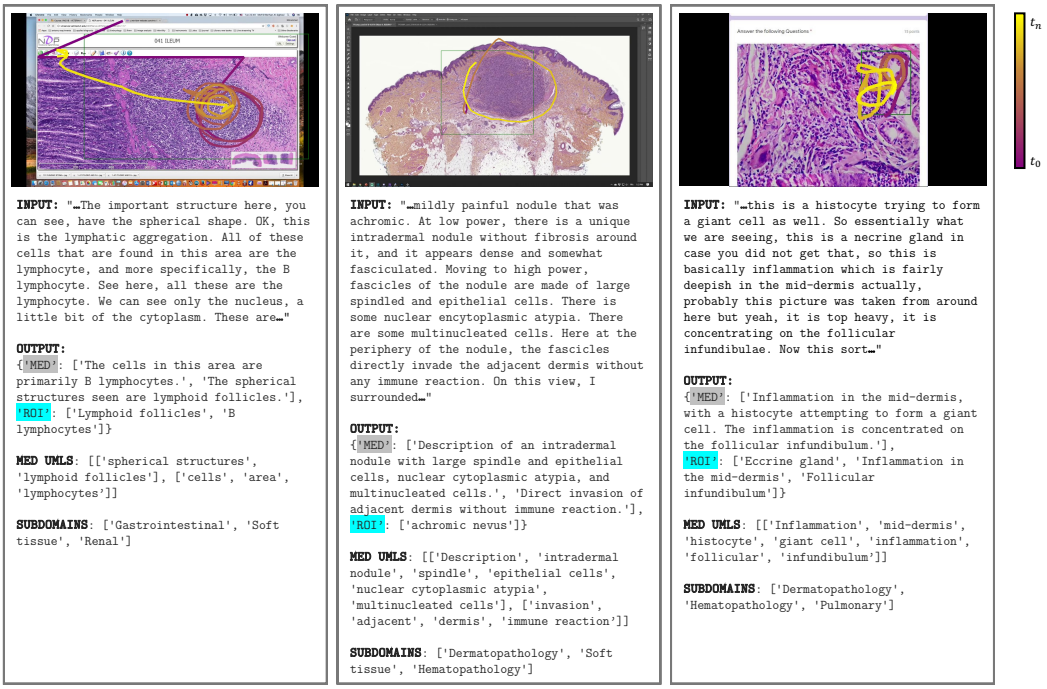


Figure 13: **Histopathology examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical and ROI text. **Traces**: Cursor traces and bounding boxes aligned in-time with the raw text. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

INPUT: "…color and power doppler images. Slide four of three. We're looking at a transverse cine clip of the bladder, and there's a large mass that seems to be growing out of the left wall. It's a polyploid mass with a frond -like margin. Also, notice that there's a faint right ureteral jet. You…"

OUTPUT:
{'MED': ['A large mass growing out of the left wall of the bladder is observed.', 'A faint right ureteral jet flow is also noted.', 'The mass has a polyploid appearance with a frond-like margin.']}

MED UMLS: [['mass', 'left wall', 'bladder', 'frond-like margin'], ['right ureteral jet flow']]

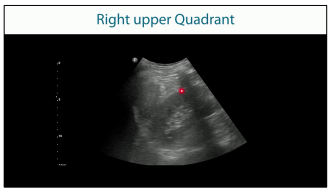
SUBDOMAINS: ['bladder ultrasound', 'abdominal ultrasound', 'pancreas ultrasound']

INPUT: "…The aorta is a pulsatile structure right above the spine, right to the to the right of the spine on our image and to the left of the patient. So here we see the aorta pulsating. And the proximal aorta, you can identify the celiac artery as it branches out into the splenic and hepatic artery, creating what they call a Siegel sign. You can see that on this patient here, the Siegel sign. And this is the proximal aorta, and you want to measure three portions of the aorta, both the proximal, the mid, and the distal. So we're going to go to the mid now as you come down. And what you're noticing what I'm doing is I'm putting a lot of pressure on to get to…"

OUTPUT:
{'MED': ['The aorta is a pulsatile structure located right above the spine.', 'Three portions of the aorta should be measured: proximal, mid, and distal.', 'The celiac artery branches out from the proximal aorta, creating a steal sign.']}

MED UMLS: [['aorta', 'pulsatile structure', 'spine'], ['celiac artery branches', 'proximal aorta', 'steal sign'], ['portions', 'aorta', 'measured', 'proximal', 'mid', 'distal']]

SUBDOMAINS: ['abdominal ultrasound', 'aortic ultrasound', 'doppler ultrasound']



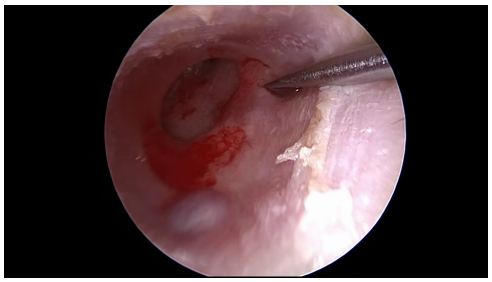
INPUT: "…resemble pleurifusion. As we have seen in the demonstration, we can take a look at the right upper quadrant view, we see the right kidney and the liver, and then we just move more cranial and more dorsal. And then we are basically imaging the zone 4. Here we see the liver, here the pleura, there is even already a B-line, here is the mirror artifact, so we have the diaphragm and here the mirrored liver in the lung. We can visualize this curtain phenomenon, so the lung moving towards the liver and over the liver. If we zoom in, we really can be sure that there is no pleurifusion present, so always look in the near field. In…"

OUTPUT:
{'MED': ['Right upper quadrant view shows the right kidney and liver. Moving cranial and dorsal, imaging zone 4.', 'Pleura is seen with a B-line and mirror artifact. Diaphragm and mirrored liver in the lung.', 'No pleural effusion present when zoomed in.']}

MED UMLS: [['Right upper quadrant view', 'right kidney', 'liver', 'cranial', 'dorsal', 'imaging zone 4'], ['Pleura', 'B-line', 'mirror artifact', 'Diaphragm', 'liver', 'lung'], ['No', 'pleural effusion', 'zoomed']]

SUBDOMAINS: ['echocardiography', 'doppler ultrasound', 'lung ultrasound']

Figure 14: **Ultrasound examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

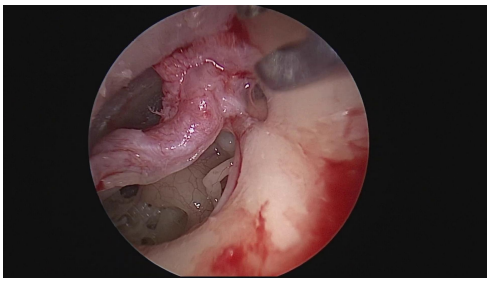


INPUT: "This video will demonstrate a transcanal endoscopic lateral cartilage perichondrograft tympanoplasty. A subtotal perforation identified with chronic meningitis involving the remnant portion of the tympanic membrane. Canal incisions are made outlining the vascular strip after the canal has been infiltrated with local anesthetic. The medial end of the canal incisions along the vascular strip are connected. A…"

OUTPUT:
{'MED': ['The video demonstrates a transcanal endoscopic lateral cartilage perichondrograft tympanoplasty.', 'There is a subtotal perforation of the tympanic membrane.', 'The patient has chronic meningitis.', 'Canal incisions are made to outline the vascular strip.', 'Local anesthetic is used to infiltrate the canal.', 'The medial end of the canal incisions along the vascular strip are connected.']}

MED UMLS: [['video', 'transcanal'], ['subtotal perforation', 'tympanic membrane'], ['patient', 'chronic meningitis'], ['Canal', 'incisions', 'vascular strip'], ['Local', 'infiltrate', 'canal'], ['medial', 'canal', 'incisions', 'vascular strip']]

SUBDOMAINS: ['ears endoscopy', 'throat endoscopy', 'nose endoscopy']



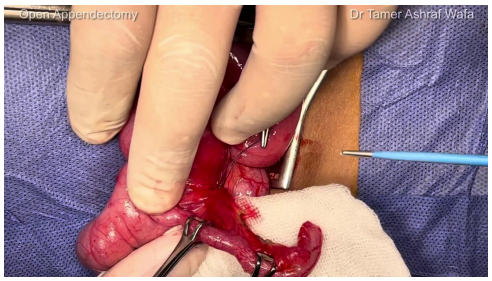
INPUT: "…The middle ear space is entered with a round knife. The middle mucosa is dissected from the corded tympani nerve. The lateral chain is palpated to assess for mobility along with the stapes. A measuring rod is used to estimate the distance between the stapes footplate and the incis long process. The…"

OUTPUT:
{'MED': ['Procedure: Middle ear space is entered with a round knife.', 'Middle mucosa is dissected from the chorda tympani nerve.', 'Evaluation of lateral chain and stapes for mobility.', 'Using a measuring rod to estimate distance between stapes footplate and incus long process.']}

MED UMLS: [['Procedure', 'Middle ear space', 'round knife'], ['Middle mucosa', 'dissected', 'chorda tympani nerve'], ['Evaluation', 'lateral chain and', 'stapes', 'mobility'], ['measuring', 'estimate', 'distance', 'stapes footplate', 'incus']]

SUBDOMAINS: ['ears endoscopy', 'nose endoscopy', 'throat endoscopy']

Figure 15: **Endoscopy examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

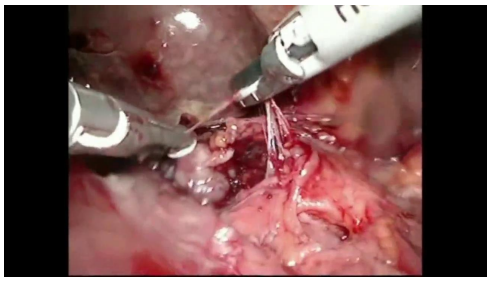


INPUT: "A mosquito forceps is passed at the base of the meso appendix to pass a suture. The artery is carefully ligated in continuity. The meso appendix is then divided by cautery. The mesenteric division continues to the base of the appendix. Any residual small vessels may be controlled by cautery. The base of the appendix is gently crushed just above its origin and the clamp is placed distally a…"

OUTPUT:
{'MED': ['Ligation of the artery in the mesoappendix', 'Division of the mesoappendix by cautery', 'Control of residual small vessels by cautery', 'Crushing of the base of the appendix']}

MED UMLS: [['Ligation', 'artery', 'mesoappendix'], ['Division', 'mesoappendix', 'cautery'], ['Control of residual', 'cautery'], ['Crushing', 'base', 'appendix']]

SUBDOMAINS: ["appendix surgery"]



INPUT: "and in this picture we are trying to dissect out the left hepatic artery and you can see the pulsation of the castor duodenum and the left hepatic artery. Left hepatic artery is on your right hand side of the screen and behind you can see the portal vein that's where the portal vein is. Now we are dissecting the portal vein you can clearly see the dark blue appearance of the portal vein there…"

OUTPUT:
{'MED': ['The left hepatic artery is on the right side of the screen.', 'The portal vein appears dark blue.', 'The pulsation of the castor duodenum can be seen.', 'The left hepatic artery and portal vein are being dissected.']}

MED UMLS: [['left hepatic artery', 'portal vein', 'dissected'], ['left hepatic artery', 'right side', 'screen'], ['pulsation', 'castor duodenum'], ['portal vein']]

SUBDOMAINS: ["gallbladder surgery", "liver surgery"]

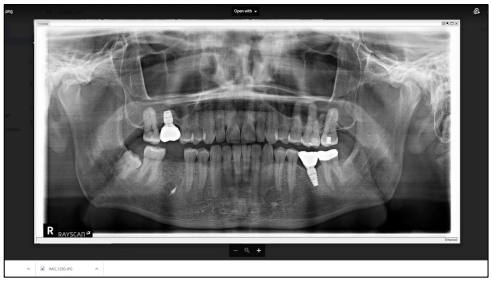
Figure 16: **Surgery examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

INPUT: "This is a 14 year old patient who's got four developing third molars and we can see here by the look of things that you know we're pretty tight for space. So this patient we can pretty safely say at the moment is going to need to have the wisdom teeth out…"

OUTPUT:
{'MED': ['The patient is a 14-year-old with four developing third molars and limited space. Wisdom teeth removal is recommended.', 'Explanation given to the patient about the need for wisdom teeth removal due to lack of space and potential damage to the nerve.', 'Development of roots in wisdom teeth over time.']}

MED UMLS: [['patient', 'molars', 'Wisdom teeth', 'removal', 'recommended'], ['Explanation', 'patient', 'wisdom teeth', 'removal', 'space', 'potential', 'damage', 'nerve'], ['Development', 'roots', 'wisdom teeth']]

SUBDOMAINS: ['orthodontics', 'oral and maxillofacial surgery', 'pediatric dentistry']



INPUT: "And again, I wish I could really show you or share his CVCT scan that he has plenty of thickness of bone in that area. And as you saw in this x -ray, he has plenty of distance, like, you know, about 20 millimeters from the edge of the bone to his nerve that they could have placed a much longer implant and they could have placed a much wider implant. But unfortunately, you…"

OUTPUT:
{'MED': ['The patient has plenty of thickness of bone in a specific area.', 'The patient has plenty of distance (about 20 millimeters) from the edge of the bone to the nerve.', 'The possibility of placing a longer and wider implant is mentioned.']}

MED UMLS: [['patient', 'thickness', 'bone', 'area'], ['patient', 'distance', 'millimeters', 'edge', 'bone', 'nerve'], ['longer', 'implant']]

SUBDOMAINS: ['periodontics', 'prosthodontics', 'oral and maxillofacial surgery']

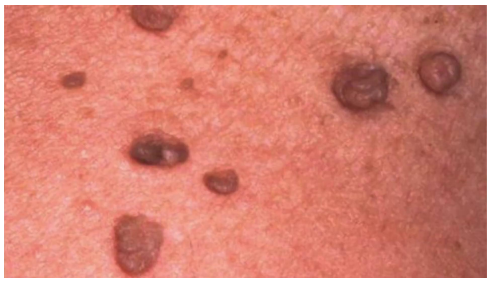
Figure 17: **Dentistry examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

INPUT: "central del representing the follicular pore. The papules are usually about 2 -4 mm in diameter and may be interspersed with small pustules demonstrating minimal inflammatory response. In male patients, lesions appear predominantly in the back in 88%, in the shoulders in 57%, in …"

OUTPUT:
{'MED': ['Papules are usually about 2-4 mm in diameter and may be mixed with small pustules.', 'In male patients, lesions appear predominantly on the back (88%) and shoulders (57%).', 'Minimal inflammatory response.']}

MED UMLS: [['Papules', 'diameter'], ['Minimal', 'inflammatory response'], ['male', 'patients', 'lesions', 'shoulders']]

SUBDOMAINS: ['dermatologic surgery', 'medical dermatology', 'dermatopathology']



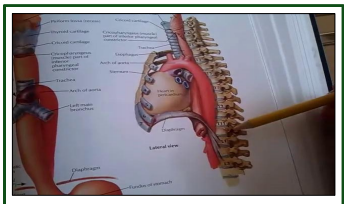
INPUT: "They usually vary in size from about two to six millimeters and are skin - colored, although larger and hyperpigmented lesions can also be seen. One study suggests that 70 % of patients with multiple skin tags may have diabetes. How …"

OUTPUT:
{'MED': ['Skin tags usually vary in size from about two to six millimeters.', 'Skin tags can be skin-colored, larger, and hyperpigmented.', 'Multiple skin tags may be associated with diabetes.']}

MED UMLS: [['Skin', 'tags', 'size', 'millimeters'], ['Skin', 'tags', 'skin-colored', 'larger', 'hyperpigmented'], ['Multiple', 'skin', 'associated with', 'diabetes']]

SUBDOMAINS: ['general dermatology', 'medical dermatology', 'dermatologic surgery']

Figure 18: **Dermatology examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

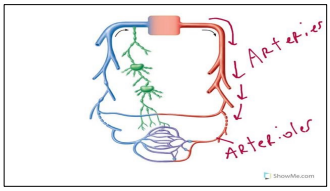


INPUT: "Question 52, oesophagus and stomach. The oesophagus goes through the mediastinum and the stomach lies in the abdominal cavity. The oesophagus lies in front of the third vertebral column from C6 to T11. It follows the curvatures of the vertebral column and it has three constrictions where it starts, the pharyngeo-oesophageal constriction. The second is where it is crossed by aortic arch and left principal bronchus, it's called the…"

OUTPUT:
{'MED': ['The oesophagus goes through the mediastinum and the stomach lies in the abdominal cavity.', 'The oesophagus lies in front of the third vertebral column from C6 to T11.', 'The oesophagus has three constrictions: pharyngo-esophageal constriction, constriction where it is crossed by aortic arch, and left principal bronchus.']}

MED UMLS: [['oesophagus', 'mediastinum', 'stomach', 'abdominal cavity'], ['oesophagus', 'third vertebral column', 'C6', 'T11'], ['oesophagus', 'constrictions', 'pharyngo-esophageal constriction', 'constriction', 'aortic arch', 'left principal bronchus']]

SUBDOMAINS: ['anatomy and physiology', 'surgery', 'clinical skills']

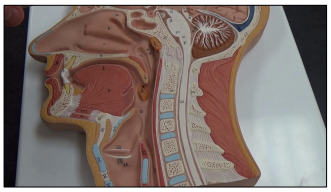


INPUT: "…there's different types. You've got your muscular arteries and you've got your elastic arteries, so on and so forth, that are for conducting and distributing to the body. Then we have the capillary bed which is the exchange vessel here. It's the capillary bed. Then we have the smaller veins called venules which are there that lead up into the veins and then those are eventually going to dump into the IVC and SVC which will then bring the blood back into the heart. So…"

OUTPUT:
{'MED': ['Muscular arteries and elastic arteries are responsible for conducting and distributing blood to the body.', 'Venules lead up into the veins.', 'IVC and SVC bring the blood back into the heart.', 'Capillary bed is the exchange vessel.']}

MED UMLS: [['Muscular arteries', 'elastic arteries', 'conducting', 'distributing', 'blood', 'body'], ['Capillary bed', 'exchange vessel'], ['Venules', 'veins'], ['IVC', 'SVC', 'blood back', 'heart']]

SUBDOMAINS: ['anatomy and physiology', 'pathology', 'internal medicine']



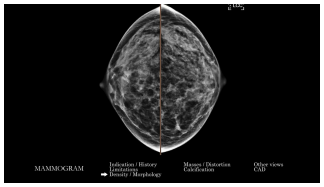
INPUT: "… So you have a superior nasal conch, a middle nasal conch, and an inferior nasal conch. You've got the hard palate here, the soft palate back here. You have the cribriform plate up here. You have the frontal sinus way up here in the frontal bone. You have the sphenoid sinus back here. The internal nare, so remember our external nare was your nostrils, the opening there. Well…"

OUTPUT:
{'MED': ['Superior, middle, and inferior nasal conchae are present.', 'Hard palate and soft palate are located.', 'Cribriform plate is located in the superior region.', 'Frontal sinus is present in the frontal bone.', 'Sphenoid sinus is located posteriorly.', 'Internal nare is the opening inside the nose.']}

MED UMLS: [['Superior', 'middle', 'inferior nasal conchae'], ['Hard palate', 'soft palate'], ['Cribriform plate', 'superior region'], ['Frontal sinus', 'frontal bone'], ['Sphenoid', 'posteriorly'], ['Internal nare', 'opening', 'nose']]

SUBDOMAINS: ['anatomy and physiology', 'respiratory medicine']

Figure 19: **General medical examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.



INPUT: "And so getting a sense of what is the densest area within each breast can help give you, start, you know, keying you into things you want to investigate further. And when you are looking for abnormality, and one of the things we want to keep in mind as we kind of go deeper and then into the tomosynthesis sort of views is anytime we do find an abnormality, we want to keep in mind to look at potential involvement of the skin, the nipple, if there's associated adenopathy, if there's trabecular thickening of the breast, you know. And…"

OUTPUT:
{'MED': ["The densest area within each breast can help in identifying areas for further investigation.", "When looking for abnormality, consider potential involvement of the skin, the nipple, associated adenopathy, and trabecular thickening of the breast."]}

MED UMLS: [["breast", "density", "investigation"], ["abnormality", "skin", "nipple", "adenopathy", "trabecular thickening", "breast"]]

SUBDOMAINS: ["screening mammography", "diagnostic mammography", "digital mammography"]

Figure 20: **Mammography examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

INPUT: "…This is a very mild example, but sometimes they are so dense that it is difficult to see details inside the eye. But people don't seem to have much problem looking outward until a posterior vitreous detachment happens, then they suddenly become noticeable. In summary, the vitreous is a gel with a particular set of structural features which affect its behavior. As time goes by, changes…"

OUTPUT:
{'MED': ['Dense vitreous can make it difficult to see details inside the eye.', 'Vitreous is a gel with specific structural features that affect its behavior.', 'Posterior vitreous detachment can suddenly make dense vitreous noticeable.']}

MED UMLS: [['Dense', 'vitreous', 'eye'], ['Posterior vitreous detachment', 'dense', 'vitreous'], ['Vitreous', 'gel', 'structural', 'behavior']]

SUBDOMAINS: [['retina ophthalmology'], ['ocular oncology ophthalmology'], ['ophthalmic pathology']]

Figure 21: **Ophthalmology examples** within the MEDICALNARRATIVES dataset. **Input**: raw input text from ASR. **Output**: the output from the LLM, with denoised medical. **UMLS**: UMLS entities extracted from the medical text. **Subdomain**: classification of the sample into finer-grained subdomains.

Domain	Dataset	Total samples	Train	Test	Image Size
CT	LIDC-IDRI [11]	10005	7004	3002	512 × 512
	TCGA-LUAD [5]	48931	34252	14679	512 × 512
	WORD [78]	30495	21347	9149	512 × 512
	Positive Videos	1612	1128	484	
X-ray	ChestX-ray14 [124]	112120	78484	33636	1024 × 1024
	GRAZPEDWRI-DX [84]	20327	14229	6098	660 × 1660
	Shoulder X-ray Classification [22]	841	589	252	
	Digital Knee X-ray [42]	1650	1155	495	300 × 162
	MURA [101]	40561	28393	12168	1500 × 2000
	Positive Videos	692	484	208	1440 × 1080
MRI	fastMRI [132]	58847	41193	17654	320 × 320
	Duke-Breast-Cancer-MRI [106]	922	645	277	256 × 256
	Medical Segmentation Decathlon [8]	2633	1843	790	256 × 256
	Positive Videos	118	83	35	
Dermatology	Dermnet [40]	19500	13650	5850	720 × 472
	DDI [30]	656	459	197	300 × 300
	7-point [64]	2045	1432	614	480 × 720
	ISIC [104]	33126	23188	9938	
	Fitzpatrick 17k [43]	16577	11604	4973	
	HAM10000 [118]	10015	7011	3005	800 × 600
Endoscopy	KVASIR [95]	8000	5600	2400	720 × 576
	ITEC LapGyn4 [71]	59439	41607	17832	256 × 256
	Red Lesion Endoscopy [28]	3895	2727	1169	320 × 320
	FetReg [15]	12334	8634	3700	
	TMEDOM [7]	956	669	287	
	Positive Videos	9496	6647	2849	
US	COVID-19 Ultrasound [19]	59	41	18	
	BUSI [4]	780	546	234	500 × 500
	DDTI [93]	134	94	40	560 × 360
	MMOTU [136]	1639	1147	492	330$\tilde{8}$88 × 218$\tilde{6}$57
	HC18 [120]	1334	934	400	
	EchoNet-Dynamic [89]	10030	7021	3009	112 × 112
	Positive Videos	1874	1312	562	
Dentistry	Panoramic radiography [102]	598	419	179	2041 × 1024
	ODSI-DB [52]	316	221	95	
	DENTEX 2023 [46]	2332	1632	700	
	Dental Calculus [94]	220	154	66	
	Vident-lab [63]	15110	10577	4533	416 × 320
	Dental condition [107]	1296	907	389	612 × 408
	Oral cancer [130]	144	101	43	
	Dental cavity [108]	176	123	53	
Surg	SARAS-ESAD [16]	27175	19023	8153	1920 × 1080
	CholecSeg8k [49]	8080	5656	2424	854 × 480
	DeSmoke-LAP [91]	6000	4200	1800	
	Surgical Hands [76]	2838	1987	851	
	m2caiSeg [79]	307	215	92	716 × 402
	NeuroSurgicalTools [20]	2476	1733	743	612 × 460
	ROBUST-MIS 2019 [103]	10000	7000	3000	960 × 540
Optha	Cataracts [6]	35127	24589	10538	1920 × 1080
	Ocular Disease Recognition [2]	3358	2351	1007	512 × 512
	MeDAL Retina [85]	2181	1527	654	768 × 768
	RFMID [90]	3200	2240	960	2144 × 1424
	Glaucoma Detection [1]	650	455	195	3072 × 2048
	DRIVE [116]	40	28	12	584 × 565
Mammo	CBIS-DDSM [110]	10239	7167	3072	
	CDD-CESM [66]	2006	1404	602	2355x1315
	CMMD [29]	5202	3641	1561	
Genmed	LAION [112]	10861	7603	3258	
Non-medical	Celeb [75]	202599	60780	28364	178 × 218
	Places [137]	10624928	637496	399497	200 × 200
	AI2D [65]	4903	3432	1471	
	DocFig [57]	33028	26422	6606	
	SciFig-Pilot [61]	263952	211162	52790	
	SlideImages [83]	3452	2762	690	
	TextVQA [115]	25119	20095	5024	
	SlideShare-1M [9]	977605	782084	195521	
	Negative Videos	23956	19165	4791	
	EgoHands [14]	4800	3840	960	720 × 1080
	11k Hands [3]	11076	8861	2215	1600 × 1200
	IPN Hand [17]	95021	76017	19004	640 × 480

Table 11: Datasets used to train ResNet50 and ViT-Small medical image classifiers, used in Section A.2 and Section B.2.

System Prompt:

```
Imagine you are a text classifier. Classify the given text into the given types.
```

User Prompt:

```
Imagine you are a text classifier. Classify the given text into the following types:
X-ray, MRI, CT, ultrasound, endoscopy, surgery, dermatology, dentistry, mammography,
ophthalmology, general medical. Output only the relevant types in an ordered python
list.
```

Few-shot examples:

```
"Scanning from the right upper quadrant to the right lower quadrant. Acute right
lower quadrant pain. Gallbladder and hyperechoic area. The clinical examination
shows the external, internal obliques, and the transversus abdominis muscles. The
presence of the hyperechoic area may indicate fat stranding or fat edema, which can
be seen in appendicitis. Fat stranding or fat edema is sometimes seen in
appendicitis, but the appendix in this case was normal. Color-doppler image shows no
flow within the fat. Color-doppler image shows no flow within the fat. Hyperechoic
fat is seen in case number two, with a bit of edema. Iliac vessels can be seen in
the clip. Hyperechoic, hyperechoic fat that appears edematous. Identification of
iliac vessels (artery and vein), right ovary. Iliac artery and iliac vein.
Hyperechoic, edematous fat. Right ovary. Abdominal CT scan in axial plane.":

"['ultrasound', 'ct']"
```

```
INPUT: "Case 52 shows the strength of hydrogen bonding. The third case is an
aminotasis and the third case is a natural hemangioma. The learning point is that
natural hemangioma appears very large and uniform in all three recipes. The third
case is impossible to be a hemangioma and the third is a hydatid cyst. The weak area
is the hemangioma. In this equal image of the three infections, we see a large
hemangioma with a central lamp. Despite the rest of the hemangioma, which is a group
of internal recipes containing blood.. The doctor also has brain notes. Possible
notes of carcinoma or hepatocellular notes. We have an old CT brain note and this is
a brain note and the correct answer is metastasis because we have a normal note sign
here. Increase in size of hepatocellular carcinoma over a period of 6 years. The
learning point is that hepatocellular carcinoma can significantly increase in size
over time."

OUTPUT: "['ct', 'ultrasound', 'mri']"
```

Figure 22: The GPT-3.5 Turbo prompts used to determine whether a video contains discussion of multiple medical domains, with few-shot examples.

System Prompt:

```
You are a computed tomography (CT) text classifier that classifies CT-related text
into CT subdomains.
```

User Prompt:

```
Imagine you are a text classifier. Classify the given text into the following ct
types: sequential CT, spiral CT, electron beam tomography, spectral CT, CT perfusion
imaging, PET CT, head CT, neck CT, lung CT, CT angiography, cardiac CT, abdominal
CT, pelvic CT, facial structure CT, spine CT, extremities CT. Output only the top 3
types in an ordered python list:
```

Few-shot examples:

```
"There is subtle enhancement of mucosin in the distal coma bowel duct. No dilatation
of the bowel duct upstream. There is subtle enhancement of mucosin in the distal
common bile duct. There is no dilatation of the bowel duct upstream.  The patient
had a follow-up visit two years later. The gallbladder wall was not enhancing on the
previous visit. There is biliary dilatation and a mass-like appearance in this area.
Biliary dilatation is present. There is mass-like irregular thickening of the common
bile duct, giving a rat-like appearance. The appearance is consistent with
cholangiocarcinoma. Biliary dilatation and mass-like irregular thickening of the
common bile duct, resembling a rat-like appearance, indicative of
cholangiocarcinoma. Cholangiocarcinoma is observed in ERCP. The cholangiocarcinoma
appears mass-like and axial...":

"['abdominal CT', 'CT angiography', 'pelvic CT']"
```

INPUT: "The patient is a 68-year-old male who has symptoms of loss of orientation, difficulty with ambulation, and changes in behavior. A head CT scan has revealed some findings. Patient presenting with loss of orientation, difficulty with ambulation, and behavior changes. Large abnormality seen on head CT scan. Thoracic imaging is the focus. Head CT scan shows a large abnormality with low attenuation areas representing edema and areas of higher attenuation suggesting hemorrhage with surrounding edema and mass effect. Low attenuation areas in the brain represent edema. Areas of higher attenuation suggest hemorrhage and surrounding edema. This could potentially be a primary brain tumor, such as an astrocytoma or glioblastoma. The presence of hemorrhage, surrounding edema, and mass effect suggests a primary brain tumor, potentially an astrocytoma or glioblastoma. The possibility of metastasis from lung cancer is also considered. Suspicion of a brain tumor as a possible metastasis, with immediate consideration of lung cancer. Impressive abnormality in the left hilar region, potentially indicating an abnormality or pathology in that area. Impressive abnormality in the left hilar region on chest CT..."

OUTPUT: "['brain CT', 'lung CT', 'head CT']"

Figure 23: The GPT-3.5 Turbo prompts used for determining which specific sub-domains are discussed in a video, with few-shot examples.

System Prompt:

```
You are a research paper processor which splits the captions of figures into sub-
captions that correspond with subfigures.
```

User Prompt:

```
Given a caption of an image containing sub-images, please decompose the caption in
accordance with each sub-image. Be sure to adhere to the following guidelines:
1. Preserve the original wording of the caption. Refrain from adding new
information, summaries, or introductions.
2. Omit references to the index or number of the sub-images, such as (xx), left,
right, etc.
3. There might be a common prefix or suffix caption shared among all sub-images at
the beginning, end, or within the caption. Please incorporate the prefix or suffix
into each sub-image's caption. If one subcaption contains context for multiple other
subcaptions, add that context to each of the relevant subcaptions.
4. The final output should be in JSON format, with an outer field 'subcaptions',
with a value that is a list of 'subfigure' and 'subcaption' dictionaries.
5. If a subfigure contains more nested figures, i.e. subfigure (A) contains
references to (left) and (right), add a field called "location" that stores the
"left" or "right".
6. If there are no references to sub-images, give a single subcaption with label
"A".
```

Few-shot examples:

```
"(a) H&E image of a breast tumor tissue. Fluorescently labeled markers superimposed
as green color on the H&E image, (b) \u03b2-catenin, (c) pan-keratin, and (d) smooth
muscle \u03b1-actin, markers.":

{"subcaptions": [
        {"subfigure": "a",
         "subcaption": "H&E image of a breast tumor tissue. Fluorescently labeled
markers superimposed as green color on the H&E image."},
        {"subfigure": "b",
         "subcaption": "H&E image of a breast tumor tissue. \u03b2-catenin marker
superimposed as green color on the H&E image."},
        {"subfigure": "c",
         "subcaption": "H&E image of a breast tumor tissue. pan-keratin marker
superimposed as green color on the H&E image."},
        {"subfigure": "d",
         "subcaption": "H&E image of a breast tumor tissue. smooth muscle \u03b1-
actin marker superimposed as green color on the H&E image."}]}
```

INPUT: "Angiography of the celiac artery showed that the dilated omental artery was revealed continuously from the splenic artery (A), turned over, headed toward the vascular sac (B), and returned to the omental vein (white arrow) and left colonic vein (white arrowhead) (C). A stenosis (black arrow) due to ligation at the time of splenectomy was observed in the splenic artery (D)."

OUTPUT:

```
{"subcaptions": [
            {"subfigure": "A",
             "subcaption": "Angiography of the celiac artery showed that the dilated
omental artery was revealed continuously from the splenic artery."},
            {"subfigure": "B",
             "subcaption": "Angiography of the celiac artery showed that the dilated
omental artery turned over and headed toward the vascular sac."},
            {"subfigure": "C",
             "subcaption": "Angiography of the celiac artery showed that the dilated
omental artery returned to the omental vein (white arrow) and left colonic vein
(white arrowhead)."},
            {"subfigure": "D",
             "subcaption": "Angiography of the celiac artery showed a stenosis
(black arrow) due to ligation at the time of splenectomy in the splenic artery."}]}
```

Figure 24: The GPT-3.5 Turbo prompts used for splitting a compound figure caption into sub-captions, with few-shot examples.

System Prompt:

```
You are a research paper processor which splits the captions of figures into sub-captions that correspond with subfigures.
```

User Prompt:

```
You are given the sub-figures and sub-captions of a figure in a medical article. You are also given the inline mentions of the figure. Return a JSON where each inline mention is paired to the most relevant subfigure, and all strings in the JSONs have double quotes. If the inline mentions are broadly applicable to all subcaptions, add the inline mentions to each subcaption. If the inline mentions aren't applicable to any subcaption, return an empty JSON.
```

Few-shot examples:

```
{"img_inline_sentences": ["As shown in Figure 7, lithium-treatment caused marked decrease in the mean protein abundance of NKCC2 in the renal medulla of WT mice.",
        "9-fold higher protein abundance as compared to the WT mice (Figure 7B).",
        "In contrast, in the cortex lithium administration caused significant decreases in NKCC2 protein abundance in both WT and KO mice with no difference in their mean values (Figure 7D).",
        "Interestingly, similar to AQP2 protein abundance, the mean NKCC2 protein abundance in control diet-fed P2Y2 KO mice was 2-fold higher as compared to the corresponding value in WT mice (Figure 7D)."],
 "subcaptions": ["A", "B", "C", "D"]}:

{"A": ["As shown in Figure 7, lithium-treatment caused marked decrease in the mean protein abundance of NKCC2 in the renal medulla of WT mice."],
 "B": ["As shown in Figure 7, lithium-treatment caused marked decrease in the mean protein abundance of NKCC2 in the renal medulla of WT mice.", "9-fold higher protein abundance as compared to the WT mice (Figure 7B)."],
 "C": ["As shown in Figure 7, lithium-treatment caused marked decrease in the mean protein abundance of NKCC2 in the renal medulla of WT mice."],
 "D": ["As shown in Figure 7, lithium-treatment caused marked decrease in the mean protein abundance of NKCC2 in the renal medulla of WT mice.", "In contrast, in the cortex lithium administration caused significant decreases in NKCC2 protein abundance in both WT and KO mice with no difference in their mean values (Figure 7D).", "Interestingly, similar to AQP2 protein abundance, the mean NKCC2 protein abundance in control diet-fed P2Y2 KO mice was 2-fold higher as compared to the corresponding value in WT mice (Figure 7D)."]}
```

INPUT:

```
{"img_inline_sentences": ["The mass demonstrated a scattered calcification and expansive bony destruction (Fig 2)."],
 "subcaptions": ["A", "B"]}
```

OUTPUT:

```
{"A": ["The mass demonstrated a scattered calcification and expansive bony destruction (Fig 2)."],
 "B": ["The mass demonstrated a scattered calcification and expansive bony destruction (Fig 2)."]}
```

Figure 25: The GPT-3.5 Turbo prompts used for pairing inline references of a figure with the most relevant sub-figures, with few-shot examples.