Supplement

Yogesh Kumar and Pekka Marttinen

Department of Computer Science, Aalto University, Finland

A Detailed Experiment Setup

A.1 Pretraining with Expert Annotations

For pretraining both CLIP and eCLIP models, we utilize the MIMIC-CXR dataset. Expert annotations, in the form of heatmaps, are derived from a subset of MIMIC-CXR, the EGD-CXR dataset [22], which comprises of 1080 samples. We employ the author's official preprocessing code to convert the eye-tracking fixation data into heatmaps. The data tuple (image, text, heatmap) is then used for training with contrastive learning. We employ two dataloader: one for the main dataset without heatmap ("main loader") and another for the subset with expert heatmaps ("expert loader"). In each training iteration, one batch is fetched from each loader; the CLIP model processed only the main batch, while the eCLIP model has the flexibility to use either just the main batch or both batches. Listing 2.1 shows the Pytorch-like pseudocode for the eCLIP model.

The utilization of the expert batch in eCLIP is determined by a curriculum probability, initially set to zero during the cold start phase. This probability linearly increases to p_{max} during the warmup phase, then linearly decreases to p_{min} during the cool-down phase, where it remains for the remainder of the training process. p_{max} was set to 0.5 for all experiments, while p_{min} was set to 0.05 for the ViT Base model and to 0.1 for all other models.

A.2 m^2-mixup vs m^3-mixup

We illustrate m^2-mixup in Fig. 3, where embeddings from image and text domains are mixed to mitigate the modality gap, as proposed by Oh et al. [33]. Oh et al. [33] further introduce m^3-mixup, which combines m^2-mixup with corresponding unimodal mixups. Specifically,

$$\mathcal{L}_{m^3} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{m^2} + \mathcal{L}_{uni}$$

For further details, please refer to Oh et al. [33].

A.3 Linear Probe Experiments

In our linear probe experiments, we utilize the CLIP and eCLIP with Swin Tiny as the image encoder following other recent similar works [54,59]. The pretrained model's image encoder is entirely frozen and we append a linear layer for classification. This layer's output dimension is set to 1 for the Pneumonia dataset, 8 for

```
def heatmap_processor(image, heatmap):
    # reshapes and transposes have been omitted for brevity
    pathches = patchify(image)
    heatmap_patches = patchify(heatmap * image)
    processed_patches = multi_head_attention(
        q=heatmap_patches, k=patches, v=patches)
    reconstructed_image = unpatchify(processed_patches)
    return reconstructed_image

class ExpertClipImageEncoder:
    self.base = ViT()
    self.projector = ProjectionBlock()

    def forward(image, heatmap=None):
        mse_loss = None
        B, C, H, W = image.size()
        # priming
        reconstructed_image = heatmap_processor(
        image, torch.ones((B, 1, H, W)))
        mse_loss = mse_loss_fn(image, reconstructed_image)

        image_features = self.base(image)
        # project to shared embedding dimension
        image_embed = self.projector(image_features)

        if heatmap is not None:
            expert_image = heatmap_processor(image, heatmap)
            # mixup augmentation
            lambda_ = beta(alpha=0.3).sample()
            expert_image_m = mixup(image, expert_image, lambda_)

            expert_image_features = self.base(expert_image_m)
            expert_image_embed = self.projector(expert_image_features)

        return image_embed, expert_image_embed, mse_loss

# include expert batch based on curriculum learning probability
# curriculum_prob is varied between cpmin to cpmax (typically, 0.1 & 0.5)
use_expert =  np.random.rand() < curriculum_prob

# compute clip loss
clip_loss = clip_loss_fn(
    concat([image_embed, expert_image_embed]) if use_expert else image_embed,
    concat([text_embed, text_embed]) if use_expert else text_embed
)
```

Listing 2.1: PyTorch-like pseudocode for eCLIP implementation

CXR-8 and 5 for OpenI-5. We allocate 10% of the training data as validation set and conduct training over 5 epochs with a cosine decay learning rate schedule with linear warmup for 10% of the total training steps. Base learning rates are set to $2e^{-5}$ for the Pneumonia dataset and $1e^{-5}$ for both CXR-8 and OpenI-5. We employ binary cross entropy as the loss function and the model selection for testing is based on the epoch with the lowest validation loss.

A.4 Zero-shot Classification

Following the CLIP paper [36], we generate descriptive prompts for each label, mirroring the patterns found in the radiology reports of our pretraining data. For example, within the pneumonia detection task, a 'normal' X-ray is prompted as "Chest radiograph with normal findings, no signs of pneumonia", while a prompt for pneumonia diagnosed X-rays would read "Radiograph of the chest displaying multifocal opacities, suggestive of viral pneumonia". We apply the

ensemble promoting technique, where we generate multiple variations of each label's prompt to create a list of text embeddings for each label. The mean of these embeddings serves as the representation for the corresponding label. The specific prompts utilized for each label have been included in the Supplement section.

These prompts are converted into embeddings using the text encoder of the trained CLIP model, while the images are processed using the corresponding image encoder to produce image embeddings. Classification is then performed by selecting the label whose text emebdding is most similar to the image embedding, as determined by cosine similarity. We use the prompts used in [15], samples from which are shown below.

ZS Prompts

Atelectasis - mild subsegmental atelectasis
Cardiomegaly - cardiac silhouette size is mildly enlarged
Consolidation - increased reticular consolidation at the lower lung zone
Edema - mild pulmonary edema
Pleural Effusion - stable right bilateral pleural effusion
Pneumonia - Bronchopneumonia pattern suggestive of bacterial infection

B Additional Results

In Table 1 we show the zero-shot classification accuracies for the CLIP, eCLIP and baseline models.

Table 1: Zero-shot classification performance on 4 X-ray datasets and model configurations, reported as accuracy 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$_{\text{Resnet50}}$	$0.498_{\pm.017}$	$0.462_{\pm.014}$	$0.731_{\pm.013}$	$0.173_{\pm.002}$
+naive	$0.409_{\pm.061}$	$0.369_{\pm.044}$	$0.669_{\pm.049}$	$0.145_{\pm.020}$
+DACL	$\underline{0.530_{\pm.016}}$	$0.438_{\pm.007}$	$0.752_{\pm.007}$	$0.179_{\pm.008}$
+m^3-mix	$0.525_{\pm.004}$	$0.469_{\pm.004}$	$0.748_{\pm.004}$	$0.171_{\pm.003}$
+expert (ours)	$0.518_{\pm.001}$	$0.436_{\pm.008}$	$0.730_{\pm.023}$	$0.179_{\pm.028}$
+expert$^{\mathcal{P}}$(ours)	$0.520_{\pm.004}$	$\underline{0.478_{\pm.008}}$	$\underline{0.753_{\pm.000}}$	$0.168_{\pm.001}$
CLIP$_{\text{Swin Tiny}}$	$0.529_{\pm.020}$	$0.454_{\pm.003}$	$0.799_{\pm.004}$	$0.188_{\pm.006}$
+naive	$0.536_{\pm.008}$	$0.459_{\pm.018}$	$0.793_{\pm.002}$	$0.188_{\pm.012}$
+DACL	$0.481_{\pm.005}$	$0.398_{\pm.020}$	$0.758_{\pm.005}$	$0.124_{\pm.019}$
+m^3-mix	$0.561_{\pm.006}$	$\underline{0.467_{\pm.005}}$	$0.802_{\pm.001}$	$0.208_{\pm.007}$
+expert (ours)	$0.549_{\pm.016}$	$0.443_{\pm.023}$	$0.799_{\pm.004}$	$0.195_{\pm.003}$
+expert$^{\mathcal{P}}$(ours)	$\underline{0.565_{\pm.004}}$	$0.465_{\pm.006}$	$\underline{\mathbf{0.810_{\pm.001}}}$	$\underline{\mathbf{0.210_{\pm.001}}}$
CLIP$_{\text{ViT Small}}$	$0.524_{\pm.023}$	$0.441_{\pm.007}$	$0.796_{\pm.000}$	$0.179_{\pm.007}$
+naive	$0.532_{\pm.017}$	$0.454_{\pm.028}$	$0.793_{\pm.003}$	$0.169_{\pm.022}$
+DACL	$0.485_{\pm.024}$	$0.402_{\pm015}$	$0.753_{\pm.009}$	$0.154_{\pm.007}$
+m^3-mix	$\underline{0.561_{\pm.002}}$	$0.455_{\pm.003}$	$0.786_{\pm.002}$	$0.182_{\pm.001}$
+expert (ours)	$0.548_{\pm.017}$	$\underline{0.456_{\pm.010}}$	$0.795_{\pm.005}$	$\underline{0.185_{\pm.015}}$
+expert$^{\mathcal{P}}$(ours)	$0.556_{\pm.001}$	$0.437_{\pm.004}$	$\underline{0.805_{\pm.001}}$	$0.180_{\pm.003}$
CLIP$_{\text{ViT Base}}$	$0.540_{\pm.011}$	$0.470_{\pm.007}$	$0.788_{\pm.012}$	$0.200_{\pm.006}$
+naive	$0.503_{\pm.010}$	$0.434_{\pm.011}$	$0.787_{\pm.002}$	$0.176_{\pm.008}$
+DACL	$0.484_{\pm.006}$	$0.406_{\pm.002}$	$0.737_{\pm.001}$	$0.204_{\pm.098}$
+m^3-mix	$0.544_{\pm.015}$	$0.458_{\pm.005}$	$0.771_{\pm.009}$	$0.188_{\pm.005}$
+expert (ours)	$\underline{\mathbf{0.564_{\pm.023}}}$	$\underline{\mathbf{0.476_{\pm.005}}}$	$\underline{0.799_{\pm.003}}$	$0.204_{\pm.020}$
+expert$^{\mathcal{P}}$(ours)	$0.557_{\pm.011}$	$0.464_{\pm.018}$	$0.782_{\pm.010}$	$0.197_{\pm.0.013}$

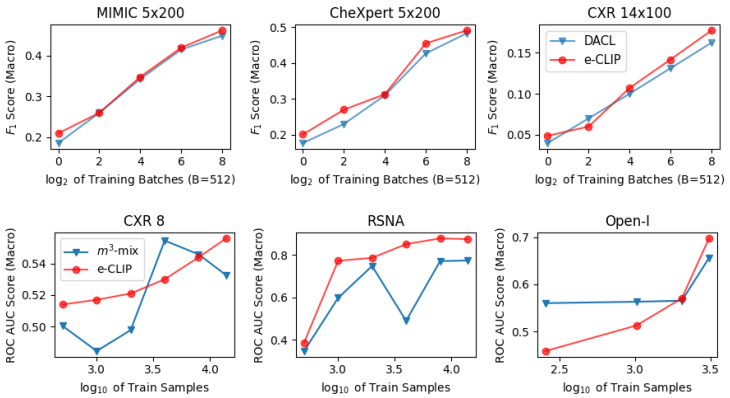

Fig. 1: Sample Efficiency. *(top row)* Zero-shot performance on three multi-label classification test sets for DACL and eCLIP Swin Tiny models, trained with varying amounts of training batches. *(bottom row)* Linear probe scores with varying amounts of training data for m^3-mixup and eCLIP Swin Tiny models.

Figure 1 demonstrates the sample efficiency of different models. The top row shows the zero-shot performance on three multi-label classification test sets for DACL and eCLIP Swin Tiny models, trained with varying amounts of training batches. This highlights each model's ability to generalize with limited data. The bottom row presents the linear probe scores for m3-mixup and eCLIP Swin Tiny models, evaluated with different amounts of training data. This illustrates how quickly each model learns and performs as more data is provided. These results underscore the importance of sample efficiency in model performance.

C Visualize Embeddings with UMAP

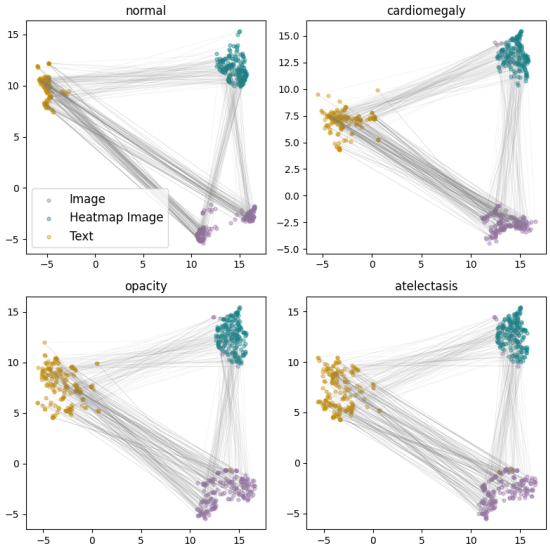

Fig. 2: 2D UMAP Projection of Embeddings Figure shows the UMAP projection of the Image, Text and heatmap processed Image embedding generated by eCLIP with Swin Tiny encoder. We use Open-I dataset for image and text and since expert annotation is unavailable for this dataset, we generate random uniform masks to simulate heatmaps.

We utilize the trained eCLIP model with the Swin Tiny encoder to examine UMAP projections of embeddings derived from the Open-I dataset. This dataset is categorized into subgroups based on the presence of specific abnormalities, as indicated in the 'Problem' column, which contains radiologist annotations for each image. Four primary abnormalities –'normal,' 'cardiomegaly,' 'atelectasis,' and 'opacity' – form the basis of our subgroup categorization. We ensure that samples within each subgroup are mutually exclusive, containing only one of these abnormalities.

To generate embeddings, we use the trained image and text encoders from our eCLIP model. Since the Open-I dataset lacks actual expert-annotated heatmaps, we simulate this condition by creating random heatmaps for each image. Thus we obtain the standard image embedding (v_i), text embedding (t_i) and expert image embedding (v_i^E) for each sample across the subgroups. These embeddings are projected into a 2D space using UMAP with cosine similarity as the metric. Subsequently, we visualize the 2D UMAP projections for each subgroup separately, facilitating a detailed inspection of the embedding distribution and the influence of expert annotations on the model's representation space. This is shown in Fig. 2

D Retrieval Augmented Generation of Radiologist Report

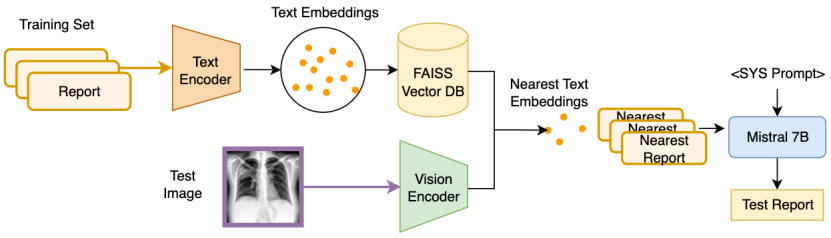

Fig. 3: Retrieval Augmented Generation of Radiology Reports. Radiology reports from the training corpus are encoded using the CLIP/eCLIP's text encoder to obtain text embeddings, which are then stored in a FAISS vector database. For a test image, the corresponding image embedding is obtained using the CLIP/eCLIP's image encoder. This image embedding is queried against the FAISS database to find the nearest text embeddings, which are used as prompts for the Mistral 7B Large Language Model (LLM). The LLM then generates a test report based on these prompts.

We detail our approach to generate radiologist reports by augmenting a Large Language Model (LLM) with retrieved report snippets from the training corpus. This method is designed to tackle the challenge of generating medically relevant and coherent radiology report without direct image examination or explicit fine-tuning on medical datasets. The process involves the following key steps:

1. **Text Embedding and Indexing:** We utilize the Open-I dataset to create the source of the radiology reports. These reports are processed through the CLIP/eCLIP trained text encoder to produce embeddings of dimension 512. These embeddings are then normalized and indexed using FAISS [8] vector database, facilitating efficient retrieval based on similarity.
2. **Text Retrieval and Clustering:** For a given test X-ray image, we first compute its embedding using the CLIP/eCLIP trained image encoder and then query the FAISS index to retrieve the closest report snippets. The "closeness" is based on cosine similarity of the normalized embeddings, ensuring that the retrieved texts are semantically relevant to the image's medical context. We use K-means to categorize the embeddings of these snippets into distinct groups. This clustering ensures the selection of representative sentences that encapsulate the primary observations within each group, thereby preserving the diversity of the retrieved reports. We require five closest snippets for prompting the LLM, so we retrieve four times this from the FAISS index for clustering.

3. **Report Generation with LLM:** Leveraging the retrieved snippets as context, we employ a frozen LLM, Mistral 7B [19], to generate a comprehensive radiology report. The aim is to produce reports that closely mimic those written by radiologists, based on the insights from the retrieved texts.

Fig. 3 shows the schematic of the steps involved in report generation using a frozen LLM.

LLM Prompts We use the following system and user prompts to use the frozen Mistral 7B model to generate the radiology report. We provide two exemplars of retrieved reports and the corresponding generated report as samples for the In-Context Learning (ICL) as the user prompt for the LLM.

<SYSTEM PROMPT>
You are to act as a radiologist, trained to generate radiology reports. Your task is to synthesize the information from the closest report snippets provided below into a comprehensive and medically accurate radiologist report for each case. Craft a comprehensive response that is concise, succinct, and focuses on the key findings and potential diagnoses. Your report should maintain a professional tone, with clarity and precision in medical terminology, suitable for medical experts. Remember to be concise, succinct, and focus on the key findings and potential diagnoses, avoiding unnecessary elaboration.

USER
The following snippets are from reports closely related to the patient's X-ray image.
< Retrieved Text >
Based on these, generate a radiologist report.

Evaluation Metrics For evaluating the generated radiology reports, we employ metrics traditionally used in text generation and translation fields, namely BLEU-2 and BERT Score [64]. We also compute the cosine similarity between the embeddings of the generated reports and the ground truth, as derived from a reference model. This approach allows for a broader assessment of semantic congruence. Specifically, we utilize sentence transformer models [38] known for their effectiveness in sentence-level comparison tasks. We employ the 'all-mpnet-base-v2' [42] model for its general semantic understanding, and the 'CheXBERT' model [41], for its domain-specific performance in medical classification tasks. These models facilitate a more comprehensive and contextually relevant evaluation of the linguistic and clinical content of the generated reports.

Generated Radiology Report Samples We provide more randomly sampled generated radiology reports in Tab. 2

Table 2: More 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.

Problems	Ground Truth	Generated
normal	The heart is normal in size. The mediastinum is unremarkable. The lungs are clear..	The lungs are clear with no evidence of consolidation, pleural effusion, or pneumothorax. Heart size is within normal limits. The visualized osseous structures are unremarkable.
normal	XXXX XXXX and lateral chest examination was obtained. One AP view is expiratory and was repeated. The heart silhouette is normal in size and contour. Aortic XXXX appear unremarkable. Lungs demonstrate no focal infiltrates. There is no effusion or pneumothorax.	The heart size is normal. The lungs are clear with no evidence of consolidation, pneumothorax, or pleural effusion. No acute abnormalities are noted in the visualized osseous structures. The previously seen right upper lobe mass lesion is not present on the current study.
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.
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.
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.
Lung; Spine	Lungs are mildly hyperexpanded. The lungs are clear. There is no focal airspace consolidation. No pleural effusion or pneumothorax. Heart size and mediastinal contour are within normal limits. There are diffuse degenerative changes of the spine.	The heart size and mediastinal contours are within normal limits. The lungs are clear, but with slightly low lung volumes and subtle left basal opacities. No pleural effusion, pneumothorax, or acute osseous abnormalities are noted. There is evidence of degenerative changes in the thoracic spine on the lateral view.
Calcified Granuloma; Opacity	Heart size and pulmonary vascularity appear within normal limits. Calcified granuloma is present in the right base. No pneumothorax or pleural effusion is seen. In the lateral right base is identified an ill-defined somewhat oblong opacity. This was not present on the previous study. The remainder of the lungs appear clear.	The lungs are clear with mild hyperexpansion, suggestive of underlying emphysema. Heart size is normal. No pleural effusions, pneumothoraces, or focal consolidations. The cardiomediastinal silhouette and osseous structures are unremarkable.