
MEDICALNARRATIVES: Connecting Medical Vision and Language with Localized Narratives

Wisdom O. Ikezogwo*
University of Washington
wisdomik@cs.washington.edu

Kevin M. Zhang*
University of Washington
kzhang20@cs.washington.edu

Mehmet Saygin Seyfioglu
University of Washington, Amazon
msaygin@cs.washington.edu

Abstract

Multi-modal models are data hungry. While datasets with natural images are abundant, medical image datasets can not afford the same luxury. To enable representation learning for medical images at scale, we turn to YouTube, a platform with a large reservoir of open-source medical pedagogical videos. We curate MedicalNarratives, a dataset 4.7M medical image-text pairs, with 1M samples containing dense annotations in the form of spatial traces (and bounding boxes), and 118K videos centered on the trace event (with aligned text), enabling spatiotemporal grounding beyond single frames. Similar to *think-aloud* studies where instructors speak while hovering their mouse cursor movements over relevant image regions, 1M images in MedicalNarratives contains localized mouse traces in image pixels, creating a spatial and temporal association between the text and pixels. To evaluate the utility of MedicalNarratives, we train GENMEDCLIP with a CLIP-like objective using our dataset spanning 12 medical domains. GENMEDCLIP outperforms previous state-of-the-art models on all 12 domains on a newly constructed medical imaging benchmark. [\[Data\]](#)

1 Introduction

Analyzing medical images requires simultaneous spatial localization and semantic understanding Morita et al. [82]. An expert has to extract visual clues from image regions and combine them with retrieved knowledge from memory, arriving at a diagnosis. This process requires connecting individual spatial image regions to clinical concepts, often utilizing a segmental approach to avoid errors. [82]. In medical image analysis, typically semantic tasks like classification, captioning, and retrieval are explored exclusively from spatial tasks like detection [96, 129, 113], or segmentation [127, 25]. This can be attributed to the lack of large grounded multimodal datasets to train such models. Recent work like MedTrinity-25M [128], attempts to address this by releasing a multimodal dataset with spatial annotations, but relies on sub-optimally pretrained models to generate text descriptions and Regions of Interests (ROIs) for medical images lacking ground truth annotations, potentially propagating model biases and errors.

While data collection costs are steep, certain protocols balance ease of collection and training utility. Specifically, Localized Narratives [96, 122] proposes a dataset of image, text, and grounding traces, curated by leveraging human annotators to describe an image vocally while simultaneously moving a computer mouse to the regions they describe, resulting in holistic grounded descriptions. This

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

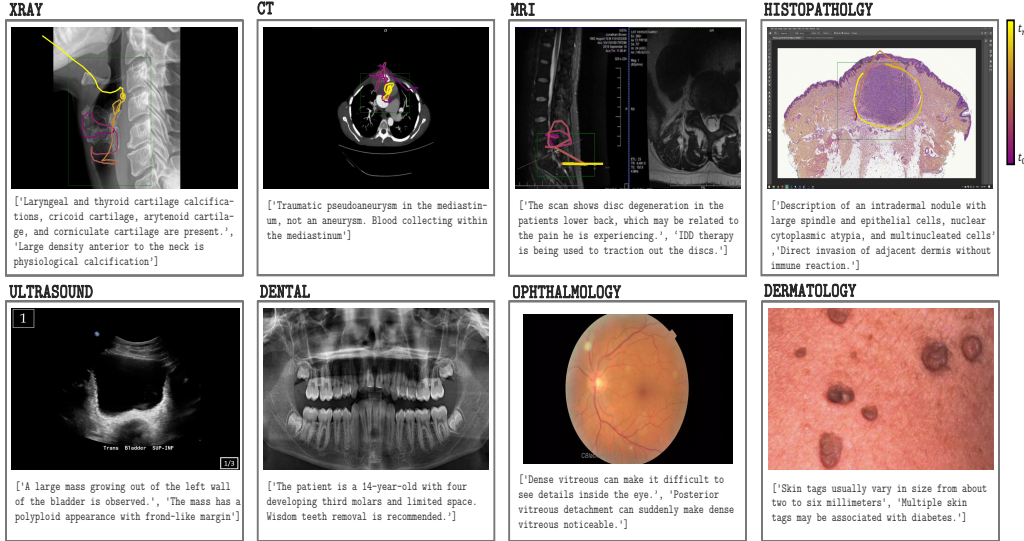


Figure 1: **MEDICALNARRATIVES**: Examples from our medical imaging modalities, excluding surgery, endoscopy, and general medical images due to their graphic nature. These samples are selected from interleaved video samples, with each sample showing the image, denoised text, and spatial traces & bbox aligned in-time on 4 domains. See section E in the Appendix for more examples and raw input text.

protocol of collecting grounded vision-language (VL) datasets does not have strict spatial annotations, yet, it captures strong spatial correlations to the description with every trace point, making the protocol uniquely easier to undertake and capture data en-mass as it appeals to the human nature to point while describing a scene [60, 123, 45]. Localized narratives have been used to train models on semantic tasks [96, 122, 133], and spatially-aware multimodal language models (MLM) like PixelLM [129], and Molmo [31] and other generative image models [68, 131].

To address these limitations, we draw inspiration from how medical experts naturally communicate and teach. In the joint field of cognitive psychology and medical imaging, studying how medical experts analyze patient data, studies leverage the think-aloud protocol [34] to capture data for various types of analysis, where experts verbalize their thoughts as they perform a task, and some studies capture their eye gaze/cursor localizing the image regions they focus on [74, 48]. This protocol has been used to collect medical datasets [92, 81], including the Tufts dental x-ray database [92], which captures a multimodal dataset incorporating radiologist expertise through eye-tracking and the think-aloud protocol.

We propose **MEDICALNARRATIVES** a dataset that leverages pedagogical medical videos where instructors narrate descriptions while pointing to relevant regions with their cursor, closely mimicking the think-aloud protocol used in clinical practice and the Localized Narratives protocol. Our dataset contains 4.7 million image-text pairs across 11 medical modalities and 1 pseudo-medical domain, with interleaving samples between varying modalities (e.g., X-ray and CT for the same patient), which we argue improves downstream performance as these samples connect multiple visual and textual concepts. Importantly, 1M of these samples are grounded in expert traces that can be reformatted into bounding boxes or masks (see Figure 5), serving as pretraining data for various tasks.

To test the base utility of our dataset, we train a vision-language model (**GENMEDCLIP**) on our dataset and evaluate it on a new benchmark of datasets that cut across 11 medical modalities. On both classification and retrieval, we see our trained **GENMEDCLIP** model outperform prior SOTA models like **BIOMEDCLIP** in both tasks with an average of 3% and 14% respectively. While the proposed dataset is a combination of data from **A**. Temporal unstructured sources like video, and **B**. Static structured sources like scientific articles, unlike prior work that solely leverage one source, our experiments show that the utility of the dataset increases with more data from video, with a net difference of 11.65% on classification tasks and 53.6% on zeroshot retrieval tasks. Finally, we show

the utility of traces with qualitative examples, converting traces into segmentation using pretrained interactive medical image segmentation (IMIS) models like ScribblePrompt [127, 25]. We hope future works leverage the dataset to train more grounded generative models similar to Quilt-LLaVA [113], LLaVA-Med++ [128], and PixelLM [129] as well as spatially-controlled medical image captioning models [96]. To bolster other use cases, we also release the constituting video clip IDs (useful for obtaining the videos) and many other metadata, including modality type and UMLS entities.

In addition to the centered still images with traces, we provide paired videos (temporal windows around trace start/end), preserving narration alignment to the pointing behavior. This addition allows models to learn spatiotemporal grounding (e.g., cursor trajectories across frames) rather than static spatial associations alone.

2 Related work

2.1 Vision Language representation

Vision-language (multi-modal) models have evolved over time in both supervised and self-supervised paradigms; in recent studies, contrastive self-supervision objectives [98, 135, 56] that learn by matching paired-modality embeddings have outperformed prior work [77, 72, 24] in downstream tasks and, more importantly, perform better at zeros-shot tasks or on emergent domains for which disparate modalities share a paired domain [39, 139]. In medical imaging, early studies in radiology [135, 50] were pre-trained on specific x-ray images and their reports, and more recently domain specific VL models have pushed the SOTA on various tasks with models developed for Ophthalmology [114], Histopathology [53, 51], Computed Tomography [47], Mammography [23], Dermatology [67], Ultrasound/Echocardiography [26]. These models work well for the specific domains they are trained on and not for other domains, which may not have enough data to train for, hence the push for more general medical models [134, 138, 119].

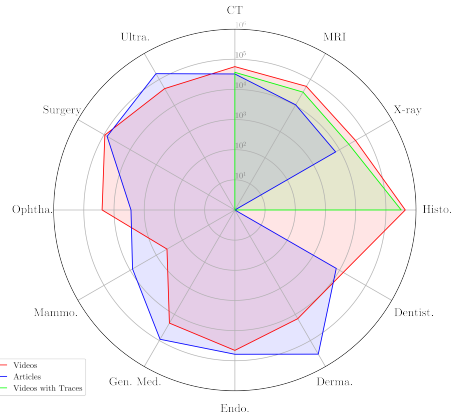


Figure 2: Breakdown of MEDICALNARRATIVES in size by modalities across both video and article subsets.

2.2 Medical (Localized) Narratives

In training these VL models, much research effort is used in sourcing, filtering, and curating medical image(s)-text(s) paired data for pre-training, mostly sourcing general and specific medical domain datasets from Medical reports [59], PubMed [35, 105, 36, 134], books [36], social-media [51, 53], YouTube/videos [53, 113] or mixtures of these [128].

The utilization of these data for dense tasks like segmentation and detection (open-closed vocabulary) is limited as they do not provide any spatial annotation localizing regions of the images to specific labels/text. In contrast, every word in a localized narrative [96, 122, 133, 113] is grounded to a

Dataset	Size	Source	Medical Only	Domains	Open Source Data/Code	Novel Images	Video	Text	Spatial Annot.
PMC-15M [134]	15M	A	✗	30	✗/✓	✓	✗	Captions	✗
PMC-FG-64M [134]	46M	A	✗	30	✗/✗	✓	✗	Captions	✗
PMC-CLIP [73]	15M	A	✗	12	✓/✗	✓	✗	Captions	✗
MedTrinity-25M [128]	25M	P	✓	10	✓/✗	✗	✗	Captions/labels	Seg. mask
MedicalNarratives	4.7M	V+A	✓	12	✓/✓	✓	✓	Expert	Traces

Table 1: **Comparison with large-scale medical datasets.** In the table, A: Articles, V: Videos, and P: pre-published datasets. Open-Source column is formatted *data/pipeline*.

region of the representative image by the point/trace captured. This datasets have been used to train models for semantic reasoning [96, 122, 133], and for dense tasks [41, 38, 32], and they also support training both generative multimodal language models [129, 113, 122] and generative image models [68, 131]. Specifically in medical image analysis, Quilt-LLaVA [113] adopts this paired data structure for training its histopathology chatbot with improved spatial reasoning, and Pathnarrative’s [133] hierarchical decision-to-reason localized narrative structure, enables classification and captioning tasks, offering explainable insights that improve human-AI collaboration in pathological diagnosis.

3 MEDICALNARRATIVES:

Curating a vision-language dataset with spatial traces from unstructured pedagogy videos is a non-trivial, as many videos either lack voiced audio, fail to contain medically relevant content, or have insufficient medical relevance. In addition, conventional automatic speech recognition (ASR) systems also struggle with the specialized requirements of medical language transcription, necessitating a non-trivial solution. The de-noising of text and image modalities adds further complexity as the videos are typically conversational and, therefore, inherently noisy. Instructors often record both relevant and irrelevant visual content in their videos, making extracting frames at static intervals non-representative of the medical data contained in the video.

To collect MEDICALNARRATIVES, we leverage insights from Quilt-1M [53] prior work, we trained models and handcrafted algorithms that leverage the nuances in the instructors’ textual and visual behavior, ensuring accurate collection and alignment of both modalities. Finally, we manually filter noisy samples out and employ other heuristics and models to remove artifacts like faces and irrelevant traces. In this section, we start by characterizing the dataset 3.1, then we discuss the methods used to source and filter the dataset 3.2, localize traces 3.3, and discuss the implicit interleaving property 3.4. See Figure 3 for our pipeline and section A and B of the appendix for how we align the data samples A.5.

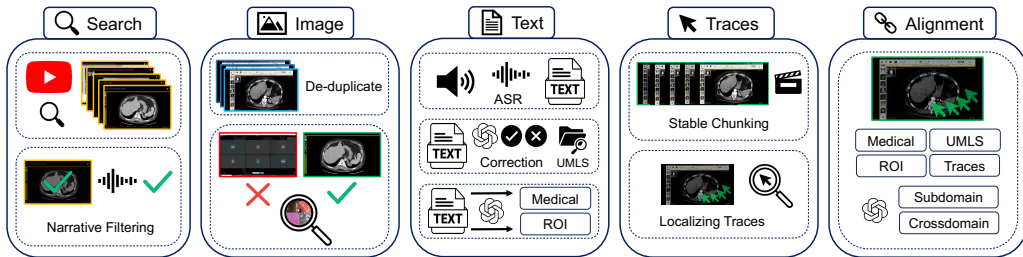


Figure 3: The data curation pipeline for the Video subset of the MEDICALNARRATIVES dataset. **Search**: YouTube video-first search strategy, with filtering by pre-trained classifiers and heuristics. **Image**: Extracting keyframes of a video, denoising, and identifying medical images. **Text**: ASR transcription, text correction with LLMs, and medical/ROI text extraction. **Traces**: Identifying stable chunks of a video, then localizing cursor traces within each chunk. **Alignment**: Mapping medical/ROI text, traces, and images together. Samples are classified into finer-grained subdomains, and samples with discussions of multiple domains are identified with LLMs.

3.1 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. We compare our dataset against other medical pretraining datasets in Table 1 across various key distinctive properties including data source, and spatial annotation. Figure 2 shows the distribution of MEDICALNARRATIVES across various medical modalities, and Table 9 and Table 10 in the appendix give detailed statistics across all medical modalities.

3.1.1 Video-Subset

We searched over 738K videos and extracted 74K narrative-styled videos that passed our heuristics and had relevant medical imaging pedagogy, a 10.1% yield making up a total of 4526 hours of video.

In total, we collect 736K unique images with an average size of H: 1493px and W: 923px and 1.63M image-text pairs from videos with 1M of these samples grounded with traces, these samples cover 101.8M number of unique trace points yielding 547K number of unique bounding boxes with an average size of H: 316px and W: 357px across the 4 domains with traces: CT, MRI, X-ray, and Histopathology. The dataset contains 118K videos, collected at the boundaries of the traces, with a min, max, and average duration of 3.3, 228.8 and 24.2 seconds. The mean length of the text captions is 29.87 words, with an average of 2.48 medical sentences per image. Our dataset spans over 4M UMLS entities from those mentioned in the text with over 300K unique entities across medical (e.g., findings, or disease) and non-medical (e.g., governmental or regulatory activity) semantic types.

3.1.2 Article Subset

We extract 5.4M articles from PubMed [86], with 23M figures, after filtering for medical figures only, we obtain 1.03M figures from 273K articles, and after sub-figure separation, we have an average of 2.62 subfigure-subcaption pairs per-article figure, with an average of 45.45 words per-caption.

3.1.3 Quality

Unlike localized narratives [96, 122] where localization accuracy can be measured by comparing against human annotation, none of our videos to our knowledge have any structured human spatial annotation to compare against. Nonetheless, to evaluate our pipeline’s performance, we assess several aspects. First, we calculate the precision of our LLM’s corrections by dividing the number of conditioned misspelled errors replaced (i.e., passed the UMLS check) by the total number of conditioned misspelled words found, yielding an average of 47.99%. We also determined the unconditioned precision of the LLM, similar to the previous step, and found it to be 17.58%. Therefore, we replace our detected incorrect words with the LLM’s correction 47.99% of the time, and 17.58% of the time we replace the LLM’s detected errors with its correction. To estimate the ASR model’s transcription performance, we compute the total number of errors replaced (both conditioned and unconditioned) and divide it by the total number of words in each video, resulting in an average ASR error rate of 0.81%. Also note that, by prompting the LLM to extract only medically relevant text, we further eliminate identifiable information, such as clinic addresses, from our dataset.

Since the dataset was collected for pretraining, we do not upsample the text after correcting for errors and filtering bad images; on average, each image is paired with approx. 83 words of relevant text and traces when available and validated.

3.2 Data Sourcing and Filtration

This involves (a) sourcing video/article data across 12 medical imaging domains, (b) filtering videos/articles, (c) denoising the captured images, captions, and trace modalities, and (d) aligning all modalities. We detail our method and highlight key contributions in sections A and A of the appendix.

3.2.1 Text Extraction and Denoising

Videos: In line with Quilt-1M [53] we leverage an open-source ASR model - Whisper [99] to transcribe all speech from the selected videos, correcting transcription errors using language model with specialized prompts (see section A.4 for details on the error-extracting algorithm).

Articles: Similarly we parse each article’s XML document, extracting each figure’s caption and inline mentions (see B.1). Since many sub-figures are typically grouped into single large figures, we split the compounded figure captions into sub-captions, leveraging an LM to find and split sub-captions due to the non-triviality of identifying enumerations in the text and splitting the captions correctly (see B.4). Furthermore, we refine the inline mentions of a figure and match them to specific sub-captions/sub-figures (see B.6).

3.2.2 Image Extraction and Denoising

Videos: For each video, we identify medical key-frames and subsequently leverage these frames’ times to split the video into time-intervals called *chunks* from which to extract representative image(s). To extract representative image(s), we use the median image of stable frames in each chunk if they

exist, else we de-duplicate the captured key-frames, exploiting the human tendency in pedagogy videos to pause while explaining and pointing [96, 45, 113].

Articles: For scientific documents, we extract the figures as images. However, many of these figures contain multiple sub-figures which can take nonconventional grid shapes and are labeled irregularly, making the task of splitting into sub-figures and pairing with the correct sub-caption non-trivial. Since most compound figure layouts are not uniform and vary in the whitespace in between sub-figures, we train an object detection model based on the YOLO architecture [58] on sub-figure annotation datasets MedlCaT and ImageCLEF 2016 [117, 37]. See more details in section B.3

3.3 Localizing Traces in Videos

Extracting the trace/cursor location from medical clips poses a significant challenge due to certain domain properties including homogeneity in color and texture, significant black/white background, and presence of artifacts in videos such as minor pixel variations and variations in the narrators’ cursor movement speed and style. We modify the methodology proposed by Quilt-LLaVA [113] centered around the observation that narrators typically pause before signaling with their cursor. We isolate segments in the video where the background is static, termed stable chunks. To do so, we develop a frame-differencing approach to detect chunks with minimal background movement. Our algorithm computes the absolute difference between consecutive frames and then applies a Gaussian filter for adaptive thresholding to detect frames with minor changes.

Due to the homogeneity of medical images, naive pixel-wise differencing produces many false positives, misidentifying changing chunks as stable. To mitigate this, we incorporate a perceptual metric, using the structural similarity index measure (SSIM) on randomly sampled patches to verify frame changes. Next, for each stable chunk, we compute a median frame (in the pixel domain) as a background reference, subtract it from individual frames, and apply a threshold to isolate the cursor. We then extract trace points by identifying the maximum pixel value coordinates. This method assumes minimal background motion, but subtle movements, such as narrator facial expressions, can interfere. To address this, we apply a face detection model [111] to mask distractions, ensuring focus remains on cursor movement. This algorithm provides a robust and generalizable approach for capturing cursor traces from medical videos.

Extracting videos

For each detected trace segment we extract the video clip at the start and end of the trace temporal window aligned with the text.

3.4 Cross-Modal Interleaved Samples

A key advantage of MEDICALNARRATIVES is its interleaved multi-modal nature. This manifests in two ways: (1) **Video-based interleaving:** Medical pedagogy videos frequently present multiple imaging modalities for the same patient. Instructors naturally explain relationships between these modalities in a single narrative, creating one-to-many mappings between textual descriptions and images. This allows our model to learn connections between modalities through shared textual context (see Figure 9 in Appendix). (2) **Sample-based interleaving:** Articles and Videos often contain images with multiple sub-images showing different modalities accompanied by a unified caption. This structure similarly reinforces cross-modal relationships. (see MRI example in Figure 1 and Figure 9 in the appendix). This interleaved nature of MEDICALNARRATIVES significantly enhances cross-modal retrieval capabilities, as shown in Sec. 4.4. We open-source our dataset with modality tags which can be used to identify cross-modal samples.

4 GENMEDCLIP: Experiments

We test the utility of MEDICALNARRATIVES on two medically relevant tasks image classification (zeroshot and linear probing) and cross-modal information retrieval (zero-shot) across all in-domain modalities. We select the Contrastive Language-Image Pre-training (CLIP) objective [98] to pre-train a VL model: GENMEDCLIP. We train several models, varying the image, and text encoders while making adaptations in line with prior work on the choice of encoders and text tokenization for

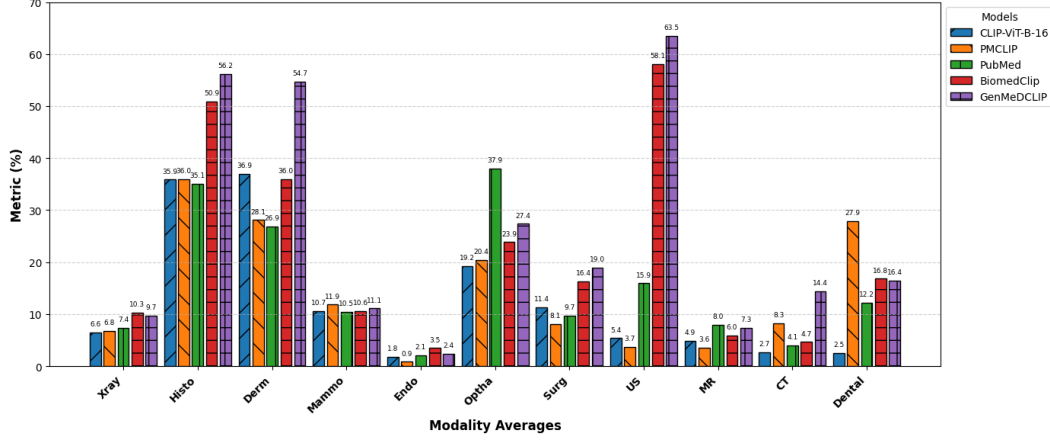


Figure 4: **Zeroshot Classification Results** shows that our model GENMEDCLIP outperforms all other baselines including the out-of-domain CLIP, and biomedical vision-language models BIOMED-CLIP, and PUBMEDCLIP across the constructed medical benchmark which covers all 11 medical domains represented. The metric for Xray and Mammography is mean average precision while the rest is accuracy.

improved performance [134, 53]. For the image tower, we finetune Vision Transformers (ViT-Base) [33] models pretrained using a supervised cross-entropy objective (ViT-Base-16 and ViT-Base-32 [126]) and unsupervised contrastive objective ViT-Base-16) [98], on 224*224 pixel images. On the text tower, we use GPT2 [97] with a context length of 77, and BioMedBert [44] with context size to 256. To train our models we utilize OpenClip [54] on 4 Nvidia A40 GPUs for 20 iterations. To ensure a fair comparison with baselines, we trained three different variants of our model: GENMEDCLIP-32: with ViT-B/32 image-tower and GPT2/77 text-tower architecture, GENMEDCLIP-PMB: with ViT-B/16 image-tower and Bert/256 BiomedBert [44] text-tower, and GENMEDCLIP-PMB: with ViT-B/16 image-tower and GPT2/77 [44] text-tower; all finetuned for 20 epochs over our train-set, while data split ablation models are trained for 6 epochs. (Details in Section D in Appendix.)

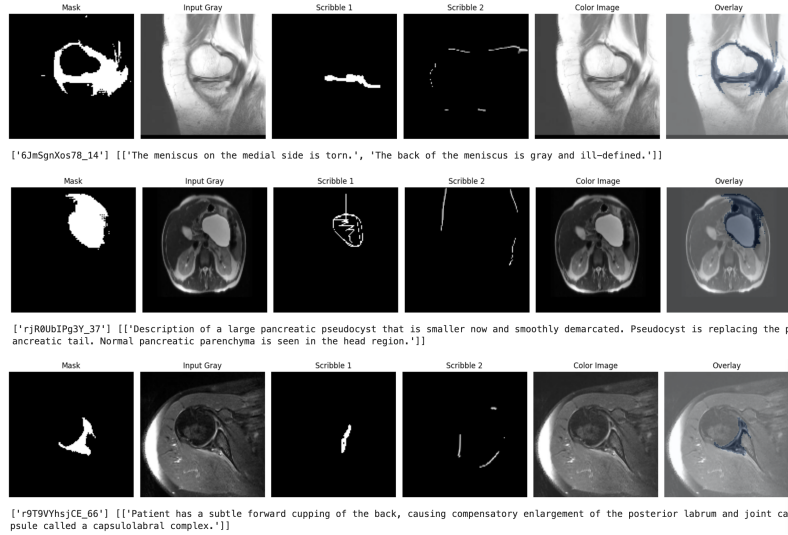


Figure 5: Using trace as prompts for segmentation using ScribblePrompt-SAM. (Right) resulting mask from trace (Center).

Model	Isic	Til	Pcam	Mhist	Nck	Mammo	Avg
CLIP-ViT-B-16 [98]	71.23	91.23	82.42	63.97	92.26	83.30	80.74
PUBMEDCLIP [35]	68.58	91.32	84.07	72.16	92.29	83.90	82.06
BIOMEDCLIP [134]	68.25	91.82	83.43	66.73	93.05	83.70	81.17
GENMEDCLIP-32	72.75	93.26	86.77	72.06	92.77	83.70	83.55
GENMEDCLIP-PMB	69.38	91.51	84.54	67.66	88.02	84.20	80.88
GENMEDCLIP	74.87	93.34	87.69	72.16	90.84	84.90	83.97

Table 2: **Linear Probing** results across datasets representing Dermatology (Isic), Histopathology (pcam, mhist, nck), and Mammography (vinDr-Mammo) classification tasks. GENMEDCLIP outperforms all baselines showing the capacity of our model to be fine-tuned for downstream tasks. The metric used is accuracy.

4.1 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, CT, and ultrasound we use their respective subsets from **RadImageNet** [80] dataset. For Xray, we evaluate on **VinDr-CXR** Chest Xrays [87] test set and report the mean average precision (mAP), similarly for Mammography we use **VinDr-Mammo** [88] and report the mAP. We evaluate surgical organ classification using **Dresden** [21], and for endoscopy, we test on all procedure images in **GastroVison** [55]. 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. To evaluate the Ophthalmology domain we evaluate on **G1020** [13] a retinal fundus glaucoma dataset and on **Optical Coherence Tomography Dataset** (OCTDL) [70]. We evaluate the Histopathology domain on the following datasets: **PatchCamelyon** [121], **NCT-CRC-HE-100K** [62], **BACH** [10], **Osteo** [12], **SkinCancer** [69], **MHIST** [125], **LC25000** [18], and on TCGA-TIL [109]. Please see section D in the appendix for more details.

Models	Data	T2I retrieval			I2T retrieval			Avg
		@5	@50	@200	@5	@50	@200	
CLIP-ViT-B-16 [98]	-	3.48	20.38	35.69	3.56	20.39	35.42	19.82
PMC-CLIP [73]	A	0.01	0.33	1.18	0.01	0.34	1.24	0.52
PUBMEDCLIP [35]	A	1.44	12.68	25.44	1.10	12.30	24.07	12.84
BIOMEDCLIP [134]	A	16.50	51.48	67.46	15.71	48.85	64.61	44.10
GENMEDCLIP-32	V+A	22.36	76.33	88.60	20.75	75.15	88.23	61.90
GENMEDCLIP-PMB	V+A	28.29	82.91	92.43	29.21	82.91	92.43	68.03
GENMEDCLIP	V+A	34.89	83.83	92.27	34.26	83.48	92.32	70.17
Data Split Ablation								
GENMEDCLIP *	A	2.11	12.89	22.36	2.35	13.66	22.81	12.70
GENMEDCLIP *	V+A	28.01	80.56	90.96	27.48	79.95	90.85	66.30

Table 3: **Retrieval** results on our held-out set of 16K samples across all medical domains, show that our model GENMEDCLIP outperforms all other baselines on both Zeroshot image-to-text and vice-versa text-to-image retrieval task. In the table, A: Articles, V: Videos, and * represents a shorter number of fine-tuning iterations

4.2 Zero-shot classification

We evaluate our model’s zero-shot performance against three state-of-the-art models: CLIP, BIOMED-CLIP, PMC-CLIP, and PUBMEDCLIP. In Figure 4 and Table 8, each domain in the benchmark is represented by a set of dataset(s). The prompts used for these evaluations are presented in Table 7 in the Appendix. Across the benchmark, our model averages the following GENMEDCLIP-32:

31.33%, GENMEDCLIP-PMB: 31.46%, and GENMEDCLIP: 32.55% metric all outperforming BIOMEDCLIP with 27.80% by 4.75%. Specifically, as shown in Figure 4, GENMEDCLIP outperforms all baselines in five medical domains: Histopathology, Dermatology, Surgery, Ultrasound, and CT, while remaining comparable to baselines in the Chest X-ray, Endoscopy, Mammography, and MRI domains.

4.3 Supervised linear probing

We assess the full-shot performance of our model by conducting linear probing with 100% of the training data; we report the average accuracy over all benchmark evaluation across five distinct datasets, specifically those with dedicated training and testing sets among our external datasets in Dermatology, Histopathology, and Mammography. Remarkably, our model, utilizing the ViT-B/32 architecture with GPT/77, outperforms its counterparts, BIOMEDCLIP, and CLIP, in most datasets. Overall, on average GENMEDCLIP outperforms all other models including BIOMEDCLIP and PUBMEDCLIP with over 2.8%, and over 1.9% respectively.

4.4 Cross-Modal Retrieval

We evaluate cross-modal retrieval performance by examining both zero-shot text-to-image and image-to-text retrieval capabilities. To do so, we leverage a randomly selected held-out partition of MEDICALNARRATIVES, not used in training our models. The held-out set contains 16K image-text pairs with the following medical modality distribution: 1756 X-ray, 1237 MRI, 1851 CT, 1351 Ultrasound, 1744 Surgery, 1346 Endoscopy, 1189 Dermatology, 1216 Dentistry, 1151 Ophthalmology, 1000 Histopathology, 1299 General Medical, 1149 Other (Mammo etc) image-text pairs. Retrieval in our study is done by identifying the nearest neighbors for each modality and then determining whether the corresponding pair is within the top N nearest neighbors, where $N \in \{1, 50, 200\}$, mimicking several medical search tasks. Results in Table 3 shows that on average GENMEDCLIP outperforms all baselines and specifically outperforms BIOMEDCLIP by 26.07%, The results also confirm the observation made in BIOMEDCLIP [134] where the general CLIP model outperforms the in-domain model PUBMEDCLIP by 6.98%

4.5 Data Split Ablation

As seen in Tables 8 and 3, we ablate the added utility of capturing pedagogy video data by training two models, one trained solely on articles and the other on both articles and video data. The results show that adding the video-derived data leads to higher average classification (11.65% higher) and retrieval (53.6% higher) performance. We suspect that the dynamic nature of *video frames* introduces diverse vantage points, partially explaining these improvements. We also see that classification performance across all Article only trained models except Biomedclip is comparable further buttressing the impact of video as a data source.

5 Discussion and Limitations

MEDICALNARRATIVES contributes a robust pipeline for grounded multi-modal data curation across noisy, unstructured, and diverse medical modalities sources. We believe it would catalyze progress in novel medical vision-language tasks, like spatially-controllable report generation [96, 129], and interactive medical image segmentation [25, 127]. Figure 5 illustrates how the captured traces, albeit noisy and not expert-validated, can serve as conditioning for semi-automatic segmentation models like ScribblePrompt [127] toward plausible object boundaries and for exploring visual grounding toward text+trace conditioned segmentation.

Spatial Reasoning Applications

Beyond retrieval/classification, the trace-aligned samples provide direct supervision for grounded language and localization tasks without dense masks. Following Localized Narratives [96, 122], each word/phrase co-occurring with a trace segment supplies weak phrase-region links for grounded captioning, referring expressions, and spatial relation inference. High-dwell (i.e. spatial regions where narrators focus on) segments of traces can be collapsed into pointing cues to train pointing-based

medical MLMs, such as Molmo [31]. Because our traces are timestamped, the same supervision naturally extends to video: the trajectory of the cursor across frames yields spatiotemporal grounding suitable for dynamic "point-while-describe" models and temporal localization (e.g., axial CT sweeps or ultrasound). The dataset also supports tasks that predict spatial traces as additional loss objectives toward imbuing models with spatial understanding [129] and panoptic narrative grounding objectives [41] are directly enabled by these trace–text alignments.

For dense prediction and controllable generation, traces act as sparse supervision that can be transformed into inputs for interactive medical image segmentation (IMIS) models (e.g., as points/scribbles for ScribblePrompt) to bootstrap pseudo-masks and iteratively refine them [127]. Coupling trace spans with their co-mentioned phrases supplies approximately localized phrase labels for open-vocabulary detection/segmentation similar to phrase-region training [32, 38].

Finally, the same signals can guide where and what to synthesize in text-to-image/volume models, using trace/point conditioning alongside clinical text to localize clinical entities, while enabling spatially controllable medical report generation [68, 131, 113, 129].

Limitations

1. Our dataset lacks human-annotated bounding boxes, limiting overlap assessment between video traces and annotations, restricting dense tasks like semantic segmentation.
2. Our dataset overrepresents abnormal cases, an underlying bias, reflected in hospital practices where imaging follows clinical suspicion. This may impact model generalizability and introduce bias in clinical decision support.
3. While we showcase the capacity of traces to be useful for IMIS task, this work does not leverage the traces to train any models toward downstream spatial or spatial aware models like PixelLM [129]. We leave this to future work.

6 Conclusion

This study proposes a robust protocol for collecting and annotating medical narratives. Our curated dataset MEDICALNARRATIVES, which follows the Narratives Annotation Protocol addresses the specific challenges of medical data collection at scale balancing the relationship between downstream utility and ease/cost of collection. We argue that this protocol can serve as the de facto standard for annotating future multimodal medical datasets, particularly given its flexibility in capturing grounded text describing medical images effectively. We demonstrate a strong performance over prior models, across both classification and retrieval tasks, establishing new state-of-the-art results and demonstrating the effectiveness of data filtration methods on model performance, as we train our GENMEDCLIP on 4.7 samples while BIOMEDCLIP trains on over 15M samples. We hope future work leverages our developed models, dataset, and protocol.

7 Acknowledgments

We thank Fateemeh Ghezloo for the initial discussions that led to this work. We thank Ranjay Krishna and Linda Shapiro for their writing contributions and support of the study. We also acknowledge Microsoft for providing OpenAI credits, and partial funding through a Population Health Initiative at University of Washington.

References

- [1] Glaucoma detection. <https://www.kaggle.com/datasets/sshikamaru/glaucoma-detection>.
- [2] Ocular disease recognition. <https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k>. Ocular Disease Recognition — Kaggle.
- [3] M. Afifi. 11k hands: Gender recognition and biometric identification using a large dataset of hand images, 2018. URL <https://arxiv.org/abs/1711.04322>.

- [4] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. ISSN 2352-3409. doi: <https://doi.org/10.1016/j.dib.2019.104863>. URL <https://www.sciencedirect.com/science/article/pii/S2352340919312181>.
- [5] B. Albertina, M. Watson, C. Holback, R. Jarosz, S. Kirk, Y. Lee, K. Rieger-Christ, and J. Lemmerman. The cancer genome atlas lung adenocarcinoma collection (tcga-luad) (version 4), 2016. URL <https://doi.org/10.7937/K9/TCIA.2016.JGNIHEP5>.
- [6] H. ALHAJJ, M. Lamard, P.-h. Conze, B. Cochener, and G. Quellec. Cataracts, 2021. URL <https://dx.doi.org/10.21227/ac97-8m18>.
- [7] A. Alhudhaif, Z. Cömert, and K. Polat. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. *PeerJ Comput Sci*, 7:e405, Feb 2021. doi: 10.7717/peerj-cs.405.
- [8] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbeláez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein, and M. J. Cardoso. The medical segmentation decathlon. *Nature Communications*, 13(1), July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30695-9. URL <http://dx.doi.org/10.1038/s41467-022-30695-9>.
- [9] A. Araujo, J. Chaves, H. Lakshman, R. Angst, and B. Girod. Large-scale query-by-image video retrieval using bloom filters, 2016. URL <https://arxiv.org/abs/1604.07939>.
- [10] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.
- [11] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. Van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Castele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke. Data from lidc-idri, 2015. URL <https://doi.org/10.7937/K9/TCIA.2015.L09QL9SX>.
- [12] H. B. Arunachalam, R. Mishra, O. Daescu, K. Cederberg, D. Rakheja, A. Sengupta, D. Leonard, R. Hallac, and P. Leavey. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706, 2019.
- [13] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.
- [14] S. Bambach, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [15] S. Bano, A. Casella, F. Vasconcelos, S. Moccia, G. Attilakos, R. Wimalasundera, A. L. David, D. Paladini, J. Deprest, E. D. Momi, L. S. Mattos, and D. Stoyanov. Fetreg: Placental vessel segmentation and registration in fetoscopy challenge dataset, 2021. URL <https://arxiv.org/abs/2106.05923>.

- [16] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo, K. Wang, L. Li, B. Wang, S. Zhao, L. Li, A. Stabile, F. Setti, R. Muradore, and F. Cuzzolin. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods, 2021. URL <https://arxiv.org/abs/2104.03178>.
- [17] G. Benitez-Garcia, J. Olivares-Mercado, G. Sanchez-Perez, and K. Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition, 2020. URL <https://arxiv.org/abs/2005.02134>.
- [18] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.
- [19] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, J. Goulet, A. Au-jayeb, M. Moor, B. Rieck, and K. Borgwardt. Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Applied Sciences*, 11(2), 2021. ISSN 2076-3417. doi: 10.3390/app11020672. URL <https://www.mdpi.com/2076-3417/11/2/672>.
- [20] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE Transactions on Medical Imaging*, 34(12):2603–2617, 2015. doi: 10.1109/TMI.2015.2450831.
- [21] M. Carstens, F. M. Rinner, S. Bodenstedt, A. C. Jenke, J. Weitz, M. Distler, S. Speidel, and F. R. Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):1–8, 2023.
- [22] W. Celniak, M. Wodziński, A. Jurgas, S. Burti, A. Zotti, M. Atzori, H. Müller, and T. Banzato. Improving the classification of veterinary thoracic radiographs through inter-species and inter-pathology self-supervised pre-training of deep learning models. *Sci Rep*, 13(1):19518, Nov 2023. doi: 10.1038/s41598-023-46345-z.
- [23] X. Chen, Y. Li, M. Hu, E. Salari, X. Chen, R. L. J. Qiu, B. Zheng, and X. Yang. Mammo-clip: Leveraging contrastive language-image pre-training (clip) for enhanced breast cancer diagnosis with multi-view mammography, 2024. URL <https://arxiv.org/abs/2404.15946>.
- [24] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Universal image-text representation learning, 2020. URL <https://arxiv.org/abs/1909.11740>.
- [25] J. Cheng, B. Fu, J. Ye, G. Wang, T. Li, H. Wang, R. Li, H. Yao, J. Chen, J. Li, et al. Interactive medical image segmentation: A benchmark dataset and baseline. *arXiv preprint arXiv:2411.12814*, 2024.
- [26] M. Christensen, M. Vukadinovic, N. Yuan, and D. Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8, 2024.
- [27] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.
- [28] P. Coelho, A. Pereira, A. Leite, M. Salgado, and A. Cunha. A deep learning approach for red lesions detection in video capsule endoscopies. In A. Campilho, F. Karray, and B. ter Haar Romeny, editors, *Image Analysis and Recognition*, pages 553–561, Cham, 2018. Springer International Publishing.
- [29] C. Cui, L. Li, H. Cai, Z. Fan, L. Zhang, T. Dan, J. Li, and J. Wang. The chinese mammography database (cmmd): An online mammography database with biopsy confirmed types for machine diagnosis of breast, 2021. URL <https://doi.org/10.7937/tcia.eqde-4b16>.
- [30] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.

- [31] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, J. Lu, T. Anderson, E. Branson, K. Ehsani, H. Ngo, Y. Chen, A. Patel, M. Yatskar, C. Callison-Burch, A. Head, R. Hendrix, F. Bastani, E. VanderBilt, N. Lambert, Y. Chou, A. Chheda, J. Sparks, S. Skjonsberg, M. Schmitz, A. Sarnat, B. Bischoff, P. Walsh, C. Newell, P. Wolters, T. Gupta, K.-H. Zeng, J. Borchardt, D. Groeneveld, J. Dumas, C. Nam, S. Lebrecht, C. Wittlif, C. Schoenick, O. Michel, R. Krishna, L. Weihs, N. A. Smith, H. Hajishirzi, R. Girshick, A. Farhadi, and A. Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models, 2024. URL <https://arxiv.org/abs/2409.17146>.
- [32] K. Desai, I. Misra, J. Johnson, and L. van der Maaten. Scaling up instance segmentation using approximately localized phrases. In *British Machine Vision Conference*, 2022. URL <https://api.semanticscholar.org/CorpusID:256904321>.
- [33] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [34] S. J. Durning, J. Graner, A. R. Artino Jr, L. N. Pangaro, T. Beckman, E. Holmboe, T. Oakes, M. Roy, G. Riedy, V. Capaldi, et al. Using functional neuroimaging combined with a think-aloud protocol to explore clinical reasoning expertise in internal medicine. *Military Medicine*, 177(suppl_9):72–78, 2012.
- [35] S. Eslami, G. de Melo, and C. Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.
- [36] J. Gamper and N. Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16549–16559, 2021.
- [37] A. García Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller. Overview of the ImageCLEF 2016 medical task. In *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, September 2016.
- [38] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Scaling open-vocabulary image segmentation with image-level labels, 2022. URL <https://arxiv.org/abs/2112.12143>.
- [39] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.
- [40] S. Goel. Dermnet. <https://www.kaggle.com/datasets/shubhamgoel27/dermnet>.
- [41] C. González, N. Ayobi, I. Hernández, J. Hernández, J. Pont-Tuset, and P. Arbeláez. Panoptic narrative grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1364–1373, October 2021.
- [42] P. S. Gornale and P. Patravali. Digital knee x-ray images, 2020.
- [43] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, 2021. URL <https://arxiv.org/abs/2104.09957>.
- [44] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.
- [45] M. Gygli and V. Ferrari. Efficient object annotation via speaking and pointing. *International Journal of Computer Vision*, 128(5):1061–1075, 2020.
- [46] I. E. Hamamci, S. Er, E. Simsar, A. E. Yuksel, S. Gultekin, S. D. Ozdemir, K. Yang, H. B. Li, S. Pati, B. Stadlinger, et al. Dentex: An abnormal tooth detection with dental enumeration and diagnosis benchmark for panoramic x-rays. *arXiv preprint arXiv:2305.19112*, 2023.

- [47] I. E. Hamamci, S. Er, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, M. F. Dasdelen, O. F. Durugol, B. Wittmann, T. Amiranashvili, E. Simsar, M. Simsar, E. B. Erdemir, A. Alanbay, A. Sekuboyina, B. Lafci, C. Bluethgen, M. K. Ozdemir, and B. Menze. Developing generalist foundation models from a multimodal dataset for 3d computed tomography, 2024. URL <https://arxiv.org/abs/2403.17834>.
- [48] L. Helle. Prospects and pitfalls in combining eye-tracking data and verbal reports. *Frontline Learning Research*, 5(3):1–12, 2017.
- [49] W. Y. Hong, C. L. Kao, Y. H. Kuo, J. R. Wang, W. L. Chang, and C. S. Shih. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80, 2020. URL <https://arxiv.org/abs/2012.12453>.
- [50] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3942–3951. IEEE, 2021.
- [51] Z. Huang, F. Bianchi, M. Yuksekgonul, T. Montine, and J. Zou. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, 2023. doi: 10.1101/2023.03.29.534834. URL <https://www.biorxiv.org/content/early/2023/04/01/2023.03.29.534834>.
- [52] J. Hyttinen, P. Fält, H. Jäsberg, A. Kullaa, and M. Hauta-Kasari. Oral and dental spectral image database—odsi-db. *Applied Sciences*, 10(20), 2020. ISSN 2076-3417. doi: 10.3390/app10207246. URL <https://www.mdpi.com/2076-3417/10/20/7246>.
- [53] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024.
- [54] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. URL <https://doi.org/10.5281/zenodo.5143773>. If you use this software, please cite it as below.
- [55] D. Jha, V. Sharma, N. Dasu, N. K. Tomar, S. Hicks, M. K. Bhuyan, P. K. Das, M. A. Riegler, P. Halvorsen, U. Bagci, and T. de Lange. Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection, 2023. URL <https://arxiv.org/abs/2307.08140>.
- [56] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [57] K. V. Jobin, A. Mondal, and C. V. Jawahar. Docfigure: A dataset for scientific document figure classification. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 1:74–79, 2019. URL <https://api.semanticscholar.org/CorpusID:207959459>.
- [58] G. Jocher, J. Qiu, and A. Chaurasia. Ultralytics YOLO, 2023. URL <https://github.com/ultralytics/ultralytics>.
- [59] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0. URL <https://doi.org/10.1038/s41597-019-0322-0>.
- [60] D. Kahneman. *Attention and effort*, volume 1063. Citeseer, 1973.
- [61] Z. Karishma. Scientific document figure extraction, clustering and classification, 2021. [32].
- [62] J. N. Kather, N. Halama, and A. Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo*10, 5281, 2018.

- [63] E. Katsaros, P. K. Ostrowski, K. Wlodarczak, E. Lewandowska, J. Ruminski, D. Siupka-Mroz, L. Lassmann, A. Jezierska, and D. Wesierski. Multi-task video enhancement for dental interventions, 2022. ISSN 1611-3349. URL http://dx.doi.org/10.1007/978-3-031-16449-1_18.
- [64] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019. doi: 10.1109/JBHI.2018.2824327.
- [65] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi. A diagram is worth a dozen images, 2016. URL <https://arxiv.org/abs/1603.07396>.
- [66] R. Khaled, M. Helal, O. Alfarghaly, O. Mokhtar, A. Elkorany, H. El Kassas, and A. Fahmy. Categorized digital database for low energy and subtracted contrast enhanced spectral mammography images, 2021. URL <https://doi.org/10.7937/29kw-ae92>.
- [67] C. Kim, S. U. Gadgil, A. J. DeGrave, J. A. Omiye, Z. R. Cai, R. Daneshjou, and S.-I. Lee. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine*, pages 1–12, 2024.
- [68] J. Y. Koh, J. Baldridge, H. Lee, and Y. Yang. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 237–246, January 2021.
- [69] K. Kriegsmann, F. Lobers, C. Zgorzelski, J. Kriegsmann, C. Janßen, R. R. Meliß, T. Muley, U. Sack, G. Steinbuss, and M. Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12, 2022.
- [70] M. Kulyabin, A. Zhdanov, A. Nikiforova, A. Stepichev, A. Kuznetsova, M. Ronkin, V. Borisov, A. Bogachev, S. Korotkich, P. A. Constable, and A. Maier. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific Data*, 11(1), Apr. 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03182-7. URL <http://dx.doi.org/10.1038/s41597-024-03182-7>.
- [71] A. Leibetseder, S. Petscharnig, M. J. Primus, S. Kletz, B. Münzer, K. Schöffmann, and J. Keckstein. Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018. URL <https://api.semanticscholar.org/CorpusID:49643457>.
- [72] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020. URL <https://arxiv.org/abs/2004.06165>.
- [73] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.
- [74] S. Littlefair, P. Brennan, W. Reed, M. Williams, and M. W. Pietrzyk. Does the thinking aloud condition affect the search for pulmonary nodules? In *Medical imaging 2012: image perception, observer performance, and technology assessment*, volume 8318, pages 366–374. SPIE, 2012.
- [75] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [76] N. Louis, L. Zhou, S. J. Yule, R. D. Dias, M. Manojlovich, F. D. Pagani, D. S. Likosky, and J. J. Corso. Temporally guided articulated hand pose tracking in surgical videos, 2021. URL <https://arxiv.org/abs/2101.04281>.

- [77] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.
- [78] X. Luo, W. Liao, J. Xiao, J. Chen, T. Song, X. Zhang, K. Li, D. N. Metaxas, G. Wang, and S. Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, Nov. 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102642. URL <http://dx.doi.org/10.1016/j.media.2022.102642>.
- [79] S. Maqbool, A. Riaz, H. Sajid, and O. Hasan. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. *arXiv preprint arXiv:2008.10134*, 2020.
- [80] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.
- [81] J. Molin, M. Fjeld, C. Mello-Thoms, and C. Lundström. Slide navigation patterns among pathologists with long experience of digital review. *Histopathology*, 67(2):185–192, 2015.
- [82] J. Morita, K. Miwa, T. Kitasaka, K. Mori, Y. Suenaga, S. Iwano, M. Ikeda, and T. Ishigaki. Interactions of perceptual and conceptual processing: Expertise in medical image diagnosis. *International Journal of Human-Computer Studies*, 66(5):370–390, 2008. ISSN 1071-5819. doi: <https://doi.org/10.1016/j.ijhcs.2007.11.004>. URL <https://www.sciencedirect.com/science/article/pii/S107158190700167X>.
- [83] D. Morris, E. Müller-Budack, and R. Ewerth. Slideimages: A dataset for educational image classification, 2020. URL <https://arxiv.org/abs/2001.06823>.
- [84] E. Nagy, M. Janisch, F. Hržić, et al. A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. *Sci Data*, 9:222, 2022. doi: 10.1038/s41597-022-01328-z. URL <https://doi.org/10.1038/s41597-022-01328-z>.
- [85] S. A. Nasser, N. Gupte, and A. Sethi. Reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data, 2023. URL <https://arxiv.org/abs/2307.10698>.
- [86] National Library of Medicine. Pmc open access subset [internet], 2003. URL <https://pmc.ncbi.nlm.nih.gov/tools/openftlist/>.
- [87] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. Tong, D. H. Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data*, 9(1):429, 2022.
- [88] H. T. Nguyen, H. Q. Nguyen, H. H. Pham, K. Lam, L. T. Le, M. Dao, and V. Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.
- [89] D. Ouyang, B. He, A. Ghorbani, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 2020. doi: 10.1038/s41586-020-2145-8. URL <https://doi.org/10.1038/s41586-020-2145-8>.
- [90] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, L. Giancardo, G. Quellec, and F. Mériaudeau. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. *Data*, 6(2), 2021. ISSN 2306-5729. doi: 10.3390/data6020014. URL <https://www.mdpi.com/2306-5729/6/2/14>.
- [91] Y. Pan, S. Bano, F. Vasconcelos, H. Park, T. T. Jeong, and D. Stoyanov. Desmoke-lap: improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *Int J Comput Assist Radiol Surg*, 17(5):885–893, May 2022. doi: 10.1007/s11548-022-02595-2.

- [92] K. Panetta, R. Rajendran, A. Ramesh, S. P. Rao, and S. Agaian. Tufts dental database: a multimodal panoramic x-ray dataset for benchmarking diagnostic systems. *IEEE journal of biomedical and health informatics*, 26(4):1650–1659, 2021.
- [93] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero. An open access thyroid ultrasound image database. In *10th International Symposium on Medical Information Processing and Analysis*, volume 9287, pages 188–193. SPIE, 2014. doi: 10.1117/12.2073532. URL <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9287/92870W/An-open-access-thyroid-ultrasound-image-database/10.1117/12.2073532.full>.
- [94] PKNU-PR-ML-Lab. Calculus. <https://github.com/PKNU-PR-ML-Lab/calculus>.
- [95] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, 2017. URL <https://doi.org/10.1145/3193289>.
- [96] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari. Connecting vision and language with localized narratives, 2020. URL <https://arxiv.org/abs/1912.03098>.
- [97] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [98] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [99] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.
- [100] R. B. Rahman, S. A. Tanim, N. Alfaz, T. E. Shrestha, M. S. U. Miah, and F. Mridha. Dental OPG XRAY Dataset. 2024. doi: 10.17632/c4hhrkxytw.4.
- [101] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng. Mura: Large dataset for abnormality detection in musculoskeletal radiographs, 2018. URL <https://arxiv.org/abs/1712.06957>.
- [102] J. Román, V. Fretes, C. Adorno, R. Silva, J. Noguera, H. Legal-Ayala, J. Mello-Román, R. Torres, and J. Facon. Panoramic dental radiography image enhancement using multiscale mathematical morphology. *Sensors*, 21(9):3110, 2021. doi: 10.3390/s21093110. URL <https://doi.org/10.3390/s21093110>.
- [103] T. Ross, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. M. Filimon, P. Scholz, T. N. Tran, P. Bruno, P. Arbeláez, G.-B. Bian, S. Bodenstedt, J. L. Bolmgren, L. Bravo-Sánchez, H.-B. Chen, C. González, D. Guo, P. Halvorsen, P.-A. Heng, E. Hosgor, Z.-G. Hou, F. Isensee, D. Jha, T. Jiang, Y. Jin, K. Kirtac, S. Kletz, S. Leger, Z. Li, K. H. Maier-Hein, Z.-L. Ni, M. A. Riegler, K. Schoeffmann, R. Shi, S. Speidel, M. Stenzel, I. Twick, G. Wang, J. Wang, L. Wang, L. Wang, Y. Zhang, Y.-J. Zhou, L. Zhu, M. Wiesenfarth, A. Kopp-Schneider, B. P. Müller-Stich, and L. Maier-Hein. Robust medical instrument segmentation challenge 2019, 2020. URL <https://arxiv.org/abs/2003.10299>.
- [104] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, A. Halpern, B. Helba, H. Kittler, K. Kose, S. Langer, K. Lioprys, J. Malvey, S. Musthaq, J. Nanda, O. Reiter, G. Shih, A. Stratigos, P. Tschandl, J. Weber, and P. Soyer. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data*, 8:34, 2021. doi: 10.1038/s41597-021-00815-z. URL <https://doi.org/10.1038/s41597-021-00815-z>.
- [105] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.

- [106] A. Saha, M. R. Harowicz, L. J. Grimm, J. Weng, E. H. Cain, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations (version 3), 2023. URL <https://doi.org/10.7937/TCIA.E3SV-RE93>
- [107] S. Sajid. Oral diseases. <https://www.kaggle.com/datasets/salmansajid05/oral-diseases>
- [108] N. Sajjad. Dental cavity. <https://www.kaggle.com/datasets/nabeel1921/dental-cavity>
- [109] J. Saltz, R. Gupta, L. Hou, T. Kurc, P. Singh, V. Nguyen, D. Samaras, K. R. Shroyer, T. Zhao, R. Batiste, J. Van Arnam, T. C. G. A. R. Network, I. Shmulevich, A. U. K. Rao, A. J. Lazar, A. Sharma, and V. Thorsson. Tumor-infiltrating lymphocytes maps from tcga h&e whole slide pathology images, 2018. URL <https://doi.org/10.7937/K9/TCIA.2018.Y75F9W1>. Data set.
- [110] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin. Curated breast imaging subset of digital database for screening mammography (cbis-ddsm), 2016. URL <https://doi.org/10.7937/K9/TCIA.2016.7002S9CY>
- [111] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [112] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL <https://arxiv.org/abs/2210.08402>
- [113] M. S. Seyfioglu, W. O. Ikezogwo, F. Ghezloo, R. Krishna, and L. Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024.
- [114] D. Shi, W. Zhang, J. Yang, S. Huang, X. Chen, M. Yusufu, K. Jin, S. Lin, S. Liu, Q. Zhang, and M. He. Eyeclip: A visual-language foundation model for multi-modal ophthalmic image analysis, 2024. URL <https://arxiv.org/abs/2409.06644>.
- [115] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read, 2019. URL <https://arxiv.org/abs/1904.08920>.
- [116] J. Staal, M. Abramoff, M. Niemeijer, M. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4): 501–509, 2004. doi: 10.1109/TMI.2004.825627.
- [117] S. Subramanian, L. L. Wang, S. Mehta, B. Bogin, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, and H. Hajishirzi. Medcat: A dataset of medical images, captions, and textual references, 2020. URL <https://arxiv.org/abs/2010.06000>
- [118] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), Aug. 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL <http://dx.doi.org/10.1038/sdata.2018.161>
- [119] T. Tu, S. Azizi, D. Driess, M. Schaeckermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.
- [120] T. L. A. van den Heuvel et al. Automated measurement of fetal head circumference using 2d ultrasound images. <https://doi.org/10.5281/zenodo.1327317>, July 2018.

- [121] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.
- [122] P. Voigtlaender, S. Changpinyo, J. Pont-Tuset, R. Soricut, and V. Ferrari. Connecting Vision and Language with Video Localized Narratives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [123] P. Voigtlaender, S. Changpinyo, J. Pont-Tuset, R. Soricut, and V. Ferrari. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471, 2023.
- [124] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3462–3471. IEEE, July 2017. doi: 10.1109/cvpr.2017.369. URL <http://dx.doi.org/10.1109/CVPR.2017.369>.
- [125] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, N. Tomita, L. Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.
- [126] R. Wightman. Pytorch image models. <https://github.com/huggingface/pytorch-image-models>, 2019.
- [127] H. E. Wong, M. Rakic, J. Gutttag, and A. V. Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In *European Conference on Computer Vision*, pages 207–229. Springer, 2024.
- [128] Y. Xie, C. Zhou, L. Gao, J. Wu, X. Li, H.-Y. Zhou, S. Liu, L. Xing, J. Zou, C. Xie, and Y. Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine, 2024. URL <https://arxiv.org/abs/2408.02900>.
- [129] J. Xu, X. Zhou, S. Yan, X. Gu, A. Arnab, C. Sun, X. Wang, and C. Schmid. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039, 2024.
- [130] S. Yadav. Oral cancer lips and tongue images. <https://www.kaggle.com/datasets/shivam17299/oral-cancer-lips-and-tongue-images>.
- [131] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=AFDcYJKhND>. Featured Certification.
- [132] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui. fastmri: An open dataset and benchmarks for accelerated mri, 2019. URL <https://arxiv.org/abs/1811.08839>.
- [133] H. Zhang, Y. He, X. Wu, P. Huang, W. Qin, F. Wang, J. Ye, X. Huang, Y. Liao, H. Chen, et al. Pathnarratives: Data annotation for pathological human-ai collaborative diagnosis. *Frontiers in Medicine*, 9:1070072, 2023.
- [134] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

- [135] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.
- [136] Q. Zhao, S. Lyu, W. Bai, L. Cai, B. Liu, G. Cheng, M. Wu, X. Sang, M. Yang, and L. Chen. Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation, 2023. URL <https://arxiv.org/abs/2207.06799>.
- [137] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding, 2016. URL <https://arxiv.org/abs/1610.02055>.
- [138] H.-Y. Zhou, S. Adithan, J. N. Acosta, E. J. Topol, and P. Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.
- [139] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The dataset characterization, methods, and examples claimed in the abstract and introduction can be seen in sections [3](#), [1](#), and [E](#), as well as the dataset links provided at submission.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Please see section [5](#)

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: Not applicable since this is not a theory paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include all the details of the data creation, and the code.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes we do, we submit links for both

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: please see sections 4 and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report this, due to the added computation and other constraints.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: please see section 4 and Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Yes it does.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Please see the discussion section 6 and 5

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not applicable as data is fully open and our license does not allow for commercial use or duplication/derivatives protecting the source.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: This work creates an asset and cites/credits all assets used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: yes.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[No\]](#)

Justification: We do not crowdsource or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[No\]](#)

Justification: We do not perform research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We outline our use of LLMs to filter our data in section 3 and in the Appendix, we also list the exact prompts used and model type in the Appendix.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.