

Table 1: Evaluation datasets.

Task	Setup	Dataset	Modality	Nb. Samples
Retrieval	I→T & T→I	Quilt-1M	Histopathology	13,559
		MIMIC-IV-CXR	Chest X-ray	3,269
		DeepEyeNet	Retina	3,140
Zero-shot classification & Linear probing	6 classes	PAD-UFES-20	Dermatology	460
	7 classes	SkinCancer	Dermatology	2,003
	2 classes	PatchCamelyon (PCam)	Histopathology	32,768
	8 classes	NCT-CRC-HE-100K	Histopathology	6,333
	3 classes	LC25000Lung	Histopathology	3,000
	2 classes	LC25000Colon	Histopathology	2,000
	4 classes	BACH	Histopathology	100
	4 classes	SICAPv2	Histopathology	2,122
	9 classes	PathMNIST+	Colon Pathology	107,180
	7 classes	DermaMNIST+	Dermatoscope	10,015
	4 classes	OctMNIST+	Retinal OCT	109,309
	2 classes	PneumoniaMNIST+	Chest X-Ray	5,856
	5 classes	RetinaMNIST+	Fundus Camera	1,600
	2 classes	BreastMNIST+	Breast Ultrasound	780
	8 classes	BloodMNIST+	Blood Cell Microscope	17,092
	8 classes	TissueMNIST+	Kidney Cortex Microscope	236,386
	11 classes	OrganAMNIST+	Abdominal CT	58,830
	11 classes	OrganCMNIST+	Abdominal CT	23,583
	11 classes	OrganSMNIST+	Abdominal CT	25,211

## 1 Pretraining Hyperparameters

For pretraining, we use the AdamW optimizer with a weight decay of 0.2,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.98$ . The learning rate is scheduled using cosine decay, with a linear warmup over the first 10% of the total training steps. We apply gradient accumulation with a frequency of 4. Training is performed on 8 NVIDIA A100 GPUs with a total batch size of 2048. The initial learning rate is set to  $5.0 \times 10^{-4}$ , and models are trained for 64 epochs.

## 2 Evaluation Datasets

We evaluate our pretrained models on three downstream tasks: image-text retrieval, zero-shot classification, and linear probing. A summary of the evaluation datasets used for these tasks is provided in Table 1.

**MIMIC-CXR:** MIMIC-CXR contains 377,110 de-identified chest X-ray images from 65,379 patients, accompanied by free-text reports. The dataset was collected from the emergency department of Beth Israel Deaconess Medical Center. Each patient typically has multiple views and a corresponding radiology report labeled using the CheXpert labeling tool (Irvin et al., 2019), which identifies 13 common conditions such as atelectasis, cardiomegaly, consolidation, pleural effusion, and pneumonia.

**Quilt-1M:** Quilt-1M comprises over one million histopathology image-text pairs. The largest subset, Quilt, includes 802,144 pairs extracted from 1,087 hours of educational histopathology videos on YouTube. Captions were generated using a combination of large language models, handcrafted rules, and automatic speech recognition. Additional subsets come from PubMed Open Access, LAION-5B, and OpenPath Twitter data, resulting in a combined dataset of over one million pairs.

**DeepEyeNet:** DeepEyeNet is a large-scale retinal image dataset comprising 15,709 images, including both color fundus photography (CFP) and fluorescein angiography (FA). Each image is annotated with three expert-defined labels: a disease or symptom name, a set of relevant keywords, and a detailed clinical description. The dataset covers 265 distinct retinal conditions.

25 **SICAP:** The Prostate Cancer Grade Assessment (SICAP) dataset comprises prostate histology  
 26 whole-slide images annotated with global Gleason scores and path-level Gleason grades, supporting  
 27 research in automated prostate cancer grading.

28 **PAD-UFES-20:** PAD-UFES-20 includes 2,298 clinical images of six types of skin lesions, each  
 29 accompanied by up to 22 patient metadata features, facilitating studies in skin lesion classification.

30 **Skin Cancer:** This dataset contains 2,357 dermatoscopic images of skin lesions, labeled with  
 31 diagnostic categories, aiding in the development of skin cancer detection models.

32 **PCam (PatchCamelyon):** PCam consists of 327,680 color images (96×96 px) extracted from  
 33 histopathologic scans of lymph node sections, each labeled to indicate the presence of metastatic  
 34 tissue.

35 **NCT-CRC-HE:** The NCT-CRC-HE dataset comprises 100,000 non-overlapping image patches  
 36 from H&E-stained histological images of human colorectal cancer and normal tissue, supporting  
 37 research in histopathological image analysis.

38 **LC-Lung:** LC-Lung includes 15,000 histopathological images of lung tissue, categorized into  
 39 benign and malignant classes, useful for lung cancer classification studies.

40 **LC-Colon:** LC-Colon comprises 10,000 histopathological images of colon tissue, labeled as benign  
 41 or malignant, aiding in colon cancer detection research.

42 **BACH:** The Breast Cancer Histology (BACH) dataset contains microscopy images of breast tissue,  
 43 annotated across four classes: normal, benign, in situ carcinoma, and invasive carcinoma, facilitating  
 44 automated breast cancer diagnosis.

45 **DermaMNIST+:** DermaMNIST+ consists of 10,015 dermatoscopic images categorized into seven  
 46 skin disease classes, serving as a benchmark for skin lesion classification tasks.

47 **OCTMNIST+:** OCTMNIST+ includes 109,309 optical coherence tomography images labeled for  
 48 retinal diseases like choroidal neovascularization, diabetic macular edema, and drusen, supporting  
 49 ophthalmic image classification.

50 **PneumoniaMNIST+:** PneumoniaMNIST+ is based on 5,856 pediatric chest X-ray images, labeled  
 51 for pneumonia detection, aiding in the development of automated pneumonia diagnosis models.

52 **RetinaMNIST+:** RetinaMNIST+ comprises 1,600 retinal fundus images labeled for common eye  
 53 diseases, useful for training models in automated retinal disease classification.

54 **BreastMNIST+:** BreastMNIST+ contains 780 ultrasound images of breast tumors, labeled as  
 55 benign or malignant, supporting breast cancer detection research.

56 **BloodMNIST+:** BloodMNIST+ consists of 17,092 microscopic images of blood cells, classified  
 57 into eight cell types, facilitating automated classification tasks in hematology.

58 **TissueMNIST+:** TissueMNIST+ includes 236,386 microscopic images of tissue samples from  
 59 different organs, labeled according to tissue type, supporting histopathological analysis.

60 **PathMNIST+:** PathMNIST+ is derived from colorectal cancer tissue slides, containing 107,180  
 61 images labeled with nine different tissue classes, aiding in multi-class classification tasks in pathology.

62 **OrganAMNIST+:** OrganAMNIST+ consists of 58,850 abdominal CT images labeled with different  
 63 anatomical organ classes, supporting organ segmentation and classification tasks.

64 **OrganCMNIST+:** OrganCMNIST+ contains 23,600 coronal CT images of various organs, labeled  
 65 for organ classification tasks, used for research in medical image understanding.

66 **OrganSMNIST+:** OrganSMNIST+ comprises 23,600 sagittal CT images of multiple organs,  
 67 annotated for classification, aiding in comprehensive medical imaging analysis.

### 68 3 Evaluation Results

#### 69 3.1 Retrieval

70 In addition to the results of recall at 200 which was included in the main body of the paper, we also  
 71 provide the results for recall at 10 and recall at 50.

Table 2: Retrieval performance (Recall@10) of all models trained on paired image-caption pairs in the medical domain. The last column, Average Recall (AR), aggregates the results across all tasks.

Model	Image-to-Text			Text-to-Image			AR
	MIMIC	Quilt	DeepEyeNet	MIMIC	Quilt	DeepEyeNet	
PMC-OA	0.014	0.020	0.026	0.010	0.016	0.017	0.017
OPEN-PMC	0.022	0.018	0.024	0.016	0.016	0.024	0.020
BioMedCLIP	0.022	0.024	0.031	0.015	0.027	0.024	0.023
BIOMEDICA	0.005	0.033	0.023	0.006	0.041	0.022	0.021
PMC-6M	0.033	0.028	0.039	0.028	0.032	0.035	0.032
OPEN-PMC-18M	0.026	0.031	0.033	0.023	0.038	0.040	0.031

Table 3: Retrieval performance (Recall@50) of all models trained on paired image-caption pairs in the medical domain. The last column, Average Recall (AR), aggregates the results across all tasks.

Model	Image-to-Text			Text-to-Image			AR
	MIMIC	Quilt	DeepEyeNet	MIMIC	Quilt	DeepEyeNet	
PMC-OA	0.054	0.062	0.071	0.044	0.056	0.070	0.059
OPEN-PMC	0.072	0.059	0.058	0.056	0.053	0.077	0.062
BioMedCLIP	0.067	0.070	0.074	0.055	0.082	0.074	0.070
BIOMEDICA	0.024	0.084	0.071	0.030	0.102	0.067	0.63
PMC-6M	0.106	0.087	0.092	0.097	0.098	0.088	0.094
OPEN-PMC-18M	0.081	0.093	0.078	0.083	0.109	0.097	0.090

#### 72 3.2 Linear Probing

73 To evaluate the quality of the learned image representations, we perform linear probing using a  
 74 single-layer MLP on the downstream task datasets. Each model is trained for 40 epochs with a cosine  
 75 annealing learning rate schedule, starting from an initial learning rate of 0.1. The results are presented  
 76 in Table 4.

Table 4: Linear-probing F1-scores across diverse medical datasets for different models.

Radiology									
Model	PneumoniaMNIST+	BreastMNIST+	OrganAMNIST+	OrganCMNIST+	OrganSMNIST+	Average			
BioMedCLIP	92.96	75.63	85.71	79.29	64.88	79.69			
BIOMEDICA	86.15	77.16	89.72	82.66	70.93	81.12			
PMC-6M	79.74	77.84	89.56	85.00	69.07	80.24			
OPEN-PMC-18M	79.74	77.84	89.51	85.00	69.07	80.23			
Visible Light Photography									
Model	PAD-UFES-20	Skin Cancer	PathMNIST+	DermaMNIST+	OCTMNIST+	RetinaMNIST+	Average		
BioMedCLIP	62.31	56.43	90.27	59.62	71.70	42.95	63.88		
BIOMEDICA	82.59	68.09	88.32	74.02	80.17	52.11	74.21		
PMC-6M	75.62	61.61	91.28	61.47	80.17	46.10	69.37		
OPEN-PMC-18M	75.62	62.92	91.35	61.47	78.73	46.59	69.44		
Microscopy									
Model	Sicap	PCam	NCT-CRC-HE	LC-Lung	LC-Colon	BACH	BloodMNIST+	TissueMNIST+	Average
BioMedCLIP	63.84	83.00	72.56	96.83	99.75	73.01	95.43	43.71	78.51
BIOMEDICA	65.15	86.41	83.57	99.26	99.95	75.65	96.92	50.69	82.2
PMC-6M	60.00	84.22	64.64	98.85	99.80	62.39	95.87	49.89	76.95
OPEN-PMC-18M	59.85	84.16	64.64	98.85	99.80	65.52	95.87	49.96	77.33