# CARZero: Cross-Attention Alignment for Radiology Zero-Shot Classification

Haoran Lai[1,2,*] Qingsong Yao[3,*] Zihang Jiang[2,†] Rongsheng Wang[2]
Zhiyang He[4] Xiaodong Tao[4] S. Kevin Zhou[1,2,3,†]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine,
University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China
[2] Suzhou Institute for Advanced Research, University of Science and Technology of China,
Suzhou, Jiangsu, 215123, P.R.China
[3] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences
(CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[4] Medical Business Department, iFlytek Co.Ltd, Hefei 230088, China

{haoranlai, rongsheng_wang}@mail.ustc.edu.cn, yaoqingsong19@mails.ucas.edu.cn
{zyh, xdtao}@iflytek.com, jzh0103@ustc.edu.cn, s.kevin.zhou@gmail.com

## Abstract

*The advancement of Zero-Shot Learning in the medical domain has been driven forward by using pre-trained models on large-scale image-text pairs, focusing on image-text alignment. However, existing methods primarily rely on cosine similarity for alignment, which may not fully capture the complex relationship between medical images and reports. To address this gap, we introduce a novel approach called Cross-Attention Alignment for Radiology Zero-Shot Classification (CARZero). Our approach innovatively leverages cross-attention mechanisms to process image and report features, creating a Similarity Representation that more accurately reflects the intricate relationships in medical semantics. This representation is then linearly projected to form an image-text similarity matrix for cross-modality alignment. Additionally, recognizing the pivotal role of prompt selection in zero-shot learning, CARZero incorporates a Large Language Model-based prompt alignment strategy. This strategy standardizes diverse diagnostic expressions into a unified format for both training and inference phases, overcoming the challenges of manual prompt design. Our approach is simple yet effective, demonstrating state-of-the-art performance in zero-shot classification on five official chest radiograph diagnostic test sets, including remarkable results on datasets with long-tail distributions of rare diseases. This achievement is attributed to our new image-text alignment strategy, which effectively addresses the complex relationship between medical images and reports. Code and models are available at https://github.com/laihaoran/CARZero.*

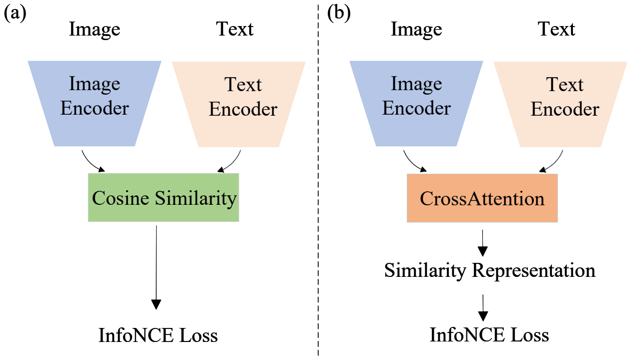

Figure 1. Comparison of the alignment scheme in Visual Language Pre-training: (left) handcrafted cosine similarity used in CLIP [6] and CheXzero [33]; (right) our proposed cross-attention alignment leveraging a novel similarity representation.

## 1. Introduction

Deep learning (DL) has achieved remarkable success in medical image recognition tasks. Prior studies [3, 18, 23, 34] have harnessed DL techniques for diagnosing diseases, yielding impressive results. However, these efforts often rely on laborious and costly annotations from clinical experts. Additionally, to reach an acceptable accuracy level, model training requires an extensive collection of labeled data for each specific disease, which can be both challenging and time-consuming. To address these issues, recent works [1, 15, 33, 38, 41, 42] have utilized paired images and reports for cost-effective disease diagnosis through zero-

[*] These authors contributed equally to this work
[†] Corresponding author

shot learning (ZSL) [29]. For rare diseases, where labeled training data is particularly scarce and difficult to obtain, ZSL emerges as a valuable tool for diagnosis.

Delving into the advanced medical ZSL methods, including ConVIRT [42], GLoRIA [15], CheXzero [33], MedKLIP [38], and KAD [41], we discover that contrastive learning serves as a foundational approach, aiming at minimizing the cosine similarity between paired image-text samples and maximizing it between unpaired ones. However, compared to natural images and texts, the relationship between medical images and reports is significantly more complex. For example, radiologists tend to describe multiple findings, diseases, and their locations within a single report, drawing upon various visual clues present in the corresponding medical images. Hence, we hypothesize that relying solely on the hand-crafted cosine similarity to measure the complex relationships between medical reports and images might be suboptimal.

To address this issue, we propose a CARZero method that leverages cross-attention alignment for radiology disease diagnosis within the setup of ZSL, a *simple yet effective* method of learning a similarity measurement that robustly represents the similarity of the medical semantic context. Specifically, as shown in Fig. 1, we use cross-attention [35] to compute global and local features from both modalities and result in features of mutual interaction, which is referred to the Similarity Representation (SimR). The SimR is then processed by a linear projector to obtain the final logit of similarity. Finally, we employ an InfoNCE loss [29] to introduce comparisons between positive and negative pairs, optimizing the model to learn discriminative features.

Furthermore, considering that medical reports are quite specialized and complex, the task of human-designed prompts poses an additional challenge to medical zero-shot classification. Recently, there has been significant research efforts [20, 21, 32, 45], dedicated to optimizing prompts, mainly with strategies that adjust the adaptability of prompts for downstream tasks. In our CARZero framework, we tackle this issue by aligning diagnostic prompts during both training and zero-shot inference phases. Fortunately, recent advancements in Large Language Models (LLMs) [43] have demonstrated significant capabilities in laguege comprehension and reformulation, enabling CARZero to standardize the diverse expressions found in reports into a cohesive and unified prompt format. This not only mitigates the challenges of manual prompt design but also unifies the format of diagnosis to improve zero-shot inference performance.

Our CARZero has been evaluated on a total of five public datasets. In particular, our CARZero achieves the state-of-the-art (SOTA) AUC performance of 0.810 in PadChest [2], which is a multi-label dataset with a long-tail distribution spanning 192 diseases. More surprisingly, our CARZero achieves zero-shot performance scores of 0.811 on ChestXray14 [37], surpassing the SOTA performances of 0.794 achieved with fine-tuning on 1% of the data.

To summarize, the contributions of this paper are listed as follows:

- We propose a novel cross-attention alignment for medical images and reports, utilizing SimR to articulate the complex relationships between medical images and reports, effectively aligning the features of both vision and text domains.
- We employ LLM to reformulate medical reports into the unified prompt template, ensuring the alignment of diagnostic expression during both the training and zero-shot inference phases.
- Tremendous experiments on five large-scale radiology diagnosis datasets confirm the zero-shot capabilities of our CARZero exceeding the SOTA zero-shot methods with a notable performance gap. Impressively, a significant improvement is achieved in diagnosing rare diseases.

## 2. Related Work

### 2.1. Zero-shot classification

For Vision Language Pretraining (VLP) tasks, previous works [24, 26] mainly use a fusion module to integrate image and text features, employing binary cross-entropy for classifying the combined features to determine if the image-text pair matches. Recently, CLIP [29] introduced contrastive learning, which measures the cosine similarity between image and text features, aiming to maximize it between the matching image-text pairs and minimize the unpaired ones. This work significantly advances the development of VLP for ZSL in visual recognition tasks. Following CLIP, many studies [4, 6, 8, 11, 14, 28, 31, 39] have utilized contrastive learning for aligning image-text, demonstrating the substantial potential of contrastive learning for VLP.

In the medical domain, VLP has demonstrated remarkable performance in ZSL for disease diagnosis. There has been a succession of outstanding works [15, 33, 38, 41, 42]. ConVIRT [42] first introduces contrastive learning to align medical images and reports. CheXzero [33] leverages the CLIP trained by nature data as pre-trained weights to achieve commendable performance on medical data. GLoRIA [15] further introduces the integration of global and local features alignment. MedKLIP [38] proposes the use of prior knowledge in the form of disease descriptions as an additional input to enhance representation learning. KAD [41] introduces word-based entity extraction to extract report information, thus improving the model's generalizability. For medical zero-shot classification tasks, these advancements represent meaningful progress. However, an important consideration they overlook is the intricate and nuanced relationship between medical images and reports, which is substantially more complex than the associations found in natural images and texts. Therefore, capturing the complex relationships between medical images and texts is key to improving the performance of medical zero-shot classification.

### 2.2. Cross-Attention in Modality Alignment

Existing works of cross-attention in modality alignment can be divided into two types. The first type uses cross-attention for modality fusion, optimizing the fused features with an image-text matching loss (ITM). For example, ALBEF [24] proposes a strategy of alignment before fusion, enabling the cross-attention module to merge the features of two modalities and optimize them using an ITM loss. BLIP [25] treats the image features as hidden states of text encoder to obtain fused modal features, which are then optimized with an ITM loss. The second type involves using cross-attention for modality projection transformation, followed by employing cosine similarity to calculate the similarity between the projected modalities and optimizing with an InfoNCE loss. For example, MGCA [36] employs cross-attention to project local features of both modalities, aligning the projected features afterward by InfoNCE loss. TEFAL [16] uses cross-attention to project video and audio on text, then optimizes these projections with an InfoNCE loss. In contrast to these established methodologies, as shown in Figure 1, this paper introduces a novel third type, which employs cross-attention to directly generate a high-level Similarity Representation between two modalities, subsequently optimized using an InfoNCE loss. This high-level SimR effectively captures the complex relationship between the two modalities, especially for medical images and reports. Moreover, InfoNCE loss, compared to the ITM loss, uses more negative samples and a softmax that is highly comparative, along with a process of logit normalization, which is more suitable for modality alignment.

### 2.3. Prompt Alignment

In VLP models, the performance of zero-shot classification is highly dependent on the design of the prompt. Typically, human-designed prompts require extensive searching to find the optimal template, which is time-consuming and labor-intensive. To improve efficiency, CoOP [45] introduced the concept of automatic prompts by learning prompts for downstream tasks. MaPLe [21] proposes a multi-modal prompt fine-tuning strategy, emphasizing the interplay between text and images in prompt construction. Imagic [20] proposes a novel framework for text-guided image editing, enabling precise alterations that resonate with the image's semantic context. GALIP [32] utilizes text-conditioned prompts to better adapt to downstream tasks, thereby improving complex image synthesis capabilities. Despite their effectiveness, these developments mainly leverage natural data, underscoring the significance of prompt design in the pre-training phase of cross-modality learning. Intuitively, using prompt templates as training data to align the training and testing prompts has been an effective method for prompt design. However, it is challenging to directly insert prompts into training data. Fortunately, Recent progress in LLMs [43] have shown enormous potential in semantic understanding, making it feasible to introduce prompt information into training texts. Therefore, leveraging large models for prompt alignment emerges as another key aspect in advancing medical zero-shot classification.

## 3. Method

In this section, we describe our proposed CARZero framework for zero-shot classification. As illustrated in Figure 2, first, we introduce a Cross-Attention Alignment that generates a SimR to represent the relationship between images and reports. Then, a linear layer projects the SimR onto a similarity matrix, which is then optimized using the InfoNCE loss. Moreover, we propose an LLM-based Prompt Alignment method that integrates prompt templates into the training data.

### 3.1. Feature Extraction

Assume that the training dataset contains $N$ samples denoted as $D_{\text{train}} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i \in \mathbb{R}^{H \times W \times C}$ represents a CXR image and $y_i$ represents its corresponding medical report. $H$, $W$, and $C$ refer to height, width, and channels, respectively. As illustrated in Figure 2, we introduce individual components of our architecture for feature extraction of different modalities, including an image encoder $\Phi_{\text{image}}$ and a text encoder $\Phi_{\text{text}}$.

**Image Encoder** The image encoder is utilized to extract the image features at different levels, as shown in Eq. (1).

$$[\boldsymbol{x}_i^l, \boldsymbol{x}_i^g] = \Phi_{\text{image}}(x_i), \tag{1}$$

where $\boldsymbol{x}_i^l \in \mathbb{R}^{L \times D}$ and $\boldsymbol{x}_i^g \in \mathbb{R}^D$ represent the local and global features, respectively. $L$ represents the number of local image patches and $D$ refers to the dimension of image features. The ViT-base [9] model is adopted as the visual encoder in our experiments.

**Text Encoder** As illustrated in Eq. (2), the text encoder is employed to extract text features from the given report. For each sample, a sentence is randomly selected from the report in each training iteration. In our experiments, to better adapt the text encoder for medical-related information extraction, we fine-tune BioBERT [22] with the clinical reports and use it for text encoding.

$$[\boldsymbol{y}_i^l, \boldsymbol{y}_i^g] = \Phi_{\text{text}}(y_i), \tag{2}$$

where $\boldsymbol{y}_i^l \in \mathbb{R}^{M \times D}$, $\boldsymbol{y}_i^g \in \mathbb{R}^D$ represent the word-based and sentence-based features, respectively. $M$ is the maximum

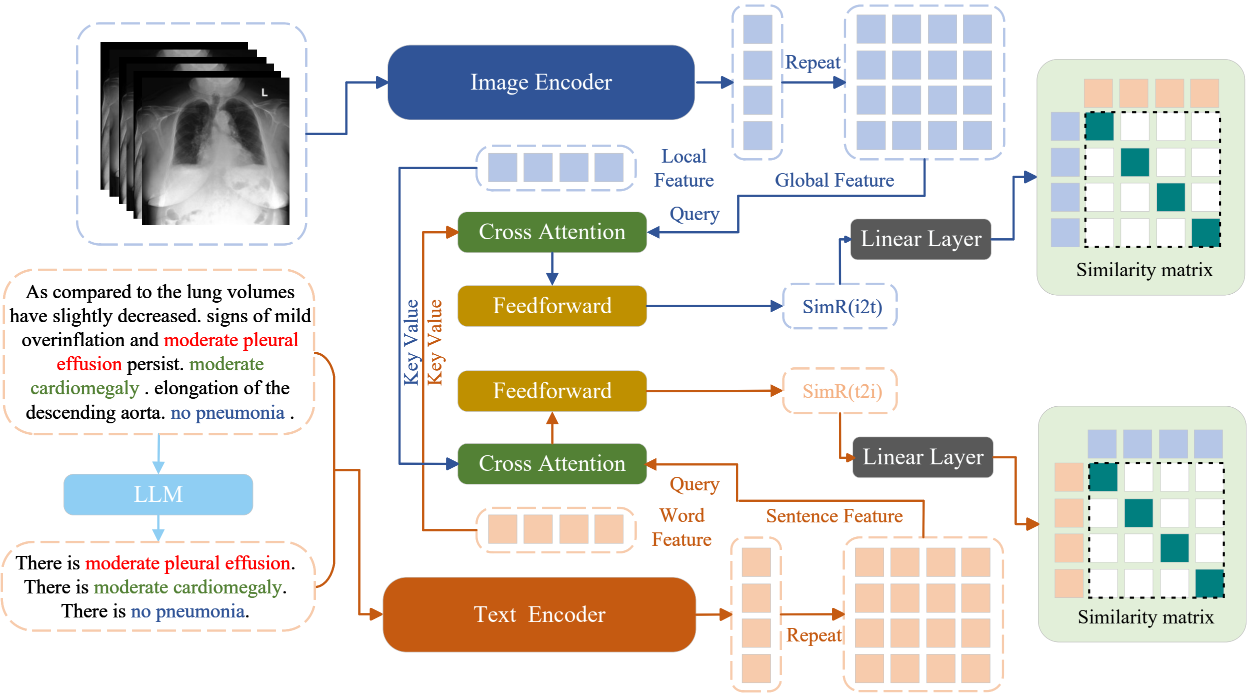
Image Encoder
Repeat
Local Feature
Global Feature
Query
Cross Attention
Feedforward
Linear Layer
SimR(i2t)
Similarity matrix
Key Value
Key Value
As compared to the lung volumes have slightly decreased. signs of mild overinflation and moderate pleural effusion persist. moderate cardiomegaly . elongation of the descending aorta. no pneumonia .
LLM
There is moderate pleural effusion. There is moderate cardiomegaly. There is no pneumonia.
Feedforward
SimR(t2i)
Cross Attention
Query
Linear Layer
Word Feature
Sentence Feature
Text Encoder
Repeat
Similarity matrix

Figure 2. The CARZero Network proposed in this paper consists of two stages. First, LLM is employed to generate prompt templates from medical reports. Second, text and vision encoders are used to extract features from image and text, which are fed into a cross-attention module to generate similarity for optimizing InfoNCE loss.

text length and $D$ refers to the dimension of text features. In our experiment, the dimensions of final output features from both the image and text encoders are the same.

### 3.2. Cross-Attention Alignment

Due to the complex relationship between medical images and reports, a cross-attention alignment module is proposed to align the feature spaces between images and reports. The objective is to employ cross-attention to obtain the SimR, which serves as a high-level representation of the correlation between the images and reports. During the training phase, the number of images and texts in a batch is equal, denoted as $I$ for the number of images and $T$ for the number of texts. For text-to-image ('t2i') alignment, sentence-based features $\boldsymbol{y}^g \in \mathbb{R}^{T\times D}$ are used as the query. To calculate SimR, the dimension of $\boldsymbol{y}^g$ is expanded and repeated to match the number of images, resulting in $\hat{\boldsymbol{y}}^g \in \mathbb{R}^{T\times I\times D}$. Then, the local features of images, $\boldsymbol{x}^l \in \mathbb{R}^{L\times I\times D}$ are used as key and value. Finally, the output from the cross-attention module is SimR, $SR_{\text{t2i}} \in \mathbb{R}^{T\times I\times D}$. This high-dimensional, learned representation is considered as a rich descriptor of the similarity between texts and images, effectively capturing their complex relationship.

$$[Q, K, V] = [W^Q\hat{\boldsymbol{y}}^g, W^K\boldsymbol{x}^l, W^V\boldsymbol{x}^l]; \tag{3}$$

$$\text{CrossAtt}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V; \tag{4}$$

$$SR_{\text{t2i}} = \text{Feedforward}\left(\text{CrossAtt}(Q, K, V)\right). \tag{5}$$

where $W^Q$, $W^K$, and $W^V$ are the weights for linear projection. $d_k$ is the feature dimension of $K$, which is equal to $D$. Following this, a linear projection is applied to map this high-dimensional representation to a low-dimensional space to derive the similarity matrix.

$$S_{\text{t2i}} = \text{Linear}(SR_{\text{t2i}}) \in \mathbb{R}^{T\times I}; \tag{6}$$

Optimization is performed using the InfoNCE loss.

$$\mathcal{L}_{\text{t2i}} = -\log\frac{e^{S_{\text{t2i}}^{i,i}}}{\sum_{k=1}^{I} e^{S_{\text{t2i}}^{i,k}}} - \log\frac{e^{S_{\text{t2i}}^{i,i}}}{\sum_{k=1}^{T} e^{S_{\text{t2i}}^{k,i}}}; \tag{7}$$

$S_{\text{t2i}}^{i,i}$ refers to a specific element in the similarity matrix, representing the similarity score between the $i$th image-text pair in the batch. Similarly, $S_{\text{t2i}}^{i,k}$ indicates the similarity score between the $i$th text and the $k$th image in the batch.

Similarly, for image-to-text ('i2t') alignment, global features from images are used as the query, and the word-based features of texts serve as the key and value. This process also yields a high-dimensional SimR, $SR_{\text{i2t}} \in \mathbb{R}^{I\times T\times D}$, for image-to-text, which is then projected using a linear layer to obtain the image-to-text similarity matrix for computation of InfoNCE loss.

$$\mathcal{L}_{\text{i2t}} = -\log\frac{e^{S_{\text{i2t}}^{i,i}}}{\sum_{k=1}^{T} e^{S_{\text{i2t}}^{i,k}}} - \log\frac{e^{S_{\text{i2t}}^{i,i}}}{\sum_{k=1}^{I} e^{S_{\text{i2t}}^{k,i}}}; \tag{8}$$

The final objective function for the CARZero model is the summation of $\mathcal{L}_{\text{t2i}}$ and $\mathcal{L}_{\text{i2t}}$, given by:

$$\mathcal{L} = \mathcal{L}_{\text{t2i}} + \mathcal{L}_{\text{i2t}}. \tag{9}$$

### 3.3. LLM-based Prompt Alignment

To align the prompts used during training and inference phases, we incorporate prompt templates into the training data using LLMs. Prompting instruction is utilized to generate the prompt template within the training data, with details shown in the supplementary. By leveraging the LLM's exceptional capability for semantic understanding, fixed prompt templates are introduced into the training data. The templates generated by the LLM are then merged with the original reports to create enhanced reports for training. During the inference phase, the prompt template, "There is [disease].", is employed for zero-shot classification. This strategy leverages the LLM's advanced capabilities in semantic comprehension to ensure that the training data is enriched with consistent and relevant prompt templates, thereby facilitating a more effective and aligned application during both training and inference stages.

## 4. Experiments

### 4.1. Dataset

**MIMIC-CXR [19]** In our experiments, we conducted model pretraining using the MIMIC-CXR dataset, a publicly available collection of chest radiographs paired with radiology text reports. The MIMIC-CXR dataset comprises 377,110 images corresponding to 227,835 radiographic studies conducted on 65,379 patients. Each radiographic study is accompanied by a radiology report and the corresponding chest X-ray image, which may be in either frontal or lateral views. The radiology report serves as a comprehensive summary provided by radiologists, encompassing various sections such as examination, indication, impression, findings, technique, and comparison. In our methodology, we selectively retain only the findings and impressions sections from these reports. Moreover, only frontal views of CXRs are used for CARZero training.

**Open-I [7]** Open-I contains 3,851 reports and 7,470 Chest X-ray images, which includes manual annotations for 18 different multi-label diseases. We evaluate the CARZero for zero-shot classification on Open-I.

**PadChest [2]** PadChest has 160,868 chest X-ray images labeled with 192 different diseases, which is a long-tailed distribution dataset. 39,053 (27%) samples are manually annotated by board-certified radiologists. For evaluation purposes, we only test on samples annotated by board-certified radiologists. Additionally, we select categories with fewer than 10 samples, totaling 20 classes, designated as **PadChest20**, to evaluate the performance of CARZero on rare diseases.

**ChestXray14 [37]** NIH ChestXray14 has 112,120 chest X-ray images with 14 disease labels from 30,805 unique patients. The official test set released by the NIH, comprising 22,433 images, are distinctively annotated by board-certified radiologists. For evaluation purposes, we only test on the official test set.

**CheXpert [17]** CheXpert has 224,316 CXRs collected from 65,240 patients. The official test set contains 500 patients annotated by a consensus of 5 board-certified radiologists [30]. We evaluate on 5 observations: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

**ChestXDet10 [27]** ChestX-Det10 is a subset of NIH ChestXray14, which is consisting of 3543 CXRs with box-level annotations provided by 3 board-certified radiologists of 10 diseases. The official test set contains 542 CXRs with 10 diseases and corresponding box-level annotations. We evaluate the zero-shot classification and grounding ability in the official test set.

### 4.2. Evaluation Metric

For the multi-label test dataset, we adopt Area under the ROC Curve (AUC), Matthews Correlation Coefficient (MCC), F1 score (F1), and Accuracy (ACC) as metrics for evaluating zero-shot classification tasks. For assessing zero-shot grounding tasks, we specifically utilize the Pointing Game [40] metric.

### 4.3. Implementation Details

In our experiments, ViT-B/16 is used as the image encoder, which utilizes M3AE [5] for pretraining on the MIMIC dataset. For the text encoder, BioBERT is fine-tuned using texts from both MIMIC and PadChest datasets. Given that PadChest reports are in Spanish, they are translated into English. The LLM named Spark[1] is employed to intelligently insert prompt templates into the training dataset.

The cross-attention module shares weights for 'i2t' and 't2i' alignments. The images are resized to a uniform shape of 224 × 224. Commonly used data augmentation including random horizontal flips, random affine transformations, and color jittering are adopted. For each report, we divide it into multiple sentences with a maximum length of 97, and one of the sentences is randomly selected at each training iteration. The optimizer used is Adam with a learning rate set to 5e-5. The code is based on the PyTorch framework. All experiments are conducted with an A800 GPU.

During the inference phase, when prompt alignment strategy is applied, we set the prompt template as "There is [disease]", denoted as $P_1$. Without prompt alignment, an empirically optimized prompt "A disease of [disease]", denoted as $P_2$, is found to yield the best performance.

[1] https://xinghuo.xfyun.cn/

| Method | Open-I | PadChest | PadChest20 | ChestXray14 | CheXpert | ChestXDet10 |
|---|---|---|---|---|---|---|
| MedCLIP [12] | 0.551 | 0.508 | 0.501 | 0.564 | 0.744 | 0.571 |
| BiomedCLIP [13] | 0.577 | 0.513 | 0.510 | 0.639 | 0.677 | 0.630 |
| GLoRIA [15] | 0.589 | 0.565 | 0.558 | 0.610 | 0.750 | 0.645 |
| BioViL [10] | 0.702 | 0.655 | 0.608 | 0.729 | 0.789 | 0.708 |
| CheXzero [33] | 0.726 | 0.648 | 0.644 | 0.712 | 0.889 | 0.640 |
| MedKLIP [38] | 0.759 | 0.629 | 0.688 | 0.726 | 0.879 | 0.713 |
| KAD [41] | 0.807 | 0.750 | 0.735 | 0.789 | 0.905 | 0.735 |
| CARZero | **0.838** | **0.810** | **0.837** | **0.811** | **0.923** | **0.796** |

Table 1. Comparative analysis of existing zero-shot classification approaches on five official multi-label CXR datasets evaluated by AUC.

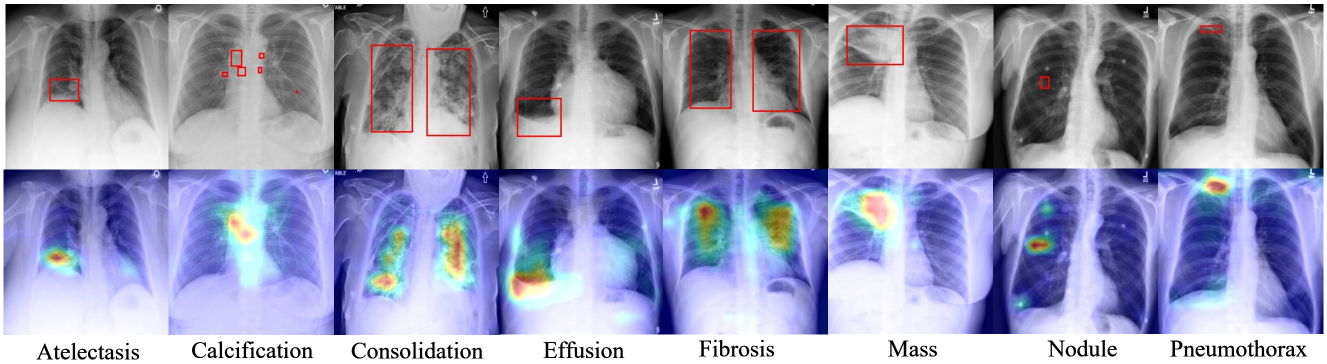

Figure 3. Visualization of attention map in CARZero on ChestXDet10. The red boxes indicate the corresponding ground truth of detection. Highlighted pixels represent higher activation weights correlating specific words with regions in the image.

| Method | ChestXray14 |
|---|---|
| GLoRIA [15] | 0.707 |
| MedKLIP [38] | 0.772 |
| MGCA [36] | 0.782 |
| KAD [41] | 0.787 |
| MRM [44] | 0.794 |
| CARZero | **0.811** |

Table 2. Comparison of the performance of existing methods fine-tuned on 1% data versus CARZero in zero-shot classification on ChestXray14 by AUC.

### 4.4. Comparison with State-of-the-art Methods

As shown in Table 1, we compare the performance of existing SOTA methods in CXR zero-shot classification on five officially released test sets. All test sets are manually annotated to ensure reliability. To ensure a fair comparison, all methods use their respective published models for inference. Among these, GLoRIA [15], MedKLIP [38], KAD [41], and our CARZero all utilize both global and local feature information for alignment. Ultimately, our proposed CARZero achieves better performance on the AUC metric across all test sets. Particularly for the PadChest dataset, which includes a long-tail distribution of 192 diseases, our method achieves an AUC of 0.810, indicating its good generalization performance on long-tail distribution data. Specifically, we achieve exceptional performance in rare disease zero-shot classification on PadChest20, attributable to our effective image-text alignment method. This demonstrates the significant potential of our approach for diagnosing rare diseases. Moreover, we compare zero-shot classification performance of our method with the results of existing works using 1% labeled data for fine-tuning. As shown in Table 2, our zero-shot classification performance even surpasses that of the existing methods fine-tuned on 1% data, demonstrating the strength of our approach. This might be due to cross-attention alignment used in CARZero generates high-level SimR to represent the relationships between medical images and reports, effectively measuring their complex relationships. Furthermore, the process of projecting SimR from high to low dimension using a learnable projection matrix fully exploits the associative information in SimR, thereby effectively aligning the feature spaces of images and texts. Additionally, our proposed prompt alignment strategy aligns prompts during the training and inference phases, eliminating the need to search for prompts, thereby enhancing the generalizability of CARZero to the prompts and improving zero-shot classification performance.

Additionally, we test the performance of CARZero in zero-shot grounding. We use the attention map from the cross-attention for grounding prediction. The aim of zero-shot grounding is to match prompt tokens with image tokens. As shown in Table 3, our CARZero achieves the best

| Method | Mean | ATE | CALC | CONS | EFF | EMPH | FIB | FX | MASS | NOD | PTX |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GLoRIA [15] | 0.367 | 0.479 | 0.053 | 0.737 | 0.528 | 0.667 | 0.366 | 0.013 | 0.533 | 0.156 | 0.143 |
| MedKLIP [38] | 0.481 | 0.625 | 0.132 | **0.837** | 0.675 | 0.734 | 0.305 | **0.224** | **0.733** | 0.312 | 0.229 |
| KAD [41] | 0.391 | **0.646** | 0.132 | 0.699 | 0.618 | 0.644 | 0.244 | 0.199 | 0.267 | **0.316** | 0.143 |
| CARZero | **0.543** | 0.604 | **0.184** | 0.824 | **0.782** | **0.846** | **0.561** | 0.184 | 0.700 | 0.286 | **0.457** |

Table 3. Comparison of various methods on the ChestXDet10 dataset for zero-shot grounding using the pointing game. The abbreviations ATE, CALC, CONS, EFF, EMPH, FIB, FX, MASS, NOD, and PTX correspond to Atelectasis, Calcification, Consolidation, Effusion, Emphysema, Fibrosis, Fracture, Mass, Nodule, and Pneumothorax, respectively.

performance. This is likely because our method directly uses SimR, generated by cross-attention alignment, to represent the similarity between images and texts. Therefore, the attention map in cross-attention effectively reflects the association between images and texts. The outstanding performance of our CARZero demonstrates that our method can effectively align the feature spaces of images and reports, and successfully match the lesion areas in the images with the prompts.

### 4.5. Visualization

As shown in Figure 3, we present the visualization results of CARZero. We perform linear interpolation on the attention map from the cross-attention alignment to obtain a pseudocolor image of the same size as the original image. At the same time, we display the names of the lesions and their corresponding locations. From Figure 3, it is evident that CARZero effectively captures the correlation between disease-related words and the corresponding lesion areas in the images, providing strong interpretability for our method. For small lesions, our method can also precisely detect them from the images. For example, as shown in Figure 3, for pneumothorax, CARZero accurately locates the corresponding lesion areas and localizes the full lesion. This may be attributed to the attention mechanism. In this paper, we propose a cross-attention alignment strategy, using the attention mechanism to directly align images and texts. This method can effectively capture text-related information from images based on a given prompt, which is demonstrated by Table 3 and Figure 3.

### 4.6. Ablation Study

**Ablation Study of Modules** To validate the effectiveness of the prompt alignment and cross-attention alignment proposed in this paper, we design experiments using the CLIP framework as the baseline, incorporating each of these modules separately. In scenarios without prompt alignment, we compare the performance of using different prompts $P_1$ and $P_2$. As shown in Table 4 (a vs. c and b vs. c), the prompt alignment strategy aligns the CARZero with $P_1$, achieving performance comparable to the empirically tuned one. This improvement is attributed to the model's alignment during the training phase with prompt $P_1$, leading to stronger generalization on $P_1$.

| # | PT | PA | CA | AUC | MCC | F1 | ACC |
|---|---|---|---|---|---|---|---|
| a | $P_1$ | | | 0.764 | 0.230 | 0.247 | 0.792 |
| b | $P_2$ | | | 0.796 | 0.255 | 0.271 | 0.821 |
| c | $P_1$ | ✓ | | 0.795 | **0.288** | **0.290** | 0.866 |
| d | $P_1$ | | ✓ | 0.781 | 0.217 | 0.230 | 0.816 |
| e | $P_2$ | | ✓ | 0.801 | 0.241 | 0.253 | 0.821 |
| f | $P_1$ | ✓ | ✓ | **0.810** | 0.257 | 0.270 | **0.867** |

Table 4. Results on ChestXray14 for the ablation study of various modules. Here, 'PT' denotes the prompt template used in the inference stage, 'PA' represents prompt alignment, and 'CA' stands for cross-attention alignment.

The comparison in Table 4 (a vs. d and b vs. e) demonstrates that introducing cross-attention alignment further enhances model performance. While CLIP utilizes global features from images and texts for alignment using cosine similarity, our method employs both local and global features for alignment. Additionally, high-level SimR is used to decipher the feature associations between images and texts, effectively aligning them and achieving superior performance. Table 4 (f) highlights the effectiveness of combining prompt alignment and cross-attention alignment for zero-shot classification. The integrated use of these techniques not only consolidates the strengths of each individual component but also helps the model to effectively generalize to diverse zero-shot scenarios.

**Ablation Study of Feature** In the context of cross-attention alignment, the query must utilize global features, while the key and value have three options: global only, local only, or a combination of both global and local features. We experiment with all three choices, as demonstrated in Table 5. The results show that using only global features yields the worst performance, while combining global and local features achieves optimal performance. The performance of using only local features is close to that of combining both global and local features. First, global features are an aggregation of local features, and this aggregation process may lead to the loss of detailed information. Hence, relying solely on global features is insufficient. Compared to global features, local features encompass richer semantic information, including details and positional information, thereby ensuring the completeness of feature information. Moreover, since global features are the aggregation of local fea-

| Global | Local | AUC | MCC | F1 | ACC |
|---|---|---|---|---|---|
| ✓ | | 0.799 | 0.248 | 0.264 | 0.856 |
| | ✓ | 0.810 | 0.257 | 0.270 | 0.867 |
| ✓ | ✓ | **0.810** | **0.259** | **0.276** | **0.880** |

Table 5. Results on ChestXray14 for key and value choice in cross-attention alignment.

| Method | AUC | MCC | F1 | ACC |
|---|---|---|---|---|
| $cos(x_y, y_x)$ | 0.804 | 0.225 | 0.228 | 0.813 |
| $cos(x_y, y) + cos(y_x, x)$ | 0.805 | 0.111 | 0.130 | 0.420 |
| Linear($SR$) | 0.810 | 0.259 | 0.276 | **0.880** |
| MLP($SR$) | **0.811** | **0.265** | **0.284** | 0.878 |

Table 6. Results on ChestXray14 for Processing of SimR. Here, $cos(\cdot)$ denotes cosine similarity, $x_y$ refers to $SR_{\text{i2t}}$ as the projection of image features onto text, and $y_x$ indicates $SR_{\text{t2i}}$ as the projection of text features onto image.

tures and match the global features of the query, integrating both global and local features can achieve the best performance, which aligns with intuition.

### 4.7. Processing SimR

To further explore the role of SimR in image-text alignment and validate the rationale behind using SimR in our study, we conduct the following investigations: (1) **Cosine similarity computation.** We consider SimR as a modality projection transformation. Following this, we can calculate the alignment of two modalities using cosine similarity, similar to the methods described in [16]. Consequently, we design two sets of experiments: one directly computes the cosine similarity between two projected features, while the other computes the cosine similarity between each projected feature and the original global features of images and texts, respectively. (2) **Direct projection to low-dimension similarity.** We continue to treat SimR as a high-level associative representation between images and texts. To more effectively extract the relationships between images and reports from SimR, we replace the simple linear layer with a multilayer perceptron (MLP).

As shown in Table 6, cosine similarity computation proves less effective than direct projection to low-dimension similarity. This may be caused by the fact that the features obtained through cross-attention alignment, as high-level SimR, can effectively represent the relationship between images and medical reports, eliminating the need for further similarity calculations, which is the core of our method. Moreover, cosine similarity, a non-parametric similarity metric, is less effective compared to our learnable similarity strategy, which better captures the complex relationships between images and medical reports, thus achieving superior performance. This indicates that using cosine similarity alone may not adequately measure the complex relationships between medical reports and images. Lastly, compared to simple linear projection, MLP employs more complex non-linear projection to reduce SimR from high to low dimensions. Experimental results demonstrate that MLP outperforms simple linear projection. This suggests that a simple linear layer might not be sufficient to fully extract the relationships between images and texts from SimR, which contains complex image-text relationships. Therefore, a complex MLP better leverages SimR's advantages.

## 5. Conclusion, Limitation and Impact

This paper proposes CARZero that achieves high-performing zero-shot classification. First, we propose a novel cross-attention alignment strategy, innovatively using the features generated by cross-attention as the SimR between images and reports. Subsequently, we use a linear layer to project SimR into a low-dimension similarity for aligning images and reports. Extensive experiments prove that the information contained in this SimR is highly effective for image-text alignment. Moreover, we propose a novel LLM-based prompt alignment strategy, integrating prompt templates into the training data to achieve training and inference prompt alignment. Finally, CARZero achieves SOTA performance on five publicly released official test sets, demonstrating its effectiveness.

**Limitations and Future Work** First, our method mainly focuses on zero-shot classification. The CARZero framework employs a cross-attention module as the core mechanism for aligning images and texts, which can be directly used as a classifier, thus offering certain advantages in fine-tuning. Therefore, future work could involve experiments in fine-tuning tasks. Secondly, our method primarily focuses on the complex relationship between medical images and reports, introducing a cross-attention alignment strategy to address it. We believe this method also has excellent generalizability to natural data. Hence, we plan to experiment with natural data in the future to further validate the effectiveness of the cross-attention alignment strategy.

**Impact** The CARZero proposed in this paper achieves state-of-the-art performance in CXR zero-shot classification tasks, including excellent performance on the long-tail dataset, PadChest, which includes 192 categories. This is significant for the diagnosis of rare diseases. In addition, ZSL greatly reduces the manpower cost of radiology experts and allows for low-cost acquisition of more training data, which is crucial for advancing chest AI diagnostics.

## 6. Acknowledgment

This work was supported by Natural Science Foundation of China under Grant 62271465 and Open Fund Project of Guangdong Academy of Medical Sciences, China (No. YKY-KF202206).

## References

[1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 1

[2] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020. 2, 5

[3] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020. 1

[4] Yihao Chen, Xianbiao Qi, Jianan Wang, and Lei Zhang. Disco-clip: A distributed contrastive loss for memory efficient clip training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22648–22657, 2023. 2

[5] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 5

[6] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. 1, 2

[7] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 5

[8] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Bannur et al. Learning to exploit temporal structure for biomedical vision-language processing. 2023. 6

[11] Varma et al. Villa: Fine-grained vision-language representation learning from real-world data. 2023. 2

[12] Wang et al. Medclip: Contrastive learning from unpaired medical images and text. *EMNLP*, 2022. 6

[13] Zhang et al. Large-scale domain-specific pretraining for biomedical vision-language processing. 2023. 6

[14] Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023. 2

[15] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena 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*, pages 3942–3951, 2021. 1, 2, 6, 7

[16] Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12054–12064, 2023. 3, 8

[17] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 5

[18] Mohammad Jamshidi, Ali Lalbakhsh, Jakub Talla, Zdeněk Peroutka, Farimah Hadjilooei, Pedram Lalbakhsh, Morteza Jamshidi, Luigi La Spada, Mirhamed Mirmozafari, Mojgan Dehghani, et al. Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. *Ieee Access*, 8:109581–109595, 2020. 1

[19] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 5

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2, 3

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 3

[22] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 3

[23] Junghwan Lee, Cong Liu, Junyoung Kim, Zhehuan Chen, Yingcheng Sun, James R Rogers, Wendy K Chung, and Chunhua Weng. Deep learning for rare disease: A scoping review. *Journal of Biomedical Informatics*, page 104227, 2022. 1

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2, 3

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 3

[26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[27] Jingyu Liu, Jie Lian, and Yizhou Yu. Chestx-det10: chest x-ray dataset on detection of thoracic abnormalities. *arXiv preprint arXiv:2006.10550*, 2020. 5

[28] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18051–18061, 2022. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[30] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexternal: Generalization of deep learning models for chest x-ray interpretation to photos of chest x-rays and external clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 125–132, 2021. 5

[31] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. *arXiv preprint arXiv:2312.03818*, 2023. 2

[32] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14223, 2023. 2, 3

[33] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022. 1, 2, 6

[34] Khoa A Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13(1):1–17, 2021. 1

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[36] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. 3, 6

[37] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017. 2, 5

[38] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01, 2023. 1, 2, 6, 7

[39] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19265–19274, 2023. 2

[40] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 5

[41] Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023. 1, 2, 6, 7

[42] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 1, 2

[43] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 2, 3

[44] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023. 6

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 3