Decomposing Disease Descriptions for Enhanced Pathology Detection: A Multi-Aspect Vision-Language Pre-training Framework

Vu Minh Hieu Phan[1], Yutong Xie[1], Yuankai Qi[2], Lingqiao Liu[1], Liyang Liu[1], Bowen Zhang[1],
Zhibin Liao[1], Qi Wu[1], Minh-Son To[3], Johan W. Verjans[1]

[1]Australian Institute for Machine Learning, The University of Adelaide; [2]Macquarie University; [3]Flinders University

[1]{vuminhhieu.phan,yutong.xie,lingqiao.liu,akide.liu,b.zhang,zhibin.liao,qi.wu01,
johan.verjans}@adelaide.edu.au, [2]yuankai.qi@mq.edu.au, [3]minhson.to@flinders.edu.au

Abstract

Medical vision language pre-training (VLP) has emerged as a frontier of research, enabling zero-shot pathological recognition by comparing the query image with the textual descriptions for each disease. Due to the complex semantics of biomedical texts, current methods struggle to align medical images with key pathological findings in unstructured reports. This leads to the misalignment with the target disease's textual representation. In this paper, we introduce a novel VLP framework designed to dissect disease descriptions into their fundamental aspects, leveraging prior knowledge about the visual manifestations of pathologies. This is achieved by consulting a large language model and medical experts. Integrating a Transformer module, our approach aligns an input image with the diverse elements of a disease, generating aspect-centric image representations. By consolidating the matches from each aspect, we improve the compatibility between an image and its associated disease. Additionally, capitalizing on the aspect-oriented representations, we present a dual-head Transformer tailored to process known and unknown diseases, optimizing the comprehensive detection efficacy. Conducting experiments on seven downstream datasets, ours improves the accuracy of recent methods by up to 8.56% and 17.26% for seen and unseen categories, respectively. Our code is released at https://github.com/HieuPhan33/MAVL.

1. Introduction

The advent of vision-language pre-training (VLP) methods, notably CLIP [30], has yielded impressive zero-shot and low-shot fine-tuning in object recognition [21, 22, 30, 42]. Pre-training on 400M image-text pairs, VLP models [21, 29, 30, 43] learn a strong semantic mapping between image and text spaces. During inference, they enable zero-shot recognition by computing the similarity between the query
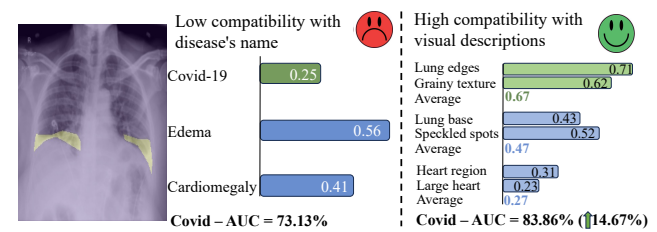

Figure 1. **Predictions of CheXzero [34] (Left), a strong CLIP-like model, and our multi-aspect matching model (Right).** `Edema` and `Covid-19` belong to the domain of lung diseases, while cardiomegaly is a heart disease. CheXzero [34] misaligns the image feature with the target `Covid-19`, while it over-aligns with `edema`. We leverage the medical knowledge base to decompose disease terms into distinct visual components, enhancing the image alignment with the representations of the target disease.

image and the text representation of each category. In the medical field, recent methods [3, 34, 35, 39] adopt CLIP and align images with their corresponding reports to enable zero-shot disease recognition, *i.e.*, classifying without further fine-tuning. As such, this strategy reduces the reliance on costly medical data annotation. Adding supervisory signals can enhance discriminative features, exemplified by a concurrent work [39] that extracts disease terms from reports and directly matches with the disease's definitions.

However, current VLP methods [3, 34, 35, 39] often produce low compatibility scores when aligning with the target disease's name, particularly for novel categories. Fig. 1 (Left) shows the misalignment between an image showing previously unseen disease, `Covid-19`, and its target disease name. This issue arises from various inherent challenges of VLP. *First*, medical imaging analysis is inherently fine-grained, which requires distinguishing between diseases that are visually similar, such as differentiating `Covid-19` from other lung diagnoses like `edema`. *Secondly*, biomedical texts contain complex clinical terminologies, and there is an imbalance between important entities
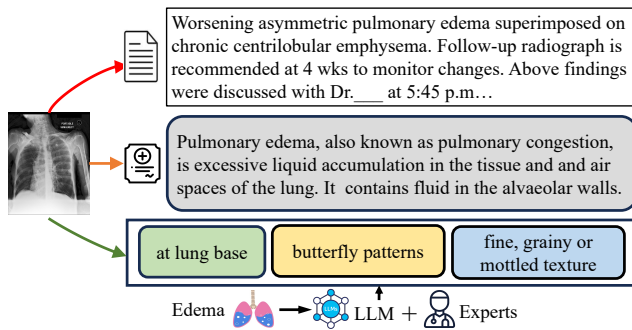
Figure 2. Illustrations of the three VLP paradigms: image-report matching [3, 6, 34] (red arrow), image-disease definition matching [39] (orange arrow), and our proposed fine-grained image-aspect matching (green arrow).

(*i.e.*, medical findings) and less relevant texts [6]. Fig. 2 shows sampled biomedical text inputs of current report matching [3, 6, 34] and disease definition matching methods [39]. Constrained by the scarcity of medical image reports [39], data-intensive models like CLIP struggle to align *fine-grained* visual features with *key* disease textual representations. *Thirdly*, current approaches represent text inputs as raw reports [6, 34] or disease descriptions [39], failing to capture generalized representations for novel medical findings. Consequently, the misalignment issue is exacerbated for diseases unmentioned in the pre-training dataset.

To resolve the image-text misalignment for both seen and unseen categories, this paper proposes to introduce a novel multi-aspect vision-language matching (MAVL) framework. We dissect disease descriptions into elemental aspects (*e.g.*, opacity level, shape) and harness pre-established knowledge about the diseases' visual profiles. We present a semi-automatic pipeline to extract descriptions of visual aspects of 75 medical findings using a large language model (LLM), GPT-4, and meticulously curate by consulting two medical experts. Our solution is motivated by two observations. First, structuring the text inputs into fine-grained visual aspects enhances alignments with the target diseases. Second, new diseases can be described by elemental visual aspects of base diseases, e.g., Covid-19 has a texture of edema (grainy texture) and typically is located in a similar position as pneumothorax (lung edges). Thus, multi-aspect decomposition improves the recognition of unseen diseases by linking visual appearances of any new diseases with the base visual knowledge. From Fig. 1 (Right), our model effectively aligns the image's features with the target `Covid-19` by leveraging the common aspects and extrapolating visual descriptions of seen lung diseases (e.g., edema) to those of a novel category.

Furthermore, we empirically observe that contrastive learning is non-optimal to capture discriminative features for differentiating fine-grained seen diseases. Capitalizing

on the aspect representations, our *secondary contribution* is to introduce a dual-head module dedicated to maximize the detection accuracy of seen and novel diseases. Each head is trained via a direct supervised loss and a contrastive loss. The supervised head maximizes the discriminative ability on base categories, while the contrastive head achieves performant zero-shot recognition of novel categories.

Our main contributions in this work are:

- A novel multi-aspect vision-language pre-training (MAVL) framework to improve the alignment between an image and textual representations of diseases, especially for unseen categories. To the best of our knowledge, this is the first study that exploits prior knowledge about pathological visual profiles to optimize fine-grained disease recognition.
- A dual-head Transformer that is trained via supervised loss and contrastive loss. Simple yet effective, we show that decoupling the learning signals is effective to maximize zero-shot recognition of unseen diseases, while protecting the discriminatory power toward base categories.
- State-of-the-art performance on zero-shot classification and grounding on both seen and novel disease categories, especially on rare diseases in a long-tailed distribution, across **seven** downstream datasets. Through fine-tuning, our model significantly surpasses previous methods, establishing a strong foundational model for both zero-shot and low-shot learning.

2. Related Works

General vision-language pre-training. Self-supervised VLP generates robust visual representations for various downstream tasks by exploiting joint vision-language supervision [21, 22, 29, 30, 42]. Current methods can be divided into two categories: a dual-stream [18, 20, 29, 30] with two encoders for each modality, or a single-stream [4, 12, 15, 23, 33] model, which enables deep V+L fusions. The supervised head in our network uses Transformer-based fusion layers to extract discriminative features.

Medical vision-language pre-training. Most medical VLP methods [3, 6, 14, 24, 35, 38, 46] adopt the two-stream approach, which applies image-report contrastive learning without using a deep fusion module. ConVIRT [46] proposes a bidirectional contrastive loss for image-report matching. GLoRIA [14] and LoVT [26] captures fine-grained alignments by optimizing the local contrastive loss between image regions and sentence-level tokens. Improving the biomedical text modelling, BioViL [6] adds clinical vocabulary and augments report-specific data. Subsequently, BioViL-T [3] extends the prior work by exploiting self-supervisory signals in a series of temporal images and reports. MGCA [35] conducts a pseudo-entity level alignment by assuming each disease prototype can be represented as a cluster in the feature space. A concurrent

work, MedKLIP [39], proposes to extract disease entities from reports and performs matching with the diseases' description. Yet, by matching reports or diseases' clinical definitions with complex terminologies, existing methods tend to misalign the image with the target disease's representations. Given the scarcity of medical data pairs and complex biomedical texts, *how to improve the image compatibility with the pathological text representation* is an open-ended question in VLP.

Textual prompt engineering. Several techniques investigate task-specific prompt engineering to adapt VLP for image classification [30, 32], object detection [32], and visual question answering [36]. In medical VLP, GloRIA [14] generates a set clinical-specific prompts describing possible sub-types, severities, and locations for each disease class. CheXzero [34] constructs a binary prompt to obtain a binary classification of a disease. MedKLIP [39] obtains clinical descriptions of medical findings from UMLS knowledge base [5]. Current medical VLP works use descriptions with domain-specific terminologies. In contrast, our proposed method leverages pathological's visual descriptions, guiding the model to effectively detect diseases in images.

Visual description for explainable VL models. Language concept bottleneck models (CBM) [25, 27, 40, 41, 44, 45] are an emerging field in explainable AI, which leverage visual descriptions of an object category to interpret VL model's decisions. Yet, they focus on *explaining* the pre-trained model in the general domain, while our framework leverages visual descriptions for *pre-training* medical models. Furthermore, extracting visual descriptions of diseases are non-trivial tasks compared to objects in the natural world. Ours is complementary to CBM studies, which potentially enables explainability by providing scores for each disease's appearance when classifying their presence.

3. Proposed Method

In this section, we first describe the VLP's reformulation, which enables supervised training (Sec. 3.1). Subsequently, we present our multi-aspect vision-language matching (MAVL) framework, which decomposes disease entities into a set of visual aspects leveraging the knowledge of medical experts and LLMs (Sec. 3.2). And thirdly, capitalizing on the aspect representations, Sec. 3.3 introduces our dual-head Transformer model, maximizing the complementary detection of both unseen and seen pathologies.

3.1. Problem setting

Given a set of B image-report pairs, $\mathcal{D} = \{(\mathcal{I}_1, \mathcal{R}_1), \ldots, (\mathcal{I}_B, \mathcal{R}_B)\}$, we aim to train a multimodal model to diagnose the presence of certain diseases in downstream datasets. At inference time, we can query the likelihood of a given disease category c, which is either seen or unseen during training, in the image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$.
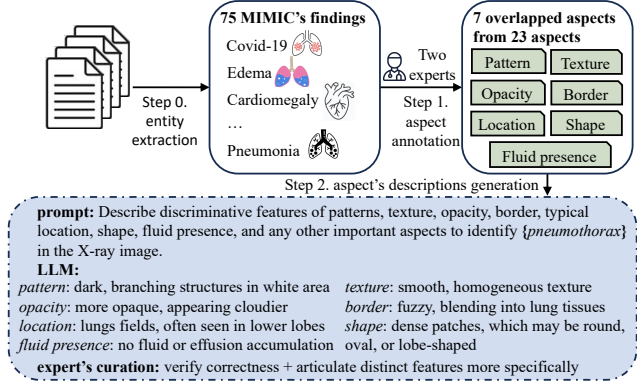

Figure 3. Pipeline to extract visual aspect's descriptions of diseases mentioned in the pre-training MIMIC dataset [19].

by prompting the textual disease description T_c:

$$\hat{p}, \hat{m} = \Phi_{\text{fusion}}(f_{\text{vision}}(\mathcal{I}), f_{\text{text}}(T_c)), \quad (1)$$

where f_{vision}, f_{text}, and Φ_{fusion} refer to vision, language and fusion modules. Here, $\hat{p}$ is the predicted likelihood of the disease; and $\hat{m} \in \mathbb{R}^{H \times W}$ denotes a heatmap with high activation on pixels indicating disease's visual presence. This heatmap is used for zero-shot visual grounding.

Reformulate VLP as a multi-label recognition. As clinical reports pose linguistic challenges (e.g., negation expressions, dense clinical terminologies), contrastively matching with raw reports struggles to capture discriminative features. Given the availability of well-established medical entity tagging models, such as RadGraph [17], we reformulate VLP as a supervised multi-label recognition. Inspired by [39], we adopt RadGraph [17] to extract entities (*i.e.*, diseases or any medical findings) in the reports:

$$\phi(\mathcal{R}_i) = \{\text{entity}_t, \text{location}_t, \text{exist}_t\}, t \in [0, t_i], \quad (2)$$

where t_i refers to the number of medical entities in the report, and location$_t$ is the location (*e.g.* lung, spine) where the entity$_t$ occurs in the image $\mathcal{I}_i$.

Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ and $L = \{l_1, \ldots, l_M\}$ denote the disease entities and locations extracted from all the report databases. Given an image $\mathcal{I}$, we reformulate VLP as a multi-label classification by directly predicting the presence of all diseases $\hat{p}_i \in \mathbb{R}^N$. Here, the ground-truth will be $Y_i = \{y_{i,1}, \ldots, y_{i,N}\}$, where $y_{i,j} = 1$ if the entity $c_j \in \mathcal{C}$ is indicated by the positive presence, *i.e.* exist$_j = 1$, in the report $\mathcal{R}_i$ and 0 otherwise.

3.2. Multi-aspect vision-language pre-training for enhanced pathology detection

To improve the image compatibility with the pathological representation, our paper proposes to dissect the disease description into a set of visual aspects by leveraging the

prior knowledge from medical experts, and a large language model (LLM). For each disease entity c_j, our proposed framework decomposes into a set of K visual aspects $S_j = \{s_{j1}, \ldots, s_{jK}\}$ by querying a knowledge base from medical professionals. We generate a description of each aspect entity Description(s_{jk}) using an LLM. To further differentiate diseases, we also add a fine-grained description of a disease $s_{j0} = $ Description(c_j), extracted from the UMLS knowledge base [5]. Our MAVL framework performs a fine-grained matching between an image $\mathcal{I}_i$ and every descriptive aspect S_j of the target disease c_j to detect its presence. An ablation study in Sec. 4 analyzes which aspect yields higher performance gain.

This paper presents a semi-automatic pipeline to decompose and describe disease's visual aspects from the 75 medical findings extracted by RadGraph [17]. Fig. 3 depicts our proposed pipeline leveraging medical experts' knowledge base and the LLM. To reduce annotation bias, two medical experts, a board-certificated cardiologist and a radiologist, independently annotated the visual aspects of 75 medical findings. Two experts independently annotate 23 visual aspects in total (see Suppl.), among which 7 aspects are in consensus, as shown in Fig. 3. We also add an extra 'other' aspect for the experts and the LLM to fill in any extra distinct features to recognize the target disease. In total, we have 8 visual aspects, plus a fine-grained disease description s_{j0}, which is extracted from the UMLS [5]

Secondly, we use GPT-4 to programmatically generate descriptions of the 8 annotated visual aspects. The pseudo-code is presented in the Supplementary. Finally, we consult two medical experts to curate the GPT-generated visual descriptions, which involves two steps. First, as LLMs can hallucinate contents, the experts correct the information. For example, GPT falsely describes 'pneumothorax' as 'more opaque, less transparent', which is then corrected as 'less opaque, more transparent'. Second, we articulate the distinct features of diseases to disambiguate fine-grained categories. For instance, GPT-generated descriptions can be vague, e.g., 'edema' has "hazy and patchy textures". We add more unique characteristics of edema: "fine, grainy, or mottled texture within the cloudy area, looking like small speckled spots".

3.3. Dual-head Transformer network.

This section presents our proposed aspect-oriented dual-head Transformer network, which takes the decomposed visual aspects of each disease and searches for their presence in the image. Given the image $\mathcal{I}$, the visual encoder f_{vision} extracts the visual features $\mathcal{V}$:

$$\mathcal{V} = f_{\text{vision}}(\mathcal{I}) \in \mathbb{R}^{h \times w \times d}, \tag{3}$$

where h, w, and d refer to the height, width and channel dimensions of the feature map. For a fair comparison with

a recent work [39], we adopt ResNet-50 for simplicity and computational efficiency. Yet, our framework is network-agnostic, and ViT backbones [11] can be applied.

Given an entity c_j, and its K visual aspects $S_j = \{s_{j1}, \ldots, s_{jK}\}$, the text encoder produces a set of d-dimension aspect embeddings $\mathbf{A}_j \in \mathbb{R}^{K \times d}$ for disease c_j:

$$\mathbf{A}_j = \{f_{\text{text}}(\text{Description}(s_{jk})) \quad | \quad k \in [0, \ldots, K]\}. \tag{4}$$

Capitalizing on the aspect representations, we develop a dual-head network with decoupled losses: a contrastive head for zero-shot recognition of unseen diseases, and a supervised head to maximize fine-grained classification of base diseases, as depicted in Fig. 4(A).

Contrastive head for zero-shot recognition of unseen diseases. This head performs fine-grained (FG) contrastive learning by matching an image representation $\mathbf{F}$ with each aspect representation $\mathbf{a}_{jk}$ of all disease c_j. The contrastive head first pools the image representation $\mathbf{f}$ from the visual feature map $\mathcal{V}$ using the attention pooling from CLIP [30]:

$$\mathbf{f} = \text{AttentionPool}(\mathcal{V}) \in \mathbb{R}^d. \tag{5}$$

Given an image feature $\mathbf{f}_i$, we compute a FG contrastive loss $\mathcal{L}_{\text{cl}}$ with each k-th aspect s_{jk} of every disease c_j:

$$\mathcal{L}_{\text{cl}} = \sum_{k=1}^{K} \log \sum_{j \in P(i)} \frac{\exp\left(\mathbf{f}_i^T \mathbf{a}_{jk}\right)}{\sum_t \exp\left(\mathbf{f}_i^T \mathbf{a}_{tk}\right)}. \tag{6}$$

Here, $P(i) = [c_j | y_{i,j} = 1]$ denotes the set of diseases c_j that are positively present in the report $\mathcal{R}_i$. During inferencing, our model programmatically extracts visual descriptions of new diseases using GPT-4, and compares its similarity with visual appearances of healthy versus pathological categories, as shown in Fig. 4(B).

Discussion. Diverging from previous report matching methods [3, 6, 34], ours matches with the textual representations of a common visual knowledge base (*e.g.*, disease's shape, opacity levels). This allows the model to link attributes of new diseases with this knowledge base and captures extensible visual representations.

Supervised head for zero-shot recognition of base diseases. Contrastive learning empirically struggles to extract discriminative features, leading to sub-optimal classification of fine-grained seen diseases. To optimize the fine-grained classification of a large set of base categories mentioned in the pre-training dataset, we develop supervised Transformer-based module Θ_{ground} to extract discriminative features by grounding the textual aspects in the image.

To extract features for a disease entity c_j, we view its descriptive visual aspects $\mathbf{A}_j \in \mathbb{R}^{K \times d}$ as a disease query set Q_j. Note, we omit the sample index i for simplicity. A Transformer-based fusion module Θ_{ground} grounds a query
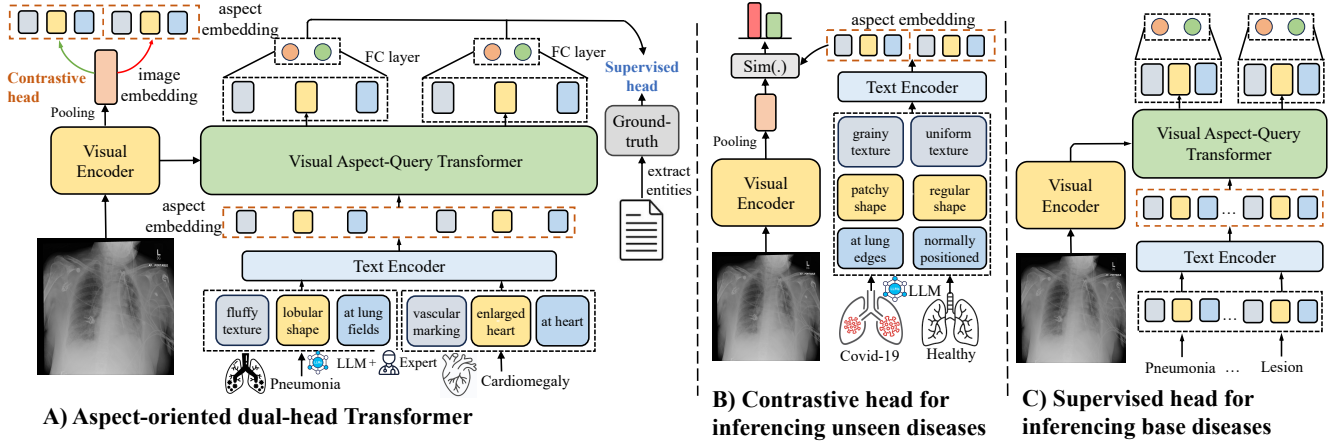
Figure 4. Multi-aspect vision-language pre-training framework (MAVL) decomposes diseases into a set of shared visual aspects using LLM and prior knowledge from two medical experts. An aspect-oriented dual-head Transformer (A) visually searches for the queried visual aspects in the image and maximizes detection accuracy of both unseen and seen diseases via two learning signals. The contrastive head (B) captures generalizable features and performs zero-shot classification of unseen diseases by comparing visual aspects between the target disease and the healthy category. The supervised head (C) captures discriminative features to classify fine-grained seen diseases.

set Q_j of textual visual aspects with image representation $\mathcal{V}$ using multiple Transformer Decoder layers [8]. The cross-attention views Q_j as Query, and $\mathcal{V}$ as Key and Value. The vision-text grounded features $\mathcal{F}_j \in \mathbb{R}^{K \times d}$ for disease c_j is formulated as

$$\mathcal{F}_j, \tilde{m}_j = \Theta_{\text{ground}}(Q_j, \mathcal{V}), \tag{7}$$

where $\tilde{m}_j$ is the heatmap for visual grounding. The feature maps are flattened into $\mathcal{F}_j \in \mathbb{R}^{Kd}$ and fed into a fully-connected layer parameterized by $W \in \mathbb{R}^{Kd \times 2}$ to predict the binary outcome of the disease $\hat{p}_j$. To detect all diseases, our fusion module takes a query set $Q = \{Q_1, \ldots, Q_N\}$ from all N medical findings that have been mentioned in the pre-training report sets, extracts the grounded features, and classifies their presence $\hat{p} \in \mathbb{R}^{N \times 2}$.

By reformulating VLP as a multi-label classification (Sec. 3.1), we directly train the visual aspect-query Transformer using supervision labels Y, thus improving zero-shot classification of seen diseases on downstream tasks. The supervised head is trained via a cross entropy loss $\mathcal{L}_{\text{sup}} = \text{CE}(\hat{p}, Y)$. Besides, the network predicts the location embedding $\tilde{e}_j$ of each disease from its features $\mathcal{F}_j$ using a simple projection layer. If the disease c_j is positively present and appears at location l_j, we adopt the location contrastive loss $\mathcal{L}_{\text{loc}}$ from [39] to match the predicted location embedding $\tilde{e}_j$ with the positive location embedding $e_j = f_{\text{text}}(\text{"It is located at } \{l_j\}\text{"})$. The network is trained end-to-end with the combined loss objectives:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cl}} + \beta \mathcal{L}_{\text{sup}} + \gamma \mathcal{L}_{\text{loc}}. \tag{8}$$

During inferencing, if the target category belongs to the set of N base categories $\mathcal{C}$, the model will use a supervised head for inferencing, as depicted in Fig. 4(C).

Discussion. While MedKLIP [39] also converts VLP into a supervised learning framework, we differ in two folds. *First*, they represent each entity c_j as a generic medical description with complex clinical terminologies (refer to Fig. 2). In contrast, our framework dissects the disease into a set of descriptive aspects, thus improving the image compatibility with the pathological representation. Importantly, we enable effective recognition of new diseases by translating them into elemental aspects, shared with base diseases. *Second*, MedKLIP [39] excludes the contrastive loss between image and text representations. While achieving high performance on base categories, their supervised model falters in unseen categories. In contrast, we decouple learning signals and introduce a dual-head network, tailored to process both known and unknown diseases.

4. Experimental Setting

4.1. Pre-training dataset

MIMIC-CXR v2 [19] consists of more than 227k studies comprising paired image-report data derived from 65,379 distinct patients who underwent scanning procedures. Each individual study may contain either one or two images, representing different scan perspectives, resulting in a cumulative dataset of 377,110 images.

4.2. Datasets for downstream tasks

ChestX-ray14 [37] comprises 112,120 frontal-view X-ray images from 30,805 individual patients, collected by NIH (National Institutes of Health) from 1992 to 2015. It includes labels for 14 prevalent diseases. We partition it into 0.8/0.1/0.1 for train/valid/test.

CheXpert [16] includes 224,316 chest X-ray images from

Table 1. Comparisons with SOTA image-text pre-training models under zero-shot classification for base diseases that have been seen during pre-training across 5 datasets.

Dataset	CheXpert [16]			ChestXray-14 [37]			PadChest-seen [7]			RSNA Pneumonia [31]			SIIM-ACR [1]		
Method	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑
ConVIRT [46]	52.10	35.61	57.43	53.15	12.38	57.88	63.72	14.56	73.47	79.21	55.67	75.08	64.25	42.87	53.42
GLoRIA [14]	54.84	37.86	60.70	55.92	14.20	59.47	64.09	14.83	73.86	70.37	48.19	70.54	54.71	40.39	47.15
BioViL [6]	60.01	42.10	66.13	57.82	15.64	61.33	60.35	10.63	70.48	84.12	54.59	74.43	70.28	46.45	68.22
BioViL-T [3]	70.93	47.21	69.96	60.43	17.29	62.12	65.78	15.37	77.52	86.03	62.56	80.04	75.56	60.18	73.72
CheXzero [34]	87.90	61.90	81.17	66.99	21.99	65.38	73.24	19.53	83.49	85.13	61.49	78.34	84.60	65.97	77.34
MedKLIP [39]	87.97	63.67	84.32	72.33	24.18	79.40	77.87	26.63	92.44	86.57	63.28	79.97	89.79	72.73	83.99
Ours	**90.13**	**65.47**	**86.44**	**73.57**	**26.25**	**82.77**	**78.79**	**28.48**	**92.56**	**86.91**	**63.41**	**82.42**	**92.04**	**77.95**	**87.14**

65,240 patients collected at Stanford Hospital. The official validation set comprises 200 chest radiographic studies annotated by three board-certified radiologists, while the official test set contains 500 studies annotated by five board-certified radiologists. Following the evaluation procedure from previous works [16, 34], we evaluate five observations in the official test set for the competition tasks.

PadChest comprises 160,868 chest X-ray images obtained from 67K patients reported at Hospital San Juan Hospital (Spain). They are annotated with 150+ distinct radiographic findings, including both unseen and seen classes during pre-training. We denote the 14 seen diseases as *PadChest-seen*, and unseen ones as *PadChest-unseen*. Among novel classes, we also evaluate rare diseases by selecting the top 20 diseases with the lowest number of samples, and the pathology recorded in the National Organization of Rare Disease (NORD) database[1], leading to 39 rare diseases in total. Approximately 27% of the data (39,053 examples) are annotated by radiologists, while the remainder are generated by a recurrent neural network. We only report the results on expert-annotated test samples.

RSNA Pneumonia [31] includes over 260,000 frontal-view chest X-rays with the annotated pneumonia masks collected by the Radiological Society of North America (RSNA). The dataset can be used for both pneumonia segmentation and classification tasks [6, 14, 39]. We split it into 0.6/0.2/0.2 for train/valid/test.

SIIM-ACR Pneumothorax [1] comprises over 12,000 frontal-view chest X-rays with pneumothorax masks, collected by the Society for Imaging Informatics in Medicine and the American College of Radiology (SIIM-ACR). We split it into 0.6/0.2/0.2 for train/valid/test.

COVIDx CXR-2 [28] and **COVID Rural** [9] provide benchmark dataset for Covid-19 detection. COVIDx CXR-2 [28] consists of 29,986 images from 16,648 COVID-19 patients, labeled for classification. Following [39], we split it into 0.7/0.2/0.1 for train/valid/test set. COVID Rural [9] consists of over 200 chest X-ray with annotated segmentation masks for COVID-19, which is used for seg-

mentation evaluation. We partition it into 0.6/0.2/0.2 for train/valid/test set.

4.3. Evaluation metrics

Following previous works [3, 6, 39], we adopt standard metrics for classification, including AUC scores, F1 scores, and accuracy, and those for segmentation, including Dice score, IoU scores, and pixel-wise accuracy. The metrics refer to the macro average on all the diseases present in the target dataset. All metrics are reported in percentage (%).

4.4. Implementation details

In *pre-training*, the text encoders are initialized with the weights of ClinicalBERT [2] and frozen. Only the visual encoder and the Transformer fusion module are trained during pre-training. All models are trained on 4 A100 GPUs. In *fine-tuning*, following previous works [3, 6, 39], we use ResNet50 [13] for classification, and ResUNet [10] for segmentation, which will be initialized with our pre-trained visual encoder. More details about hyper-parameter analysis can be found in the supplementary material.

5. Experimental results and discussion

This section presents the experimental results for zero-shot and fine-tuning setting. In the zero-shot case (Sec. 5.1), we benchmark our method against SOTA models for seen and unseen diseases on classification and segmentation tasks. In the fine-tuning case (Sec. 5.2), we evaluate the model's transferability on classification and segmentation tasks.

5.1. Zero-shot evaluation

Classification for seen diseases. Table 1 presents the zero-shot classification benchmark of *seen* diseases. Supervised methods, including MedKLIP and ours surpass other contrastive methods for base disease classification, showing that contrastive learning is non-optimal to capture discriminative features for fine-grained base disease classification. Compared to MedKLIP, our model consistently improves F1 scores by up to 8.56% across 5 datasets, collected

---
[1]https://rarediseases.org/rare-diseases/

Dataset	Covid-19 CXR-2			PadChest-unseen		PadChest-rare	
Method	AUC ↑	F1 ↑	ACC ↑	AUC ↑	ACC ↑	AUC ↑	ACC ↑
ConVIRT [46]	62.78	71.23	63.84	51.17	61.51	50.37	60.17
GLoRIA [14]	64.52	70.78	60.21	49.96	60.95	48.25	58.49
BioViL [6]	61.40	70.92	58.20	57.95	62.50	52.82	60.60
BioViL-T [3]	62.43	69.64	57.65	58.94	68.56	57.44	65.38
CheXzero [34]	73.13	76.13	71.45	66.70	81.19	65.08	81.17
MedKLIP [39]	76.28	76.54	71.96	60.31	76.69	59.75	77.84
MAVL - Sup	78.74	80.21	75.08	67.84	80.22	67.20	82.94
MAVL - Con	**83.86**	**81.73**	**78.07**	**70.42**	**84.00**	**70.06**	**84.64**

Table 2. Zero-shot classification results of unseen classes on the two datasets. Results on PadChest's rare diseases are reported. PadChest (from Spain) has a large distribution shift from the pre-training MIMIC data (from US), which is challenging for zero-shot classification. Following a prior work [34], we only consider AUC scores and accuracy. 'Sup' and 'Con' denote supervised and contrastive heads.

from various demographics, e.g., US population for CheXpert [16] and RSNA [31], and Spanish for PadChest [7]. By providing detailed visual aspects, our method excels at disambiguating fine-grained diseases, offering a robust foundational model against different imaging distributions.

Classification for unseen diseases. Table 2 presents the zero-shot classification benchmark of unseen diseases. We include the performance of both supervised and contrastive heads in our model for comparisons. We note three observations. *First*, both our heads consistently outperform previous methods on both datasets. Our supervised head drastically improves the AUC of a contrastive CheXzero, and a supervised MedKLIP by up to 7.67% and 12.49%. By decomposing the unstructured disease description into a common set of structured elemental aspects, our model effectively links new disease aspects with base diseases, improving feature representations of unseen diseases.

Second, contrastive models are generally more effective for novel disease recognition than the supervised counterparts. While the supervised MedKLIP struggles to classify many unseen diseases in PadChest, our contrastive head improves the AUC score of MedKLIP and CheXzero by 16.76% and 5.58%, respectively. This shows the effectiveness of our dual-head Transformer design, tailored to process both known and unknown diseases. *Third*, both of our heads surpass previous methods in classifying rare diseases. Our contrastive head significantly improves the AUC scores of MedKLIP and CheXzero by 17.26% and 7.65%. Leveraging prior knowledge about pathological visual profiles boosts the classification of the long-tailed distribution.
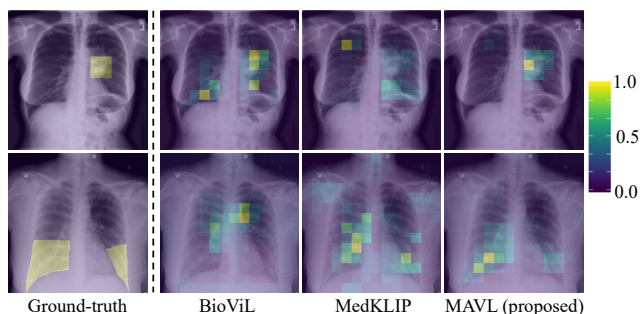

Figure 5. Visual grounding prediction scores of RSNA Pneumonia (Top) and Covid-19 (Bottom).

Visual grounding. In addition to the diagnosis, explainability is equally important in the medical domain. To this end, our model offers explainability by grounding the abnormal regions in the image. Table 3 presents visual grounding results of different models for both *seen* and *unseen* diseases: Pneumonia on RNSA dataset [31] and Covid-19 on Covid-19 Rural dataset [9]. Our model consistently surpasses previous methods on visual grounding. By dissecting disease entities into descriptive aspects, our model exhibits enhanced effectiveness in localizing abnormalities in images. In particular, we improve the Dice score of MedKLIP on a novel disease, Covid-19 by 5.34%. Decomposing the disease into a set of common visual aspects as in our MAVL improves the generalization of unseen diseases. Fig. 5 shows the visual grounding results of different methods on Pneumonia and Covid-19. Our model precisely localizes the abnormality, yielding lower false positive detection compared to other methods.

Table 3. Comparison with SOTA image-text pre-training models on zero-shot region grounding tasks on Pneumonia and Covid-19.

Dataset	RSNA Pneumonia [31]			Covid-19 Rural [9]		
Method	IoU ↑	Dice ↑	ACC ↑	IoU ↑	Dice ↑	ACC ↑
GLoRIA [14]	21.82	34.68	75.14	8.18	12.49	66.73
BioViL [6]	30.29	43.86	82.15	11.52	15.77	70.48
MedKLIP [39]	34.41	49.23	86.90	20.88	32.38	76.23
MAVL (ours)	**34.72**	**50.04**	**88.53**	**21.97**	**34.11**	**84.29**

5.2. Fine-tuning evaluation

This section provides fine-tuning evaluations by using the pre-trained models of different VLP methods to initialize weights and fine-tune downstream datasets. Table 4 presents fine-tuning results on 4 datasets with 1%/10%/100% data portion, which is consistent with previous works [6, 39]. Our model yields consistent improvements over previous works across all settings, especially under 1% data portion for both classification and segmentation. Notably, the proposed MAVL significantly outperforms the second-best method, MedKLIP, by 7.38% and 15.38% on SIIM Pneumothorax classification and RSNA-Pneumonia segmentation when fine-tuning on 1% data. On a fine-grained ChestXray-14 with 14 lung diseases, our

Table 4. Results of different VLP models under fine-tuning classification and segmentation with different data portions. AUC scores and Dice scores are respectively reported for the two tasks.

Task	Classification												Segmentation					
Dataset	Pneumonia [31]			Pneumothorax [1]			Covid-19 CXR-2 [28]			ChestXray-14 [37]			Pneumonia [31]			Covid-19 Rural [9]		
Data portion	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Scratch	68.94	83.31	87.12	53.11	76.18	87.48	85.11	93.65	98.86	45.88	56.27	67.03	45.29	58.42	69.75	14.09	25.97	37.83
ConVIRT [46]	78.86	85.42	87.64	72.39	80.41	91.67	90.30	97.74	99.70	57.23	72.53	79.13	56.48	63.94	71.87	16.97	30.79	42.71
GLoRIA [14]	79.13	85.59	87.83	75.85	86.20	91.89	92.74	97.18	99.54	58.94	72.87	79.92	58.13	67.71	72.06	16.12	31.20	43.85
BioViL [6]	80.27	86.04	88.29	70.29	79.45	88.05	92.39	98.39	99.68	60.83	72.94	80.16	60.25	68.72	72.18	17.65	37.75	47.34
MedKLIP [39]	82.11	87.14	88.58	85.24	89.91	93.02	95.58	98.77	99.77	62.09	74.02	80.90	62.36	70.24	73.88	18.58	39.28	48.65
MAVL (ours)	**86.09**	**87.90**	**88.94**	**91.53**	**93.00**	**94.48**	**97.18**	**99.15**	**99.90**	**68.65**	**80.50**	**86.22**	**71.95**	**73.51**	**76.97**	**24.51**	**40.71**	**50.25**

MAVL offers drastic performance gain even when fine-tuning with 100% data. Injecting prior knowledge about the disease's visual aspects during pre-training improves the model's transferability under low-shot learning, especially on fine-grained disease recognition.

5.3. Ablation study

Effectiveness of multi-aspect decomposition for VLP. This section answers two questions: (i) whether adding more concepts improves zero-shot performance, and (ii) which concept matching yields higher performance gain. Fig. 6(a) presents the performance when adding a single aspect. Matching with any single aspect boosts the accuracy of matching with a disease's definition as in [39]. Notably, the AUC score gain on unseen Covid CXR-2 is 4.02% when using the aspect 'texture'. More distinct aspects such as texture, patterns, and shapes yield higher gains. Fig. 6(b) shows the AUC scores when adding increasing numbers of visual aspects. Matching with higher numbers of aspects improves the zero-shot classification.
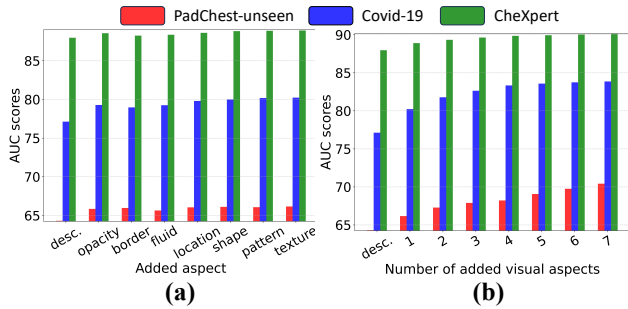

Figure 6. AUC scores of our method when (a) adding a single aspect, besides a disease clinical description and (b) when gradually adding higher numbers of aspects by their increasing importance order. The scores of using clinical description (desc.) of diseases are included as the baseline.

Effectiveness of a dual-head design. This section analyzes the effectiveness of our dual-head network, which decouples contrastive and supervised learning. We implement a single-head model using only either a contrastive branch as in [3, 34] or a supervised branch as in [39] *First*, Table 5

shows that using a single-head network design as in current literature compromises the performance of either seen and unseen classes. Capitalizing both branches in our dual-head design offers the highest F1 scores of 81.73% and 26.25% respectively on unseen Covid-19 and seen ChestXray-14. *Second*, adding supervised learning in our Dual-Con boosts the AUC scores of the single contrastive head by 14.63% on ChestXray-14. Similarly, adding CL in our Dual-Sup improves single supervised network. Integrating both learning signals as in our dual-head design complements the learning on both seen and unseen diseases.

Table 5. Zero-shot classification results between a single-head (Single) network, and our dual-head network. Con and Sup denote contrastive and supervised heads of our dual-head network.

Dataset		Covid-19 CXR-2 [9]			ChestXray-14 [37]		
Method	Loss	AUC ↑	F1 ↑	ACC ↑	AUC ↑	F1 ↑	ACC ↑
Single - Con	CL	81.65	80.75	76.98	62.08	18.27	62.53
Dual - Con	CL/CE	**83.86**	**81.73**	**78.07**	71.16	25.58	75.81
Single - Sup	CE	77.85	78.85	76.85	72.71	25.06	79.09
Dual - Sup	CL/CE	78.74	80.21	75.08	**73.57**	**26.25**	**82.77**

6. Conclusion

This study introduces a Multi-Aspect Vision-Language Pre-training (MAVL) framework, designed to intricately dissect pathological entities into distinct, visually descriptive aspects, thereby significantly enhancing pathology detection. Our MAVL framework incorporates a dual-head, aspect-oriented Transformer, which decouples learning signals to optimize performance across both seen and unseen diseases. This approach is particularly effective for novel diseases, achieving a remarkable 16.76% performance gain over existing methods. Moreover, MAVL's aspect-based matching leads to an impressive 17.26% AUC improvement for rare unseen diseases.

References

[1] Society for imaging informatics in medicine: Siim-acr pneumothorax segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation, 2019. 6, 8

[2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. 6

[3] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *CVPR*, pages 15016–15027, 2023. 1, 2, 4, 6, 7, 8

[4] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *NeurIPS*, pages 32897–32912, 2022. 2

[5] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004. 3, 4

[6] 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 *ECCV*, pages 1–21. Springer, 2022. 2, 4, 6, 7, 8

[7] 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. *Med. Imag. Analys.*, 66:101797, 2020. 6, 7

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 5

[9] Shivang Desai, Ahmad Baghal, Thidathip Wongsurawat, Piroon Jenjaroenpun, Thomas Powell, Shaymaa Al-Shukri, Kim Gates, Phillip Farmer, Michael Rutherford, Geri Blake, et al. Chest imaging representing a covid-19 positive rural us population. *Scientific Data*, 7(1):414, 2020. 6, 7, 8

[10] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, 2020. 6

[11] 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. 4

[12] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, pages 6616–6628, 2020. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] 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 *CVPR*, pages 3942–3951, 2021. 2, 3, 6, 7, 8

[15] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021. 2

[16] 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 *AAAI*, pages 590–597, 2019. 5, 6, 7

[17] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS*, 2021. 3, 4

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[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. 3, 5

[20] 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. In *NeurIPS*, pages 9694–9705, 2021. 2

[21] 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. 1, 2

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[23] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. pages 2592–2607, 2021. 2

[24] Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. 2023. 2

[25] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2023. 3

[26] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *ECCV*, pages 685–701. Springer, 2022. 2

[27] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023. 3

[28] Maya Pavlova, Naomi Terhljan, Audrey G Chung, Andy Zhao, Siddharth Surana, Hossein Aboutalebi, Hayden Gunraj, Ali Sabri, Amer Alaref, and Alexander Wong. Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Frontiers in Medicine*, 9:861680, 2022. 6, 8

[29] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. In *CVPR*, pages 6967–6977, 2023. 1, 2

[30] 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. 1, 2, 3, 4

[31] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 6, 7, 8

[32] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, pages 14274–14289, 2022. 3

[33] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Machine Learning*, 2020. 2

[34] 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, 3, 4, 6, 7, 8

[35] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *NeurIPS*, pages 33536–33549, 2022. 1, 2

[36] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 3

[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 *CVPR*, pages 2097–2106, 2017. 5, 6, 8

[38] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. 2022. 2

[39] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. MedKLIP: Medical knowledge enhanced language-image pre-training. In *ICCV*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[40] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *ICCV*, pages 3090–3100, 2023. 3

[41] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pages 19187–19197, 2023. 3

[42] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 1, 2

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Research*, 2022. 1

[44] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. 2023. 3

[45] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun. Do vision-language pretrained models learn composable primitive concepts? *Trans. Mach. Learn. Research*, 2023, 2022. 3

[46] 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. 2, 6, 7, 8