Improving Medical Multi-modal Contrastive Learning with Expert Annotations

Yogesh Kumar and Pekka Marttinen

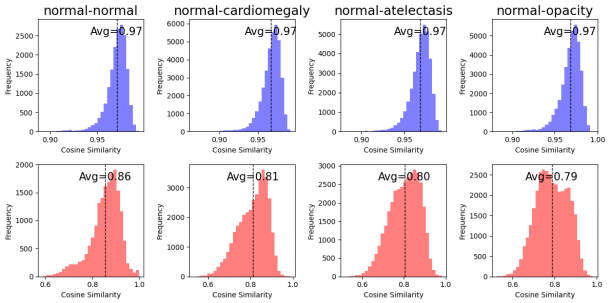
Department of Computer Science, Aalto University, Finland
{firstname.lastname}@aalto.fi

Abstract. We introduce eCLIP, an enhanced version of the CLIP model that integrates expert annotations in the form of radiologist eye-gaze heatmaps. It tackles key challenges in contrastive multi-modal medical imaging analysis, notably data scarcity and the "modality gap" – a significant disparity between image and text embeddings that diminishes the quality of representations and hampers cross-modal interoperability. eCLIP integrates a heatmap processor and leverages mixup augmentation to efficiently utilize the scarce expert annotations, thus boosting the model's learning effectiveness. eCLIP is designed to be generally applicable to any variant of CLIP without requiring any modifications of the core architecture. Through detailed evaluations across several tasks, including zero-shot inference, linear probing, cross-modal retrieval, and Retrieval Augmented Generation (RAG) of radiology reports using a frozen Large Language Model, eCLIP showcases consistent improvements in embedding quality. The outcomes reveal enhanced alignment and uniformity, affirming eCLIP's capability to harness high-quality annotations for enriched multi-modal analysis in the medical imaging domain.

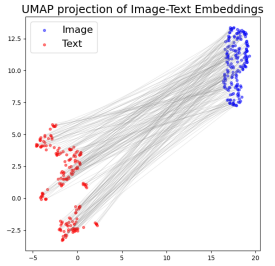
Keywords: Contrastive Learning · Medical Imaging · Zero-shot Inference

1 Introduction

Pretraining foundation models on multi-modal data – particularly leveraging the relationships between text and images – has proven to be a robust strategy for generating versatile embeddings [18, 36]. These embeddings enhance the efficacy in several downstream tasks, from image generation to advanced vision-language integration [26, 37, 40]. Central to this approach is the employment of a contrastive learning (CL) loss objective [4, 34, 61], where models are trained to align positive pairs (e.g., an image and its corresponding caption) while diversifying negative ones. A significant hurdle in this approach is the necessity of vast datasets, often comprising several millions of data points, for competitive results. Models such as CLIP [36] have been trained on internet-scale datasets, estimated to encompass hundreds of millions of image-text pairs [5, 57]. Acquiring datasets of this magnitude poses substantial challenges in specialized fields that require expert knowledge for data collection, processing and annotation.



(a) Histogram of Cosine Similarities Among Subgroups. The model exhibits high cosine similarity among embeddings within the text (*red*) and image (*blue*) modalities, regardless of the differences among subgroups. This underscores the model's challenge in capturing the subtleties inherent in medical data.



(b) Modality Gap. Despite the CLIP contrastive loss aiming at closely aligning image and text embeddings within a shared space, the modalities remain segregated into distinct regions.

Fig. 1: Analysis of CLIP Embeddings in Medical Imaging The figure presents embeddings generated by a CLIP model, pretrained on an internet-scale dataset, applied to the Open-I dataset pairing X-rays with corresponding radiology reports.

The medical imaging domain exemplifies these difficulties, where acquiring even a single data point, such as a chest X-ray, involves complex processes requiring expertise and significant resources. Moreover, the procurement of such data for machine learning research is further complicated by ethical considerations, patient privacy concerns and the need for extensive de-identification procedures.

This has led to the prevalent use of foundation models, initialized with weights from models trained on extensive internet-scale datasets, for tasks in the medical domain [15, 23, 54, 65]. However, the areas of interest within medical images are often nuanced and require expert knowledge to interpret, rendering them indistinguishable to a general-purpose model. In Fig. 1a, we investigate the embeddings generated by a CLIP model – initially pretrained on internet data – using samples from the Open-I dataset [6], which includes X-rays and corresponding radiology reports. We categorize the samples into subgroups based on the primary abnormality identified in each report, such as 'normal', cardiomegaly, atelectasis and opacity. A histogram of the cosine similarities between embeddings from different groups indicates a high degree of similarity, with values approaching 1. This could lead to potential challenges in downstream zero-shot inference tasks, which rely on the spatial segregation of embeddings from different groups [36]. Typically, continual pretraining on medically relevant data is employed to enhance the model's ability to differentiate between various abnormalities.

Recent studies [11, 25, 33, 46, 67] have identified a "modality gap" in multimodal contrastive representation learning, where the embeddings from different modalities (e.g., images and text) fall in distinct regions in the shared embedding space. This separation, which arises from factors such as initial model weights and the objectives of contrastive learning [25, 67], leads to the "cone effect" where

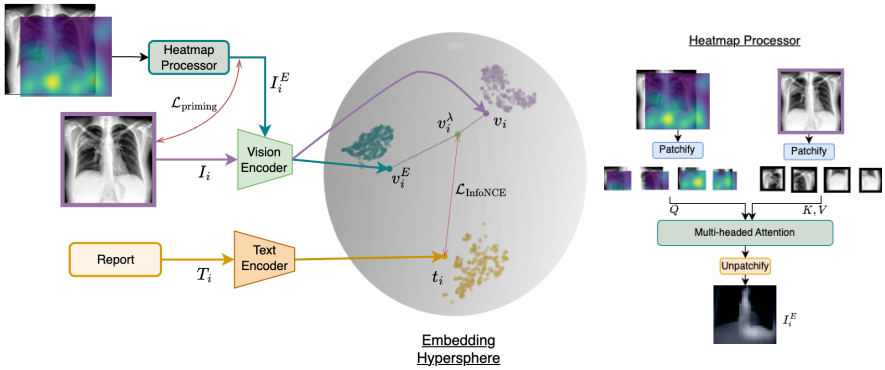


Fig. 2: eCLIP Pretraining with Expert Annotations. eCLIP adds a Heatmap Processor *(right)*, featuring a multi-headed attention layer, to the standard Image and Text encoders in CLIP. This processor, along with vision and text encoders, maps inputs into a shared hypersphere. Here, the original image (I_i), its text (T_i) and the heatmap-processed image (I_i^E) are positioned within a tripartite area (shown here after 2D UMAP projection, please refer to the Supplement for a scaled version). We employ mixup between I_i and I_i^E to generate the embedding v_i^λ, which gives us additional positive pairs to enhance the CLIP InfoNCE loss optimization. An auxillary loss, $\mathcal{L}_{\text{priming}}$, is used during the initial training steps to "prime" the heatmap processor to imitate an identity function when the heatmap is composed of all ones.

embeddings of each modality are restricted to a narrow region of the embedding hypersphere. In Fig. 1b, we illustrate this within the medical domain with a 2D UMAP [27] projection of image and text embeddings. This example highlights how embeddings from the same modality but different semantic groups, such as X-ray images of varying abnormalities, cluster closely together. This makes it difficult for a model to distinguish between semantically different images, undermining its performance in medical image analysis.

We investigate the potential of integrating expert annotations, specifically radiologist eye-gaze heatmaps, to alleviate these issues. Processing the eye-gaze data from radiologists [22] provides us with heatmaps indicative of the radiologist's attention across different regions of the X-ray images. This heatmap reflects areas of clinical interest aligned with details present in radiology reports. We posit that this could help capture nuanced visual cues in the X-rays and therefore pairing it with reports can enrich the CLIP training data with high-quality positive pairs. Due to the scarcity of such expert annotated data, we employ the mixup strategy, a data augmentation technique which has been effective in both supervised [13,48,62] and contrastive learning [33,49], to create additional synthetic samples.

We present eCLIP (expert-annotated CLIP), an adaptation of the CLIP model that incorporates expert eye-gaze heatmaps, without modifying the CLIP model's core architecture. The operational workflow of eCLIP is depicted in Fig. 2. Our **contributions** are as follows:

- **Utilization of Expert Annotations.** We harness radiologist eye-gaze heatmaps to create additional embeddings, effectively introducing valuable positive and negative pairs for enhancing the contrastive learning process.
- **eCLIP Architecture.** Our implementation features a heatmap processing mechanism utilizing multi-headed attention (MHA), optimized for handling both heatmaps and original images. This is complemented by a mixup strategy to addresses the challenge of data scarcity, and curriculum learning to ensure a gradual introduction of expert annotations.
- **Comprehensive Evaluation.** We assess eCLIP's zero-shot classification accuracy, sample efficiency and cross-modal retrieval performance and embedding quality across multiple chest X-ray datasets. We also evaluate the cross-modal embeddings to generate radiology reports using a frozen Large Language Model (LLM) without explicitly fine-tuning on medical data.

2 Related Work

Modality Gap: Liang *et al.* [25] pinpoint the origins of the modality gap to the nuances of model initialization and the objectives of contrastive learning, underscoring its impact on downstream tasks and fairness. Oh *et al.* [33] highlight poor uniformity and alignment in CLIP's embeddings and propose a finetuning method for robust representations. Zhang *et al.* [67] explore the geometry of this embedding space, and provide both theoretical and empirical insights on the nature of this geometry. Subsequent research has produced methods to mitigate the modality gap through diverse and creative approaches [11, 12, 32, 46, 66].

Improving Contrastive Learning: Several methods have improved upon the CLIP objective by introducing auxilliary losses, e.g., SLIP [30] uses SimCLR [4] loss, M3MAE [10,55] augment the Masked Autoencoder [14] reconstruction loss, FLIP [24] randomly masks out input images to improve scaling, DACL [49] proposes a domain agnostic mixup strategy, SILC [31] uses self-distillation, Mo *et al.* [28] utilized specialist captions to generate pseudo labels for unpaired images, Zhang *et al.* [63] propose Multi-task Paired Masking with Alignment to improve cross-modal interaction. Similarly there have been works that have identified the need to make the CLIP model focus on sub-regions in order to enhance its utility and downstream performance, GLoRIA [15] considers the loss from local regions from within the image and reports, Alpha-CLIP [45] uses the alpha channel to guide the CLIP model to focus on different regions of the image and generate the masks for all the images in the corpus using an image segmentation pipeline and TIER [35] uses a regularization term to improve the local focus of the model.

Multi-modal Contrastive Learning in Medical Imaging: Zhang *et al.* [65] demonstrated enhanced downstream performance by jointly using chest X-ray and report pairing for training a contrastive learning model. This was further improved by Huang *et al.* [15], by exploiting local and global features from both modalities; Wang *et al.* [54] and You *et al.* [59] achieved impressive results by using a Swin Tiny model as the image encoder and by adding modifications

to the contrastive loss. Several other works have developed similar contrastive learning foundation models while utilizing the biomedical image texts to achieve impressive results [16, 29, 44, 47, 56, 58]. Karargyris *et al.* [22] and Bigolin *et al.* [3] augment a subset of the MIMIC-CXR [21] samples with high quality eye-tracking and verbal transcripts from several radiologists. van Sonsbeek *et al.* [43] and Wang *et al.* [50] utilize the heatmaps from eye-gaze to improve image classification.

3 Method

Notations. For a given chest X-ray image I_i and its radiology report T_i, indexed by i in our dataset, we denote their L2-normalized embeddings as v_i and t_i respectively, residing in a d-dimensional space ($v_i, t_i \in \mathcal{R}^d$). Image embeddings are obtained through an encoder $v_i = f(I_i)$ and text embeddings via $t_i = g(T_i)$. Applying an expert heatmap E_i to an image results in the corresponding image embedding v_i^E. We denote the loss value for the i-th sample as $\mathcal{L}_i$. For the case of contrastive loss, this is computed in terms of some similarity measure between the embeddings, $sim(v_i, t_i)$, typically cosine similarity defined as $v_i \cdot t_i$.

3.1 Background

Central to CLIP's effectiveness is the InfoNCE loss [34], a mechanism engineered to optimize the similarity measures between corresponding (positive) pairs and to minimize those among non-corresponding (negative) pairs. The formulation of the CLIP loss objective is as follows:

$$\mathcal{L}_{\text{text}} = \mathop{\mathbb{E}}_{(t_i, v_i) \sim pos} \left[-\log \frac{\exp(sim(t_i, v_i)/\tau)}{\exp(sim(t_i, v_i)/\tau) + \sum_{j \neq i} \exp(sim(t_i, v_j)/\tau)} \right] \tag{1}$$

Total loss is then defined as, $\mathcal{L}_{\text{total}} = \frac{1}{2}(\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{image}})$, where $\mathcal{L}_{\text{image}}$ denotes the corresponding loss for the image to text mapping. Here, τ represents the temperature parameter that controls the scale of the similarity scores, typically framed as a learnable parameter during training. The loss expectation is taken over all the positive pairings in the dataset.

Theoretical results on CL indicate the concepts of *alignment* and *uniformity* as critical for the quality of embeddings [51, 52]. Alignment focuses on reducing the distance between positive pairs while uniformity seeks to evenly distribute the embeddings across the unit hypersphere, preventing extreme clustering that could impair the model's generalizability and discriminative capabilities. The alignment and uniformity can be defined formally as follows [33]:

$$\text{Alignment} = -\mathbb{E}_{(v_i, t_i) \sim pos} \left[\|v_i - t_i\|_2^2 - \min_{j \neq i} \|v_i - t_j\|_2^2 \right] \tag{2}$$

$$\text{Uniformity} = -\log \mathbb{E}_{(v_i, t_j) \sim \mathcal{D}} \left[\exp(-2\|v_i - t_j\|_2^2) \right] \tag{3}$$

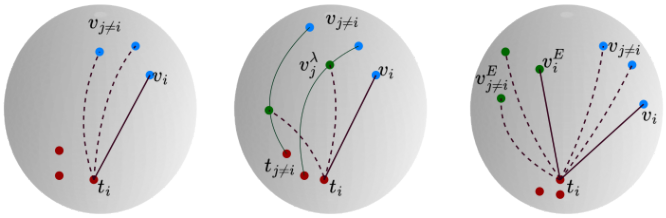


Fig. 3: Comparing eCLIP with m^2-mixup [33]. *(left)* Standard CLIP showing image-text positive pairs (v_i, t_i) (solid line), while the other image embeddings serve as negative pairs (dashed line). *(center)* the m^2-mixup creates negative pairs (v_j^λ, t_i) via interpolation between embeddings along the geodesic. *(right)* eCLIP adds expert image embedding, v_i^E, in addition to v_i for text t_i, forming additional positive and negative pairs

High intra-modal similarity, e.g. between two images as seen in Fig. 1, can inadvertently enhance similarity among negative pairs, inflating the denominator of the loss function in Equation (1), and, consequently, hurting the model's ability to differentiate between positive and negative pairs during training. A conventional method to counter this involves incorporating hard negative pairs, a strategy Oh *et al.* [33] employ by mixing embeddings from different modalities. While effective, this cross-modal mixup may obscure the semantic clarity of embeddings. As an alternative we propose increasing the dataset with additional positive pairs that exhibit minimal semantic overlap by integrating expert annotations. Fig. 3 compares the positive and negative pair creation in eCLIP with traditional CLIP and m^2-mixup.

3.2 Introducing Expert Annotations to CLIP

Our objective with eCLIP is to enhance the CLIP framework by integrating expert annotations – radiologist eye-gaze heatmaps – to diversify the pool of positive samples. The eCLIP model is designed to be compatible across all CLIP variants without modifying its core architecture. The **heatmap processor** (Fig. 2, *right*) first converts the images and heatmaps into a sequence of patches and applies multi-headed attention (MHA) over the sequences. The patchified heatmap overlaid images serve as queries, while the original image's patches act as keys and values. The processed output is then reconstructed back to its original image format, enabling the standard CLIP image encoder to obtain expert image embeddings. These new embeddings and their text embedding pair introduce additional positive samples for the contrastive loss objective (Fig. 3, *right*).

However, the size of the expert annotated data is orders of magnitude smaller than the data available for CLIP training. To effectively leverage the scarce expert-annotated data, we implement **mixup** augmentation [62]. As illustrated in Fig. 2 *(left)*, this involves blending an original image I_i with its expert version

Algorithm 1 eCLIP Algorithm

Require: Image Encoder $f(.)$
Require: Text Encoder $g(.)$
$\mathcal{L}_{\text{priming}} \leftarrow 0$; $n_{\text{p}} \leftarrow 0$
for minibatch $\{x_i\}_{i=1}^{N}$ **do**
  Unpack x_i to (T_i, I_i) and optionally E_i
  $t_i \leftarrow g(T_i)$
  $v_i \leftarrow f(I_i)$
  $p_{\text{uni}} \sim \text{Uniform}(0, 1)$
  if $p_{\text{uni}} < p_{\text{curr}}$ **and** E_i is provided **then**
    // Process expert image
    $\lambda \sim \text{Beta}(\alpha, \alpha)$
    $I_i^E \leftarrow \text{HeatmapProcessor}(I_i, E_i)$
    $I_i^\lambda \leftarrow I_i\lambda + I_i^E(1-\lambda)$
    $v_i^\lambda \leftarrow f(I_i^\lambda)$
  end if
  if E_i is entirely ones **then**
    $I_i^{\text{R}} \leftarrow \text{HeatmapProcessor}(I_i, E_i)$
    $\mathcal{L}_{\text{priming}} \leftarrow \mathcal{L}_{\text{priming}} + (I_i - I_i^{\text{R}})^2$
    $n_{\text{p}} \leftarrow n_{\text{p}} + 1$
  end if
end for
// Compute Total Loss
Require: Temperature τ
$V \leftarrow$ List of v_i for all i
$T \leftarrow$ List of t_i for all i
for $i = 1$ **to** N **do**
  if v_i^λ exists **then**
    $V \leftarrow \text{Append}(V, v_i^\lambda)$
    $T \leftarrow \text{Append}(T, t_i)$
  end if
end for
$\mathcal{L}_{\text{clip}} \leftarrow \text{ClipLoss}(V, T, \tau)$
if $n_{\text{p}} > 0$ **then**
  $\mathcal{L}_{\text{priming}} \leftarrow \frac{1}{n_{\text{p}}}\mathcal{L}_{\text{priming}}$
end if
$\mathcal{L}_{\text{total}} \leftarrow (1-w_p)\cdot\mathcal{L}_{\text{clip}} + w_p\cdot\mathcal{L}_{\text{priming}}$
Hyperparameters:
- Batch size N
- Mixup Alpha α
- Curriculum Prob. p_{curr}
- MSE Loss weight w_{p}

I_i^E to create $I_i^\lambda = \lambda I_i + (1-\lambda)I_i^E$, where $\lambda \sim \text{Beta}(\alpha, \alpha)$. (We set $\alpha = 0.3$ in all our experiments.) The eCLIP image encoder then processes I_i^λ to produce the image embedding $v_i^\lambda = f(I_i^\lambda)$. These expert embeddings form new positive pairs (v_i^λ, t_i) as well as corresponding negative pairs, which are added with existing pairs (v_i, t_i) during the computation of the CLIP InfoNCE Loss, $\mathcal{L}_i$.

To seamlessly integrate expert annotations without disrupting the foundational training of the eCLIP model, we employ a phased **curriculum learning** strategy [2]. This approach comprises a cold start phase where the model is initially trained without the expert annotations to establish a robust baseline. This phase accounts for about 10% of the total training iterations. It then transitions into a warmup phase, gradually increasing the inclusion of expert examples from 0.05 to 0.5 probability over the next 30% of iterations. Finally, a cooldown phase reduces expert example probability to 0.1 for the subsequent 40% of iterations, fine-tuning the model's performance by balancing foundational and expert-driven insights.

Additionally, we regularize the heatmap processor to behave as an identity function in scenarios where the heatmap is entirely composed of ones. We achieve this through a **priming** phase that coincides with the curriculum learning's cold start phase. We setup an auxillary mean-squared error loss to force the heatmap processor to reconstruct the original image I_i when the heatmap $E_i = 1$. This priming ensures the heatmap processor's adaptability, allowing it to process expert annotations effectively when available, while falling back to the

model's original performance in their absence. The total loss during this phase is, $\mathcal{L}_{\text{total}} = w_{\text{p}} \cdot \mathcal{L}_{\text{priming}} + (1 - w_{\text{p}}) \cdot \mathcal{L}_{\text{clip}}$, where w_{p} is a hyperparameter which we set to 0.1. The pseudocode for eCLIP is shown in Algorithm 1.

4 Experiments

Our experiments are designed to evaluate the influence of expert heatmap annotations on the quality of the learned representations. Unless stated otherwise, we assume that a large set of image-text pairs, of which a small fraction is annotated with eye-gaze heatmaps, is used for training, but for test samples no annotations are used. We utilize both quantitative measures and qualitative assessments to study the contributions of these annotations towards enhancing model performance. The source code is available online.

4.1 Setup

Baselines To validate our approach, we compare eCLIP against a model trained using traditional CLIP (referred to as the base model) and a "naive" baseline, where the expert annotated samples are directly added to the training set without using mixup or curriculum learning. We also examine the impact of two mixup methods: Domain Agnostic Contrastive Learning (DACL) [49] and m^3-mixup [33] which blends image and text embeddings to improve alignment and uniformity across modalities. While DACL is integrated during pretraining, m^3-mixup is applied post-pretraining, in a manner akin to fine-tuning. eCLIP can be applied to any variant of CLIP, which we demonstrate also with GLoRIA [15] which has a Resnet50 image encoder. We introduce **two variants** of our technique: eCLIP, which integrates expert annotations during the initial CLIP pretraining phase, and eCLIP$^{\mathcal{P}}$, which instead continually finetunes a trained CLIP model with expert annotations, similar to m^3-mixup.

Datasets For pretraining phase we utilize the **MIMIC-CXR** dataset [21], which pairs roughly 200K chest X-rays with free-text radiology reports. The images were processed into JPEG format as described in [20], and the accompanying reports were stripped of unnecessary punctuation and tokenized using the Wordpiece scheme [7]. We obtain the eye-gaze heatmap from the EGD-CXR dataset [6] and process the eye-tracking data to obtain the normalized eye-gaze heatmap which are available for 1080 datapoints.

Our evaluation setup includes multiple publicly available chest X-ray datasets, specifically **CheXpert** [17], **RSNA Pneumonia** [39], **NIH CXR** [53] and **Open-I** [6], each offering a distinct set of imaging and reporting characteristics. Following previous works [15, 54, 59], we prepare the test sets from MIMIC and CheXpert, selecting 200 random samples for five specific pathologies from the CheXpert competition, resulting in 1000 samples for each dataset (MIMIC 5x200 and CheXpert 5x200, respectively). For the NIH-CXR dataset, we assembled a subset of 100 samples for each of 14 abnormalities, thereby creating CXR 14x100

test set. The Open-I dataset is utilized for text retrieval and radiology report generation tasks. For linear probe evaluations, we use CXR-8 [53], RSNA dataset and construct an OpenI-5 dataset by extracting labels from the 'Problems' field within the reports that match the CheXpert competition labels.

Training We employed CLIP pretraining on the MIMIC-CXR dataset, utilizing a subset with roughly 1000 images with expert eye-gaze heatmap annotations, while validation and all other downstream evaluations proceed without these annotations. Our model architecture includes Swin Tiny, following recent studies [54,59], alongside Vision Transformer (ViT) Small and Base, with image encoders pretrained on ImageNet and Clinical BERT [1] with max length of sequences set to 256 as the text encoder. We cropped images to $(224, 224)$ using random resized crop augmentation and turned off all other image augmentations. Pretraining utilized 8 AMD MI250X GPUs, maintaining an effective batch size of 512 for 10,000 steps. The learning rate was $1e^{-4}$ for standard CLIP, increased to $2e^{-4}$ for eCLIP variants, with cosine annealing plus a linear warmup for the first 10% of iterations, weight decay of $1e^{-3}$, and learnable temperature parameter in the contrastive loss initialized at 0.07. Models for m^3-mixup and eCLIP$^{\mathcal{P}}$ are initialized with weights from CLIP pretraining and further finetuned for 1,000 iterations with a learning rate of $1e^{-5}$. Detailed setup information is available in the Supplement.

4.2 Zero-shot Image Classification

Following CLIP [36], our zero-shot classification method categorizes images into predefined classes without direct finetuning, thus relying on the quality of embeddings generated during pretraining. To formulate the embedding of each class label, we first generate descriptive prompts to obtain a list of text embeddings corresponding to the label using the text encoder [15, 54, 59]. The mean of these embeddings is taken as the representation of the label. For each image, classification is then performed by matching the image embedding with its closest label embedding through cosine similarity. More details of the prompts used are provided in the Supplement.

Tab. 1 illustrates the zero-shot classification performance on CheXpert 5x200, MIMIC 5x200, RSNA, and CXR 14x100 datasets. The results, based on macro-averaged F1 scores from three random initializations, highlight eCLIP variants' superior performance over the base models across all datasets. While the m^3-mixup excels in MIMIC for certain architectures, eCLIP variants show broader generalization. Notably, eCLIP's advantages are more pronounced in multi-class scenarios (CheXpert, MIMIC and CXR14) compared to binary classification on RSNA.

4.3 Sample Efficiency

Sample efficiency measures how well a model learns from limited amount of training data. eCLIP improves this efficiency by using expert annotated im-

Table 1: Zero-shot classification performance on 4 X-ray datasets and model configurations, reported as macro-averaged F1 scores from three independent random seeds. The highest score per dataset and model configuration is underlined. The overall best-performing model for each dataset is highlighted in bold.

Model	**Dataset**			
	Chexpert 5x200	MIMIC 5x200	RSNA	CXR 14x100
GLoRIA$_{\text{Resnet50}}$	$0.478_{\pm.023}$	$0.457_{\pm.016}$	$0.736_{\pm.024}$	$0.155_{\pm.001}$
+naive	$0.391_{\pm.069}$	$0.334_{\pm.057}$	$0.731_{\pm.023}$	$0.113_{\pm.024}$
+DACL	$0.506_{\pm.029}$	$0.430_{\pm.011}$	$0.736_{\pm.007}$	$0.158_{\pm.004}$
+m^3-mix	$0.512_{\pm.005}$	$0.467_{\pm.004}$	$0.760_{\pm.006}$	$0.160_{\pm.003}$
+expert (ours)	$0.507_{\pm.004}$	$0.430_{\pm.009}$	$0.761_{\pm.017}$	$0.156_{\pm.023}$
+expert$^{\mathcal{P}}$(ours)	$0.507_{\pm.005}$	$0.475_{\pm.008}$	$0.775_{\pm.001}$	$0.159_{\pm.001}$
CLIP$_{\text{Swin Tiny}}$	$0.517_{\pm.025}$	$0.452_{\pm.002}$	$0.808_{\pm.000}$	$0.169_{\pm.003}$
+naive	$0.532_{\pm.010}$	$0.452_{\pm.022}$	$0.807_{\pm.007}$	$0.167_{\pm.007}$
+DACL	$0.465_{\pm.008}$	$0.389_{\pm.015}$	$0.768_{\pm.018}$	$0.101_{\pm.013}$
+m^3-mix	$0.554_{\pm.006}$	$0.469_{\pm.008}$	$0.802_{\pm.004}$	$0.179_{\pm.008}$
+expert (ours)	$0.549_{\pm.016}$	$0.445_{\pm.021}$	$0.818_{\pm.004}$	$0.172_{\pm.006}$
+expert$^{\mathcal{P}}$(ours)	$0.558_{\pm.004}$	$0.463_{\pm.007}$	$\mathbf{0.819_{\pm.000}}$	$0.192_{\pm.003}$
CLIP$_{\text{ViT Small}}$	$0.525_{\pm.024}$	$0.441_{\pm.006}$	$0.807_{\pm.006}$	$0.159_{\pm.007}$
+naive	$0.534_{\pm.016}$	$0.440_{\pm.019}$	$0.805_{\pm.004}$	$0.156_{\pm.017}$
+DACL	$0.475_{\pm.025}$	$0.398_{\pm015}$	$0.761_{\pm.009}$	$0.133_{\pm.007}$
+m^3-mix	$0.557_{\pm.002}$	$0.454_{\pm.003}$	$0.809_{\pm.002}$	$0.164_{\pm.002}$
+expert (ours)	$0.545_{\pm.016}$	$0.452_{\pm.013}$	$0.803_{\pm.003}$	$0.165_{\pm.007}$
+expert$^{\mathcal{P}}$(ours)	$0.559_{\pm.001}$	$0.439_{\pm.004}$	$0.817_{\pm.001}$	$0.165_{\pm.004}$
CLIP$_{\text{ViT Base}}$	$0.540_{\pm.017}$	$0.465_{\pm.004}$	$0.805_{\pm.001}$	$0.183_{\pm.011}$
+naive	$0.506_{\pm.011}$	$0.426_{\pm.006}$	$0.805_{\pm.004}$	$0.151_{\pm.009}$
+DACL	$0.474_{\pm.007}$	$0.400_{\pm.002}$	$0.759_{\pm.001}$	$0.106_{\pm.003}$
+m^3-mix	$0.542_{\pm.021}$	$0.465_{\pm.013}$	$0.798_{\pm.004}$	$0.183_{\pm.020}$
+expert (ours)	$\mathbf{0.563_{\pm.021}}$	$\mathbf{0.477_{\pm.004}}$	$0.814_{\pm.003}$	$\mathbf{0.193_{\pm.017}}$
+expert$^{\mathcal{P}}$(ours)	$0.549_{\pm.003}$	$0.452_{\pm.007}$	$0.810_{\pm.001}$	$0.185_{\pm.002}$

ages to form new positive and negative pairs, aiming to improve the quality of the learned embeddings. To test this, we first looked at zero-shot classification performance, adjusting the number of training batches available for pre-training. Results, shown in Fig. 4 *(top row)*, reveal that eCLIP is more sample efficient across MIMIC 5x200, CheXpert 5x200 and CXR 14x100 datasets compared to the base model. Additionally, by applying supervised fine-tuning (SFT) with a linear probe on class-imbalanced datasets – CXR-8, RSNA and OpenI-5 (Sec. 4.1) – eCLIP demonstrates stronger performance in multi-label classification tasks, CXR-8 and OpenI-5 and remains competitive in binary classification for RSNA. This is shown Fig. 4 *(bottom row)* where we plot the ROC AUC scores against different training sample sizes for linear probing. These findings highlight eCLIP's ability to effectively learn from fewer samples.

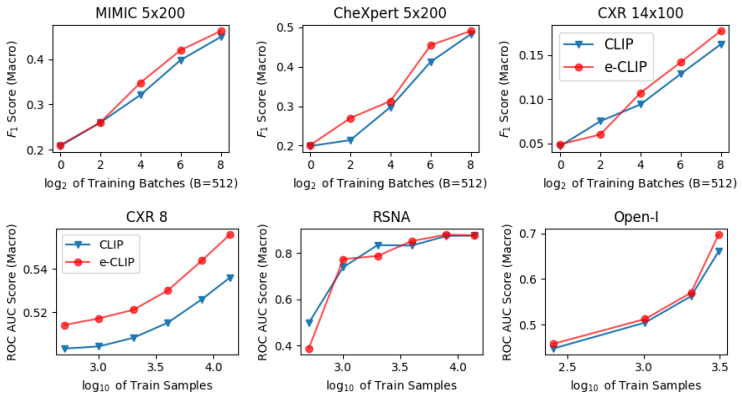


Fig. 4: Sample Efficiency. *(top row)* Zero-shot performance on three multi-label classification test sets for CLIP and eCLIP Swin Tiny models, trained with varying amounts of training batches. *(bottom row)* Linear probe scores with varying amounts of training data.

4.4 Text retrieval and retrieval augmented generation (RAG)

To compare eCLIP's cross-modal functionality with that of CLIP we focused on text retrieval task using the Open-I dataset, which consists of pairs of X-rays and radiology reports. We used the FAISS vector database [8] to index the text embeddings generated by the text encoder. For a given X-ray image I_i, we then retrieve the closest text reports from the database based on the cosine similarity in the embedding space, $min_j(v_i \cdot t_j)$. Results in Tab. 2 compare the performance of eCLIP against CLIP in text retrieval measured in Recall@1, 5, and 10. The performance of eCLIP indicates a notable improvement in its embedding quality. Note that our evaluation followed a strict criterion for recall computation, where a retrieval was counted as successful only if the exact correct report was identified. While more nuanced measures based on semantic similarity could be employed [59], we opted this approach to maintain a clear and simple evaluation framework.

Next we extend our analysis from retrieval to report generation using a frozen Large Language Model (LLM), Mistral 7B Instruct v2 [19], aiming to generate radiology reports through Retrieval Augmented Generation (RAG). This setup tests the CLIP model's capacity to retrieve texts which can be used to prompt an LLM to generate a report without finetuning on medical data. First we randomly selected 389 samples from the Open-I dataset for testing and utilized the FAISS database to index the reports from the remaining samples (i.e., training set). Given a test image we retrieve five closest reports from the training set and use them in the prompt for the LLM to generate a report for the test image. The eCLIP variant showed a small but consistent improvement over the base model in generating reports, as indicated in Tab. 3. A comparative analysis of generated report versus ground truth shown in Tab. 5, with discrepancies marked, further

Table 2: Performance on Image to Report retrieval with Open-I dataset. We report the Recall@{1, 5, 10}

Model	R@1	R@5	R@10
$\text{CLIP}_{\text{Swin Tiny}}$	3.1	7.6	11.3
$+m^3$-mix	2.4	6.5	9.8
+expert (ours)	<u>3.7</u>	<u>9.4</u>	<u>13.4</u>
$+\text{expert}^{\mathcal{P}}$(ours)	3.1	8.2	11.7
$\text{CLIP}_{\text{ViT Base}}$	3.7	9.2	13.2
$+m^3$-mix	4.1	8.8	13.0
+expert (ours)	**<u>4.4</u>**	**<u>10.3</u>**	**<u>13.5</u>**
$+\text{expert}^{\mathcal{P}}$(ours)	3.9	9.4	13.2

Table 3: Performance on report generation with Open-I dataset. We report the BLEU-2 score (BL-2), BERT recall score [64] (B-R), Cosine similarity between the sentence embeddings of the generated and ground-truth report for MPNet [42] Sentence Transformer model [38] (S_{emb}), and for the CheXBERT model [41] (CB_{emb})

Model	BL-2	B-R	S_{emb}	CB_{emb}
CLIP	0.172	**0.713**	0.791	0.492
$+ m^3$-mix	0.172	0.711	0.788	0.496
eCLIP	**0.177**	0.712	**0.795**	**0.506**

Table 4: Ablation Study with Swin Tiny. Zero-shot performance on CheXpert (CXP) and CXR14 (C14) datasets for the base CLIP and models with expert annotation integration (+E) is presented. Methods include mask multiplication ($\odot$), CNN, and Multi-headed Attention (MHA) encoders. Key augmentations: Mixup (+M), Curriculum Learning (+C), and Encoder Priming (+P) demonstrate performance gains. A control with a random mask (*rand*) confirms the significance of expert annotations. We report macro-averaged F1 scores from three random initialization

Method	CXP	C14
Base	$0.517_{\pm.024}$	$0.169_{\pm.003}$
$\odot$ Mask (+ E)	$0.540_{\pm.019}$	$0.165_{\pm.006}$
CNN Encoder (+ E)	$0.534_{\pm.012}$	$0.163_{\pm.008}$
MHA Encoder (+ E)	$0.534_{\pm.013}$	$0.153_{\pm.002}$
MHA Encoder (+ E, M)	$0.532_{\pm.018}$	$0.160_{\pm.010}$
MHA Encoder (+ E, M, C)	$0.545_{\pm.008}$	$0.173_{\pm.018}$
MHA Encoder (+ rand, M, C, P)	$0.537_{\pm.020}$	$0.166_{\pm.003}$
MHA Encoder (+ E, M, C, P)	$0.549_{\pm.016}$	$0.172_{\pm.006}$

validates the effectiveness of eCLIP's embeddings in supporting complex cross-modal tasks. Additional details, including LLM prompts and generated report samples are available in the Supplement.

4.5 Embedding Quality

For qualitative evaluations, we first examine the histogram of the cosine similarities of the embeddings from different abnormality subgroups obtained from the CLIP image encoder. In Fig. 5 *(top row)*, we can see that the similarities for the CLIP model has considerably dropped below 1 after continual pretraining on MIMIC-CXR compared to Fig. 1. This indicates that the model's ability to distinguish between different conditions has improved. The introduction of expert annotations in the eCLIP variant further improves this with mean cosine similarities for 'normal' versus cardiomegaly, atelectasis and opacity dropping to 0.36, 0.4 and 0.4 respectively.

Our evaluation of uniformity and alignment reveals that eCLIP surpasses both the MIMIC-pretrained and internet-pretrained models in these key metrics,

Table 5: Random samples of generated report. For each image in the Open-I dataset, the five closest text snippets based on embedding cosine similarity is used as prompts for Mistral 7B LLM. Utilizing in-context learning, we prompt the LLM with two such snippet-report pairs. The conditions that the generated report identified correctly are highlighted in green while those it missed are shown in red.

Problems	Ground Truth	Generated
normal	The lungs and pleural spaces show no acute abnormality. Heart size and pulmonary vascularity within normal limits.	The lungs are clear with no evidence of consolidation, pneumothorax, or pleural effusion. Heart size and mediastinal contour are normal. No acute abnormalities noted in the visualized osseous structures.
Pulmonary Atelectasis. Pleural Effusion. Cardiomegaly. Pulmonary Congestion	AP and lateral views were obtained. Bibasilar atelectasis and small left-sided pleural effusion. Stable cardiomegaly. No pneumothorax. Mild pulmonary vascular congestion.	The lungs show bibasilar airspace opacities, possibly due to infection or atelectasis. Heart size is mildly enlarged. No pneumothorax or pleural effusion.
Lung. Density. Emphysema	There is hyperexpansion. The heart size is normal. There is no pleural effusion or pneumothorax. Two circular densities overlying the right ribs which were not present in the XXXX CT. No focal infiltrates	The lungs are hyperexpanded with clear fields and no acute consolidations or pleural effusions. Heart size and pulmonary vascularity are within normal limits. Degenerative changes are present in the thoracic spine. A calcified granuloma is noted in the left perihilar region. No pneumothorax or large pleural effusion is observed.

indicating a marked improvement in the quality of embeddings (Fig. 5, *bottom left*). We also note a modest decrease in the modality gap with eCLIP (Fig. 5, bottom center). Clustering analysis via K-means (with k=5 for 5 abnormalities in data) highlights eCLIP's superior performance in grouping abnormalities, as seen from improved scored in Normalized Mutual Information (NMI), Silhouette score, and Calinski-Harabasz (CH) index (Fig. 5, bottom right).

4.6 Ablation Study

Our ablation study with the Swin Tiny encoder shows the impact of key components in our eCLIP model: multi-headed attention (MHA) layer for heatmap procesing, curriculum learning for phased introduction of expert annotations, mixup augmentation to compensate for limited number of export annotated data and priming of heatmap processor during initial training phase. Results shown in Tab. 4 reveal that the MHA-based heatmap processor improves zero-shot classification performance on CheXpert 5x200 and CXR 14x100 datasets compared to basic methods like direct application of heatmaps as mask ($\odot Mask$) or using a CNN encoder. We note a significant performance drop with randomly generated heatmaps versus expert eye-gaze heatmaps. This highlights that while our methodological improvements contribute to the performance gains, the in-

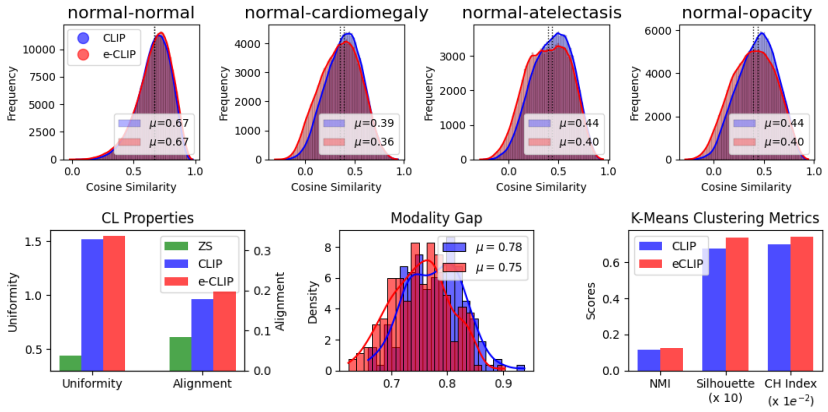


Fig. 5: Qualitative Analysis of CLIP Pretraining. *top row* illustrates the cosine similarity distributions for CLIP and eCLIP image embeddings. *bottom left and center* sections display uniformity, alignment, and modality gap comparisons among the internet pretrained model (ZS), CLIP pretrained on MIMIC, and eCLIP. *bottom right* details K-means clustering metrics for image embeddings with k=5 for both CLIP and eCLIP models.

tegration of meaningful, expert-derived signals is essential for achieving optimal results.

5 Conclusion

We introduce eCLIP, an adaptation of CLIP, demonstrating the integration of radiologist eye-tracking heatmap to overcome challenges faced in multi-modal contrastive learning. This study highlights the impact of integrating these high-quality expert annotations on improving the quality of learned embeddings and assess its influence on sample efficiency and cross-modal retrieval tasks. An important future research direction would be extending this approach to include expert annotations from the text modality (e.g., by adapting SimCSE [9]) and to leverage the temporal dynamics of eye-tracking data by aligning the sequential frames with the corresponding report snippets.

Limitations Our study is limited by the small size of expert annotated data and thus does not comprehensively analyze the impact of size or distribution of expert annotations across different abnormalities. eCLIP also incurs extra computational costs during training due to the additional forward pass required for processing expert images in the warmup and cool-down phases. Additionally, the clinical relevance of generated radiology reports has not been validated by medical experts, relying instead on standard metrics known for potential biases and inaccuracies in reflecting clinical accuracy [60].

Acknowledgements

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, and grants 352986, 358246) and EU (H2020 grant 101016775 and NextGenerationEU). We acknowledge the computational resources provided by Aalto Science-IT project. We acknowledge CSC for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through Finland.

References

1. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323 (2019) 9
2. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 41–48 (2009) 7
3. Bigolin Lanfredi, R., Zhang, M., Auffermann, W.F., Chan, J., Duong, P.A.T., Srikumar, V., Drew, T., Schroeder, J.D., Tasdizen, T.: REFLACX, A dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. Scientific Data **9**(1), 350 (2022) 5
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020) 1, 4
5. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023) 1
6. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016) 2, 8
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 8
8. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. arXiv preprint arXiv:2401.08281 (2024) 11
9. Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021) 14
10. Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P.: M3AE: Multimodal masked autoencoders learn transferable representations. Tech. rep., Technical Report 4
11. Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., Grover, A.: CyCLIP: Cyclic contrastive language-image pretraining. Advances in Neural Information Processing Systems **35**, 6704–6719 (2022) 2, 4

12. Gu, S., Clark, C., Kembhavi, A.: I can't believe there's no images! learning visual tasks using only language supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2672–2683 (2023) 4
13. Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., Yao, J.: Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. Advances in Neural Information Processing Systems **35**, 37704–37718 (2022) 3
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022) 4
15. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: 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. pp. 3942–3951 (2021) 2, 4, 8, 9
16. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**(9), 2307–2316 (2023) 5
17. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019) 8
18. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021) 1
19. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7B. arXiv preprint arXiv:2310.06825 (2023) 11
20. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR-JPG-Chest radiographs with structured labels. PhysioNet (2019) 8
21. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019) 5, 8
22. Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J.T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E.A., et al.: Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. Scientific Data **8**(1), 92 (2021) 3, 5
23. Krishnan, R., Rajpurkar, P., Topol, E.J.: Self-supervised learning in medicine and healthcare. Nature Biomedical Engineering **6**(12), 1346–1352 (2022) 2
24. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23390–23400 (2023) 4
25. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022) 2, 4
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 1
27. McInnes, L., Healy, J., Melville, J.: UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018) 3

28. Mo, S., Kim, M., Lee, K., Shin, J.: S-clip: Semi-supervised vision-language learning using few specialist captions. Advances in Neural Information Processing Systems **36** (2024) 4
29. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023) 5
30. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European Conference on Computer Vision. pp. 529–544. Springer (2022) 4
31. Naeem, M.F., Xian, Y., Zhai, X., Hoyer, L., Van Gool, L., Tombari, F.: SILC: Improving vision language pretraining with self-distillation. arXiv preprint arXiv:2310.13355 (2023) 4
32. Nukrai, D., Mokady, R., Globerson, A.: Text-only training for image captioning using noise-injected clip. arXiv preprint arXiv:2211.00575 (2022) 4
33. Oh, C., So, J., Byun, H., Lim, Y., Shin, M., Jeon, J.J., Song, K.: Geodesic multi-modal mixup for robust fine-tuning. Advances in Neural Information Processing Systems **36** (2023) 2, 3, 4, 5, 6, 8
34. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 1, 5
35. Palepu, A., Beam, A.: Tier: Text-image entropy regularization for medical clip-style models. In: Machine Learning for Healthcare Conference. pp. 548–564. PMLR (2023) 4
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 1, 2, 9
37. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 1
38. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992 (2019) 12
39. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., Arteaga, V., Galperin-Aizenberg, M., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019) 8
40. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022) 1
41. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1500–1519 (2020) 12
42. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: MPNet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems **33**, 16857–16867 (2020) 12

43. van Sonsbeek, T., Zhen, X., Mahapatra, D., Worring, M.: Probabilistic integration of object level annotations in chest x-ray classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3630–3640 (2023) 5
44. Sowrirajan, H., Yang, J., Ng, A.Y., Rajpurkar, P.: MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models, 2021. URL https://arxiv. org/abs (2010) 5
45. Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-CLIP: A clip model focusing on wherever you want. arXiv preprint arXiv:2312.03818 (2023) 4
46. Tschannen, M., Mustafa, B., Houlsby, N.: CLIPPO: Image-and-language understanding from pixels only. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11006–11017 (2023) 2, 4
47. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. NEJM AI **1**(3), AIoa2300138 (2024) 5
48. Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., Bengio, Y.: Manifold mixup: Better representations by interpolating hidden states. In: International Conference on Machine Learning. pp. 6438–6447. PMLR (2019) 3
49. Verma, V., Luong, T., Kawaguchi, K., Pham, H., Le, Q.: Towards domain-agnostic contrastive learning. In: International Conference on Machine Learning. pp. 10530–10541. PMLR (2021) 3, 4, 8
50. Wang, B., Pan, H., Aboah, A., Zhang, Z., Keles, E., Torigian, D., Turkbey, B., Krupinski, E., Udupa, J., Bagci, U.: GazeGNN: A gaze-guided graph neural network for chest x-ray classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2194–2203 (2024) 5
51. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2495–2504 (2021) 5
52. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020) 5
53. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2097–2106 (2017) 8, 9
54. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022) 2, 4, 8, 9
55. Weers, F., Shankar, V., Katharopoulos, A., Yang, Y., Gunter, T.: Masked autoencoding does not help natural language supervision at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23432–23444 (2023) 4
56. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training. medRxiv pp. 2023–01 (2023) 5
57. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying CLIP data. arXiv preprint arXiv:2309.16671 (2023) 1

58. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.H., Kiraly, A., Kazemzadeh, S., Melamed, Z., et al.: Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. arXiv preprint arXiv:2308.01317 (2023) 5
59. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: CXR-CLIP: Toward large scale chest X-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023) 4, 8, 9, 11
60. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023) 14
61. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021) 1
62. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 3, 6
63. Zhang, K., Yang, Y., Yu, J., Jiang, H., Fan, J., Huang, Q., Han, W.: Multi-task paired masking with alignment modeling for medical vision-language pre-training. IEEE Transactions on Multimedia (2023) 4
64. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. In: International Conference on Learning Representations (2019) 12
65. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022) 2, 4
66. Zhang, Y., HaoChen, J.Z., Huang, S.C., Wang, K.C., Zou, J., Yeung, S.: Diagnosing and rectifying vision models using language. arXiv preprint arXiv:2302.04269 (2023) 4
67. Zhang, Y., Sui, E., Yeung, S.: Connect, Collapse, Corrupt: Learning cross-modal tasks with uni-modal data. In: The Twelfth International Conference on Learning Representations (2024) 2, 4