ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data

Maya Varma Jean-Benoit Delbrouck Sarah Hooper Akshay Chaudhari Curtis Langlotz
Stanford University
{mvarma2, jbdel, smhooper, akshaysc, langlotz}@stanford.edu

Abstract

*Vision-language models (VLMs), such as CLIP and ALIGN, are generally trained on datasets consisting of image-caption pairs obtained from the web. However, real-world multimodal datasets, such as healthcare data, are significantly more complex: each image (*e.g. *X-ray) is often paired with text (*e.g. *physician report) that describes many distinct attributes occurring in fine-grained regions of the image. We refer to these samples as exhibiting high pairwise complexity, since each image-text pair can be decomposed into a large number of region-attribute pairings. The extent to which VLMs can capture fine-grained relationships between image regions and textual attributes when trained on such data has not been previously evaluated. The first key contribution of this work is to demonstrate through systematic evaluations that as the pairwise complexity of the training dataset increases, standard VLMs struggle to learn region-attribute relationships, exhibiting performance degradations of up to 37% on retrieval tasks. In order to address this issue, we introduce ViLLA as our second key contribution. ViLLA, which is trained to capture fine-grained region-attribute relationships from complex datasets, involves two components: (a) a lightweight, self-supervised mapping model to decompose image-text samples into region-attribute pairs, and (b) a contrastive VLM to learn representations from generated region-attribute pairs. We demonstrate with experiments across four domains (synthetic, product, medical, and natural images) that ViLLA outperforms comparable VLMs on fine-grained reasoning tasks, such as zero-shot object detection (up to 3.6 AP50 points on COCO and 0.6 mAP points on LVIS) and retrieval (up to 14.2 R-Precision points)*[1].

1. Introduction

Vision-language models (VLMs), which jointly learn relationships between images and text, have been shown in recent years to be highly effective on a variety of classification, retrieval, and robustness tasks [28, 17, 27, 8, 42, 9]. VLMs are trained on large-scale datasets consisting of image-text pairs, where the text takes the form of a concise caption (*e.g.* alt-text) describing salient attributes in the image [28, 33, 34]. During training, VLMs generally model the relationship between a paired image-text sample as a *one-to-one* mapping: a single embedding of the entire image is contrastively aligned with a single embedding of the entire caption [28, 42, 17].

However, real-world multimodal datasets, such as those obtained from healthcare settings or product databases, consist of samples that are significantly more complex than standard image-caption pairs [19, 3, 1, 25]. In particular, real-world image-text samples include text that describes many distinct attributes occurring in fine-grained regions of the paired image. For example, medical images (*e.g.* X-rays) are accompanied by detailed text reports describing a variety of attributes (*e.g.* characteristics of organs, signs of disease) that map to specific regions of the image [19, 3, 4]. We refer to these samples as exhibiting high *pairwise complexity*, since each image-text pair can be decomposed into a large number of fine-grained region-attribute pairings. Figure 1 provides a visual depiction of pairwise complexity.

Models with knowledge of region-attribute relationships have been shown to exhibit numerous advantages, ranging from improved performance on fine-grained tasks (such as object detection) to improved subgroup robustness [43, 32, 29]. However, the extent to which standard one-to-one VLMs can learn these fine-grained relationships when trained on complex, real-world datasets is not currently well understood[2]. Prior work [43] has observed that standard one-to-one VLMs often struggle to learn region-attribute relationships[3], yet the specific effects of training dataset complexity on the ability of a VLM to capture these relationships have not been previously evaluated.

[1] Code: https://github.com/StanfordMIMI/villa

[2] As a motivating example, a VLM trained on X-rays should learn to link the region of the heart to the attribute "enlarged heart" in the report.

[3] When CLIP is applied to a region-level object detection task, performance is 40% lower than simply applying CLIP at the image level [43].

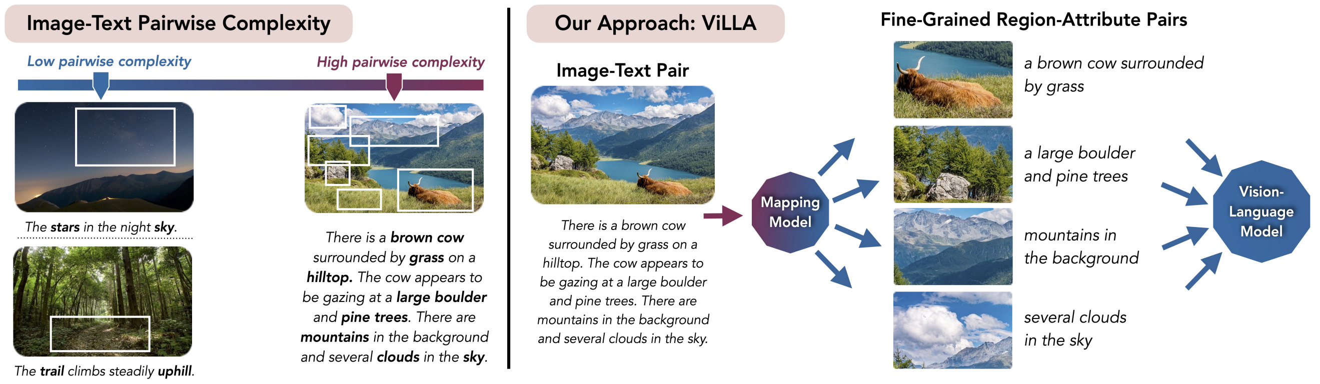


Image-Text Pairwise Complexity
Low pairwise complexity
High pairwise complexity
The stars in the night sky.
The trail climbs steadily uphill.
There is a brown cow surrounded by grass on a hilltop. The cow appears to be gazing at a large boulder and pine trees. There are mountains in the background and several clouds in the sky.
Our Approach: ViLLA
Image-Text Pair
There is a brown cow surrounded by grass on a hilltop. The cow appears to be gazing at a large boulder and pine trees. There are mountains in the background and several clouds in the sky.
Mapping Model
Fine-Grained Region-Attribute Pairs
a brown cow surrounded by grass
a large boulder and pine trees
mountains in the background
several clouds in the sky
Vision-Language Model

Figure 1. (Left) We provide examples of image-text samples with varying complexities. Examples with low pairwise complexity are from the CC3M dataset [34]. Textual attributes are in bold type, and corresponding image regions are marked with bounding boxes. (Right) We introduce ViLLA, a representation learning approach that captures fine-grained relationships between image regions and textual attributes.

As our first key contribution in this work, we conduct a systematic evaluation of the effects of training dataset complexity on the fine-grained reasoning ability of standard one-to-one VLMs. To this end, we introduce the *pairwise complexity score*, which measures the number of region-attribute pairings that can be composed from an image-text sample. Then, we create a synthetic dataset DOCMNIST, where the average pairwise complexity can be directly controlled by altering the number of region-attribute pairs per sample. We use DocMNIST to demonstrate that as the complexity of the training data increases, a standard one-to-one VLM demonstrates a performance drop of 36.9% on a text → region retrieval task and 20.5% on a region → text retrieval task, as measured by R-Precision scores. Our evaluation suggests that standard one-to-one VLMs are not effective at capturing region-attribute relationships when trained on datasets exhibiting high pairwise complexity.

Our findings demonstrate the need for a VLM that can learn accurate relationships between image regions and textual attributes when trained on complex multimodal datasets. We establish the following two desiderata for such a model. First, fine-grained relationships should be learned without the need for ground-truth region-attribute pairings, which are generally not provided in datasets and are expensive to manually label (particularly in specialized domains like healthcare). Second, representation learning should be performed using standard one-to-one VLMs. Despite their current inability to capture region-attribute relationships from complex datasets, one-to-one VLMs are highly effective [28], widely-used, and less computationally expensive than alternative fine-grained learning approaches [37].

As our second key contribution, we present **Vi**sion-**L**anguage **L**earning with **A**ttributes (ViLLA), a self-supervised multimodal representation learning approach that satisfies the above desiderata. Our key insight is that providing region-attribute pairs as training data to a standard one-to-one VLM helps improve the fine-grained reasoning ability of the model. ViLLA employs a two stage pipeline, as shown in Figure 1. The first stage involves training a lightweight *mapping model* to decompose image-text samples into region-attribute pairs. Here, we model the relationship between a paired image-text sample as a *many-to-many* mapping; given a set of many candidate image regions and a set of many textual attributes, we leverage self-supervision to learn a mapping between these sets. Then, the second stage involves training a standard one-to-one VLM on the generated region-attribute mappings.

We demonstrate with four multimodal training datasets across various domains (synthetic images, product data, medical images, and natural images) that ViLLA outperforms comparable one-to-one VLMs on fine-grained reasoning tasks, such as zero-shot object detection (up to 3.6 AP50 points on COCO and 0.6 mAP points on LVIS), text→region retrieval (up to 14.2 R-Precision points), and region→text retrieval (up to 7.8 R-Precision points). We show that these improvements are a result of our region-attribute mappings, which are up to 25.8 points more accurate than prior approaches.

Our contributions are summarized below:

- We demonstrate through a series of systematic evaluations that standard one-to-one VLMs struggle to learn relationships between image regions and textual attributes as training dataset complexity increases (leading to performance degradations of up to 37% on retrieval tasks). We also introduce DOCMNIST, a synthetic, customizable training dataset that we hope will be useful for future research on VLMs.
- We present ViLLA, a self-supervised multimodal representation learning approach that can effectively learn

fine-grained region-attribute relationships, particularly when training datasets exhibit high pairwise complexity. We demonstrate that our approach works effectively across a variety of real-world domains, outperforming comparable methods across tasks including zero-shot object detection (COCO and LVIS) and retrieval (CheXpert 5x200).

The rest of this paper is organized as follows. In Section 2, we discuss related work. We formally introduce the problem setting in Section 3, followed by our analysis with DOCMNIST in Section 4. In Section 5, we introduce ViLLA, and in Section 6, we present experimental results. Finally, we conclude in Section 7.

2. Related Work

Our work builds on several recent research directions. An extended discussion is provided in Appendix Section A.

One-to-One VLMs: Recent works have applied contrastive self-supervised learning methods to multimodal datasets [28, 17, 27, 42, 9]. During training, each image is pulled towards an associated caption and pushed away from dissimilar captions in the latent space.

Fine-Grained Representation Learning: Prior work has shown that one-to-one VLMs often struggle to capture fine-grained region-level information [26]. In particular, [43] shows that CLIP, a widely-used one-to-one VLM, achieves 60% accuracy on an image-level classification task yet only 19% on a region-level classification task with a similar number of classes. The authors attribute the performance drop to the fact that CLIP does not capture fine-grained relationships between image regions and textual attributes. Additionally, [20] demonstrates that CLIP often fails to understand subtle differences between images. Our work extends these lines of research by systematically evaluating the effect of training dataset complexity on the fine-grained reasoning ability of one-to-one VLMs.

Several prior approaches have been proposed for learning fine-grained region-level information from image-text datasets. One line of recent work leverages large quantities of human-labeled region-text pairs during training [22, 39, 35, 36]. However, obtaining human-annotated region-text pairs is expensive, time-consuming, and difficult to extend to other domains. In order to mitigate the need for human-annotated region-text pairs, [43] proposes RegionCLIP, which uses the pretrained CLIP model [28] in a zero-shot fashion to match candidate image regions with plausible textual attributes. However, this approach relies heavily on the CLIP model, which (a) has been shown to work poorly on localizing regions to text [43] and (b) cannot be accurately applied in a zero-shot fashion to out-of-domain data (such as medical images) [28]. Here, ViLLA aims to address these issues by introducing a specific training phase to learn region-attribute mappings, rather than directly using an off-the-shelf pretrained VLM model. Our work is also inspired by open vocabulary object detection [38, 23, 11, 29] and self-supervised patch-token alignment [37, 21, 15] methods.

Learning from Real-World Multimodal Data: Our work relates closely to prior studies that have developed VLMs for medical [42, 40, 2, 15] and product datasets [1, 5, 41]. We extend these lines of research by developing an approach that can effectively learn fine-grained signal from datasets with high pairwise complexity. We show that ViLLA works effectively across multiple real-world domains.

3. Preliminaries

In this section, we formally describe our problem setting. Datasets used for training VLMs can be expressed in the form $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^n$, where $x \in \mathcal{X}$ represents image inputs and $t \in \mathcal{T}$ represents text. $\mathcal{T}$ often takes the form of concise captions, which are simple phrases describing salient attributes in the associated image. Standard VLMs learn a *one-to-one* alignment between images and captions, which involves learning an embedding function $\psi_{img} : \mathcal{X} \rightarrow R^d$ that maps input images $\mathcal{X}$ to a latent space with dimension d. The function ψ_{img} is learned jointly with a function $\psi_{txt} : \mathcal{T} \rightarrow R^d$ that maps text data $\mathcal{T}$ to the same latent space. Current state-of-the-art approaches learn ψ_{img} and ψ_{txt} in a self-supervised manner by leveraging contrastive learning [28, 42, 17, 27].

In this work, we observe that real-world sources of multimodal data often include image-text pairs (x_i, t_i) that are complex, where the text refers to a large collection of attributes in various regions of the accompanying image. We can express this formally by decomposing each image into r_i regions, expressed as $x_i = \{x_i^0, x_i^1, \ldots, x_i^{r_i}\}$. Similarly, we decompose each textual description into a_i attributes, expressed as $t_i = \{t_i^0, t_i^1, \ldots, t_i^{a_i}\}$. We note that r_i does not necessarily equal a_i, since each textual attribute may manifest in one or more image regions. A set of fine-grained region-attribute pairs of size m_i can be obtained from the original sample pair (x_i, t_i); we refer to m_i as the image-text *pairwise complexity score*.

Given these definitions, we can quantitatively express the average pairwise complexity of a dataset as $s = \frac{1}{n}\sum_{i=1}^n m_i$, which characterizes the average number of region-attribute pairs per sample. Complex datasets have large values of s. Our goal in this work is to introduce an approach that can accurately learn fine-grained relationships between regions and attributes from training datasets with high pairwise complexity.

In the following section (Section 4), we first demonstrate that as the complexity of a dataset increases, standard one-to-one VLMs struggle to learn fine-grained representations. Then, in Section 5, we present an approach to improve vision-language representation learning on datasets exhibiting high pairwise complexity.

4. Understanding Dataset Complexity

In this section, we aim to better understand the challenges associated with using standard one-to-one VLMs to learn from datasets exhibiting high pairwise complexity. We introduce a synthetic training dataset in Section 4.1 with a variety of controllable dataset-level properties, such as the number of textual attributes. Then, in Section 4.2, we use this dataset to demonstrate that as complexity increases, representations learned using standard one-to-one VLMs degrade in fine-grained reasoning ability.

4.1. DocMNIST: A synthetic training dataset with controllable pairwise complexity

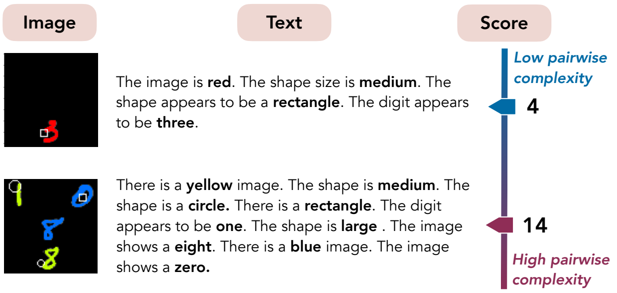
Here, we introduce our synthetic vision-language training dataset DocMNIST, which is an adaptation of the popular MNIST benchmark [7]. The purpose of DocMNIST is to enable systematic, controlled evaluations of learned vision-language representations as various dataset-level properties, such as the average pairwise complexity, are modified.

DocMNIST consists of images paired with textual descriptions (as shown in Figure 2). We set the size of each image to be $3 \times 84 \times 84$, subdivided into 9 square regions of size $3 \times 28 \times 28$. We define A as the set of possible attributes that can be assigned to each region. We consider the following attributes ($|A| = 20$) across 4 categories: digits (0-9), digit colors (purple, blue, green, yellow, red), shapes (rectangle, circle), and shape sizes (small, medium, large).

To create a DocMNIST training dataset, we first generate an image by randomly assigning attributes from set A to the nine image regions (*e.g.* a *red six*, which includes two attributes, may be assigned to the top left region). Additional constraints for this assignment process are discussed in further detail in Appendix Section B (*e.g.* all digits must have an associated color). An associated textual description is automatically generated for the image by filling attributes into pre-defined templates (*e.g.* "The image shows a six."). The average number of region-attribute pairs per sample is controlled by a user-specified variable c, which defines the average pairwise complexity of the dataset.

The final size of the training dataset is constrained by a pre-defined attribute budget b, which represents the total number of attributes across all images. We continue generating image-text pairs until the budget b is reached.

In summary, the following dataset-level variables can be controlled when generating a DocMNIST training dataset: the set of possible attributes A, the attribute budget b, and the average pairwise complexity score c.



The image is **red**. The shape size is **medium**. The shape appears to be a **rectangle**. The digit appears to be **three**.

There is a **yellow** image. The shape is **medium**. The shape is a **circle.** There is a **rectangle**. The digit appears to be **one**. The shape is **large** . The image shows a **eight**. There is a **blue** image. The image shows a **zero.**

Figure 2. Example image-text pairs from DocMNIST, where each textual attribute (bolded) appears in at least one region of the image. The pairwise complexity score measures the number of distinct region-attribute pairs.

4.2. Evaluating one-to-one VLMs with DocMNIST

In this section, we explore the following key question: *How does the complexity of the training dataset influence the fine-grained reasoning ability of a one-to-one VLM?* We measure fine-grained reasoning ability by determining if learned representations can be used to accurately localize textual attributes to corresponding image regions (text → region retrieval) and image regions to corresponding textual attributes (region → text retrieval).

First, we generate a set of DocMNIST training datasets that vary in complexity. We define A as the set of twenty attributes discussed in Section 4.1, and we then generate six versions of the training dataset with c ranging from 5.0 to 29.4. A training dataset with $c = 29.4$ has approximately 24 more region-attribute pairs per sample than a training dataset with $c = 5.0$. A fixed attribute budget of $b = 300K$ is maintained, which ensures that the total number of attributes in each of the six training datasets remains constant.

For each training dataset, we use a one-to-one VLM to contrastively learn alignments between images and the associated text. The image embedding function ψ_{img} is learned using a ResNet-50 model initialized with pretrained CLIP weights [28, 14]. The text embedding function ψ_{txt} is learned using a CLIP text encoder with frozen weights.

We measure the fine-grained reasoning ability of the resulting representations by using a held-out test set to measure text → region retrieval and region → text retrieval performance. Results are summarized in Figure 3. We observe that text → region retrieval performance drops by 36.9% (32.4 R-Precision points) and region → text retrieval performance drops by 20.5% (20.0 R-Precision points) as average pairwise complexity (c) increases from 5.0 to 29.4.

In summary, our analysis demonstrates that as the average pairwise complexity of a dataset increases, standard VLMs that assume a one-to-one relationship between images and text struggle to learn fine-grained representations. Additional analysis is provided in Appendix Section B.

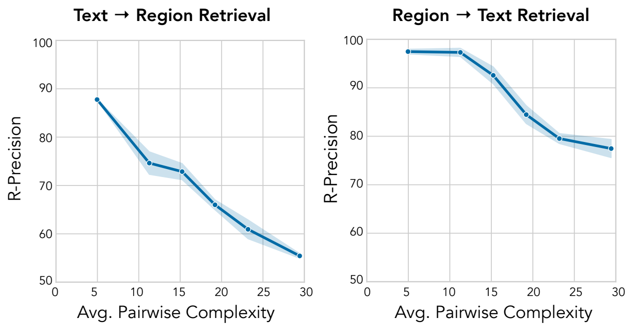


Figure 3. As the complexity of the dataset increases, representations generated using one-to-one VLMs demonstrate lower performance on text → region retrieval (left) and region → text retrieval (right). We report R-Precision scores.

5. Our Approach: ViLLA

Given our findings from Section 4, we establish the following desiderata as a set of ideal characteristics for a VLM:

- A VLM should have knowledge of fine-grained relationships between image regions and textual attributes. These relationships should be learned without supervision, since ground-truth region-attribute pairings are generally not labeled in datasets.
- Standard one-to-one VLMs should be used for learning representations. These models have been shown to be highly effective across a range of tasks [28], are widely used across numerous applications, and are less computationally expensive to train than previously-proposed fine-grained learning methods [37].

In order to satisfy both desiderata, we introduce ViLLA, a self-supervised multimodal representation learning approach that uses standard one-to-one VLMs to learn fine-grained region-attribute relationships from complex datasets. ViLLA follows a two-stage pipeline:

1. *Stage 1: Mapping Image Regions to Attributes* (Section 5.1): Given a set of candidate image regions and textual attributes, we first introduce a self-supervised *mapping model* to generate region-attribute pairs.
2. *Stage 2: Learning Vision-Language Representations* (Section 5.2): We then use the generated region-attribute mappings as training data for a standard one-to-one VLM. Our key insight is that providing accurate region-attribute pairs during training will improve the fine-grained reasoning ability of the VLM.

5.1. Mapping Image Regions to Textual Attributes

In this section, we discuss Stage 1 of ViLLA, which involves (a) decomposing image-text samples into image regions and textual attributes, (b) constructing a mapping model that learns attribute-specific projections for each input region, and (c) training the mapping model to pair regions and attributes in a self-supervised fashion. We begin with a training dataset $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^n$ consisting of n image-text pairs (x_i, t_i).

Decomposing Images and Text: We model the relationship between each paired image-text sample (x_i, t_i) as a *many-to-many* mapping. We first decompose each image x_i into r_i regions, expressed as $x_i = \{x_i^0, x_i^1, \dots, x_i^{r_i}\}$. Candidate regions can be identified in several different ways, such as by using off-the-shelf region proposal networks (RPNs), dividing images into equal-sized segments (such as quadrants), or using random bounding boxes. Similarly, we decompose each textual description t_i into a_i attributes, expressed as $t_i = \{t_i^0, t_i^1, \dots, t_i^{a_i}\}$. Attributes can also be extracted in multiple different ways, such as by leveraging structured labels provided with the dataset or using off-the-shelf entity extraction tools. Optimal approaches for region and attribute selection are dependent on the composition of the training data (*e.g*. whether domain-specific RPNs are available), and we discuss dataset-specific implementation details in Appendix Section C. Importantly, ViLLA operates under the assumption that each attribute in t_i occurs in at least one region in x_i.

Constructing the Mapping Model: Given image regions $\{x_i^0, x_i^1, \dots, x_i^{r_i}\}$ and textual attributes $\{t_i^0, t_i^1, \dots, t_i^{a_i}\}$, our goal is to learn mappings between regions and attributes. The mapping model consists of an image encoder to generate embeddings for image regions and a text encoder to generate embeddings for textual attributes.

In order to generate embeddings for image regions, we begin by initializing a CNN with pretrained weights from a domain-specific one-to-one VLM; we use ConVIRT weights [42] for our medical dataset and CLIP-RN50 weights [28] for all non-medical datasets. Given image x_i as input, we use RoIAlign [13, 43] to extract a representation for each region, resulting in region embeddings expressed in matrix form as $e_i \in \mathbb{R}^{r_i \times d}$. We use the variable d to represent the output embedding dimension, which is set to 1024 for the CLIP-based models and 768 for the ConVIRT-based model. Then, region embeddings e_i are provided as input to a set of p projection heads. Each projection head consists of a linear layer, a ReLU function, and a second linear layer. In most cases, we set $p = |A|$, where A represents the set of all attributes in the input dataset; as a result, each projection head is associated with a single textual attribute and yields an embedding that characterizes regions with respect to the attribute[4]. We represent the projection head for attribute k with the function P_k, which

[4] In some cases, particularly when $|A|$ is large, we set $p < |A|$ and assign each projection head to multiple textual attributes. Analysis on the selection of p is provided in Appendix Section C.

outputs final region embeddings of size $r_i \times d$.

Next, in order to generate embeddings for textual attributes, we insert each attribute into pre-defined prompt templates (*e.g.* "a photo of an [attribute]"). We extract representations from a pretrained text encoder, which is initialized with weights from sentence-BERT (SBERT) [30] for our medical dataset and CLIP for all non-medical datasets [28]. This yields an embedding $h_i \in \mathbb{R}^{a_i \times d}$ for textual description t_i. Additional implementation details are provided in Appendix Section C.

Training Procedure: Given region embeddings e_i, projection heads P, and attribute embeddings h_i, we now use the following procedure to train the mapping model. In order to ensure that the mapping model is lightweight, we freeze all parameters in the image and text encoder, leaving only the projection head parameters as trainable. Given a textual attribute $k \in t_i$, we use the notation $P_k(e_i) \in \mathbb{R}^{r_i \times d}$ to refer to the region embeddings resulting from the projection head corresponding to attribute k. Similarly, we use the notation $h_i^k \in \mathbb{R}^{1 \times d}$ to refer to an embedding of textual attribute k.

We train the mapping model with the following contrastive loss function. Let B represent the batch (with $|B|$ image-text pairs) and let $\sigma(a, b) = \exp(\max(\langle a, b \rangle / \tau))$.

$$L(x_i, t_i) = -\sum_{k \in t_i} \log \frac{\sigma(P_k(e_i), h_i^k)}{\sigma(P_k(e_i), h_i^k) + \sum_{j=1; k \notin t_j}^{|B|} \sigma(P_k(e_j), h_i^k)}$$

For sample (x_i, t_i) and attribute $k \in t_i$, this loss function encourages the maximum pairwise similarity between region embeddings $P_k(e_i)$ and the attribute embedding h_i^k to be high, since at least one region in x_i depicts attribute k. Simultaneously, for an image x_j where $k \notin t_j$, the maximum pairwise similarity between the region embeddings $P_k(e_j)$ and the attribute embedding h_i^k is encouraged to be low, since no regions in x_j depict attribute k.

5.2. Learning Vision-Language Representations

In this section, we discuss Stage 2 of ViLLA, which involves computing region-attribute pairs based on similarity scores assigned by the mapping model; then, a one-to-one VLM is trained on the generated pairs.

The mapping model from Section 5.1 can be used to assign attributes to regions as follows. For a sample (x_i, t_i) and textual attribute $k \in t_i$, we compute the pairwise dot product between $P_k(e_i)$ and h_i^k, resulting in a score vector $v \in \mathbb{R}^{r_i \times 1}$. We then assign k to all regions with a score greater than $max(v) - \epsilon$, where ϵ is a pre-defined threshold. This could assign zero or multiple attributes to a region[5].

Dataset	Domain	Regions	Attributes	s
DocMNIST	Synthetic	8.9	20	29.4
DeepFashion	Product	4	58	7.9
MIMIC-CXR	Medical	3	50	5.0
COCO	Natural	300	4.7k	6.8

Table 1. We use ViLLA to learn representations from 4 multimodal training datasets. Datasets vary in the average number of regions per image and total number of attributes. We also estimate the average pairwise complexity score (s) of each dataset.

We then augment the training dataset to include generated region-attribute pairs in addition to the original image-text samples. We use the augmented dataset to train a one-to-one VLM, which is optimized using a standard bidirectional contrastive loss function [42, 28].

6. Experiments

We evaluate our approach using four training datasets from various domains (synthetic images, product data, medical images, and natural images) and three fine-grained reasoning tasks (zero-shot object detection, text → region retrieval, region → text retrieval). Our experiments show that (1) our approach outperforms prior methods across all three tasks (Section 6.2) and (2) our region-attribute mappings are more accurate than prior approaches (Section 6.3). We provide extended results in Appendix Section D.

6.1. Datasets

We apply ViLLA to learn vision-language representations from four training datasets: DocMNIST (synthetic images), DeepFashion (product data), MIMIC-CXR (medical images), and COCO (natural images). Table 1 includes summary statistics, and further details are provided below:

DocMNIST: We create the synthetic DocMNIST dataset using the procedure described in Section 4. We generate a training dataset with an average pairwise complexity of 29.4.

DeepFashion [25, 18]: The DeepFashion-MultiModal dataset consists of 44k images extracted from clothing retail websites. Each image is accompanied with multi-sentence textual descriptions and structured labels (*e.g.* sleeve length, hats, etc.). During training, we use the 58 provided structured labels as our set of relevant attributes, and we divide each image lengthwise into 4 regions.

MIMIC-CXR [19, 10]: The MIMIC-CXR dataset consists of 377k chest X-ray images and associated physician reports obtained from the Beth Israel Deaconess Medical Center. We train an anatomy-specific RPN to divide each image into 3 regions: right lung, left lung, and heart. In order to create our attribute set, we use an off-the-shelf entity

[5] There are some settings, such as when regions are tight bounding boxes from an RPN, where each region is likely to capture exactly one attribute. In these cases, we invert this process and instead assign regions to attributes. Details are provided in Appendix Section C.

extractor (RadGraph) to identify the 50 entities that occur most frequently in the reports [16].

COCO [24]: The Microsoft COCO training dataset consists of 114k natural images. Each image is associated with five captions. In line with prior work, we extract 4.7k textual attributes (*e.g*. giraffe, man, bicycle, etc.) from the captions [43]. For each image, we use a pretrained RPN to extract 300 candidate regions [31].

6.2. Downstream Task Evaluations

We evaluate ViLLA on three fine-grained reasoning tasks: (1) zero-shot object detection, (2) text → region retrieval, and (3) region → text retrieval.

6.2.1 Zero-Shot Object Detection

Task: Given a bounding box containing an object, the zero-shot object detection task involves localizing the bounding box to an object category without performing any task-specific fine-tuning. We evaluate performance on the COCO validation dataset, which consists of 4.8k images annotated with ground-truth object bounding boxes corresponding to 65 classes [24]. We also evaluate on the LVIS validation set, which consists of 19k images across 1000 classes [12]. We set up both tasks as described in [43].

Evaluation: We train ViLLA on the COCO training dataset. We use a previously-developed evaluation framework [43] to compare our learned representations against three prior methods: OVR-CNN [38], CLIP [28], and RegionCLIP [43]. In line with prior work, we report AP50 scores on COCO across 17 novel categories, 48 base categories, and all 65 categories; we also report AP_{small} in order to characterize performance on small objects, a particularly challenging subgroup. On LVIS, we report AP across 337 rare categories (APr), 866 common and frequent categories (APc and APf), and all categories (mAP).

Results: Results are in Table 2. On COCO, our approach contributes to 8.1 points of improvement in AP50 over CLIP and 3.6 points of improvement over RegionCLIP. On the challenging subgroup of small objects, we note improvements of 8.6 points over CLIP and 4.6 points over RegionCLIP, suggesting that ViLLA is accurate even when regions are small. Similarly, on LVIS, we observe 2.8 points of improvement in mAP over CLIP and 0.6 points of improvement over RegionCLIP. Our results indicate that ViLLA is able to effectively reason over region-attribute relationships.

6.2.2 Text → Region Retrieval

Task: Given a textual query (*e.g*. "The person is wearing a hat"), the text → region retrieval task determines if we can retrieve image regions that capture the content of the query. We evaluate text → region retrieval on a held-out DocMNIST test set consisting of 5.9k regions as well as a held-out DeepFashion test set consisting of 7.9k regions. We obtain textual queries for both datasets by inserting attributes into pre-defined prompt templates. We consider 20 queries for DocMNIST and 46 queries for DeepFashion.

Evaluation: We train ViLLA on the DocMNIST and DeepFashion training datasets. We compare our representations against four baselines: (1) CLIP-ZS, which applies CLIP to this task in a zero-shot manner, (2) CLIP-FT-Img, which applies a CLIP model fine-tuned on image-text pairs, (3) CLIP-FT-Reg, which applies a CLIP model fine-tuned by aligning each region to the entire textual description, and (4) CLIP-ZS-Map, which first uses CLIP to generate region-attribute pairs in a zero-shot manner and then fine-tunes a CLIP model with the generated samples. We note that CLIP-ZS-Map is comparable to the RegionCLIP approach explored in Section 6.2.1. We report Precision@25, Precision@100, and R-Precision.

Results: Results are summarized in Table 3. On DocMNIST, our approach contributes to 13.4 points of improvement in P@100 and 14.2 points of improvement in R-Precision over CLIP-FT-Img, which is the next highest baseline. Similarly, on DeepFashion, our approach contributes to 8.0 points of improvement in P@100 and 4.2 points of improvement in R-Precision over CLIP-ZS-Map, which is the next highest baseline. In particular, we note that CLIP-ZS-Map, which is a baseline that emulates the design of RegionCLIP, performs particularly poorly on the DocMNIST dataset since the generated region-attribute mappings are limited by the performance of the original CLIP model. We also note that our performance improvements on DocMNIST are higher than DeepFashion; this is in line with the statistics presented in Table 1, which indicate that DocMNIST has a higher average pairwise complexity than DeepFashion. Our approach is most effective on datasets with high pairwise complexity scores.

6.2.3 Region → Text Retrieval

Task: Given an image region, the region → text retrieval task determines if we can identify the textual attributes depicted in the region. Again, we use 5.9k regions and 20 attributes for DocMNIST and 7.9k regions and 46 attributes for DeepFashion.

We additionally evaluate retrieval performance on the CheXpert 5x200 benchmark, which consists of 1000 chest X-rays across five disease categories [15]. Each disease label is converted into text using pre-defined prompts. Then, given a chest X-ray, the goal is to retrieve the textual phrase corresponding to the correct disease.

Evaluation: We train ViLLA on the DocMNIST and DeepFashion datasets, and we compare our representations

		COCO				LVIS			
Method	Pretraining Dataset	AP_{small}	$\text{AP50}_{\text{Novel}}$	$\text{AP50}_{\text{Base}}$	AP50_{All}	APr	APc	APf	mAP
OVR-CNN	COCO	-	46.7	43.7	44.5	-	-	-	-
CLIP	CLIP400M	43.3	58.6	58.2	58.3	40.3	41.7	43.6	42.2
RegionCLIP	CC3M	47.3	60.5	61.7	61.4	40.7	**43.5**	47.0	44.4
RegionCLIP	COCO	-	-	-	62.8	-	-	-	-
ViLLA (Ours)	COCO	**51.9**	**63.5**	**67.4**	**66.4**	**42.6**	42.4	**49.0**	**45.0**

Table 2. *Zero-shot object detection results.* We compare our approach with prior zero-shot object-detection methods. We report average precision (AP) scores on COCO (small, novel, base, and all object classes) and LVIS (rare, common, frequent, and all object classes).

	Text → Region						Region → Text	
	DocMNIST			DeepFashion			DocMNIST	DeepFashion
Method	P@25	P@100	R-Prec	P@25	P@100	R-Prec	R-Prec	R-Prec
CLIP-ZS	46.4	42.1	30.2	30.6	31.8	19.0	54.2	10.9
CLIP-FT-Img	84.4	78.2	55.2	51.6	58.0	34.4	78.7	27.3
CLIP-FT-Reg	47.0	38.8	29.0	33.6	36.3	23.8	38.4	21.0
CLIP-ZS-Map	55.6	47.6	35.2	53.9	59.4	35.3	49.9	28.8
ViLLA (Ours)	**91.6**	**91.6**	**69.4**	**56.5**	**67.4**	**39.5**	**86.5**	**32.3**

Table 3. *Retrieval on DocMNIST and DeepFashion:* We report text → region and region → text retrieval results on the DocMNIST and DeepFashion datasets. We compare our approach with several CLIP-based baselines, which ablate various components of our method.

Method	One-to-One	Accuracy
ConVIRT	✓	49.0
BioViL	✓	47.6
GLoRIA-Global Only	✗	52.1
GLoRIA-Local Only	✗	41.7
GLoRIA-Global+Local	✗	48.8
ViLLA (Ours)	✓	**55.9**

Table 4. *Retrieval on CheXpert 5x200*: We compare ViLLA with prior models, only two of which are one-to-one VLMs.

against the four baselines described in Section 6.2.2. Here, we report R-Precision scores.

For the CheXpert 5x200 task, we train ViLLA on the MIMIC-CXR dataset. At evaluation time, we consider each X-ray as a set of four regions - right lung, left lung, heart, full image - and perform retrieval by computing the maximum pairwise similarity with the text phrases. We compare with three prior methods: (1) ConVIRT [42, 6], (2) GLoRIA [15, 6], and (3) BioViL [2]. We report accuracy.

Results: Results on the DocMNIST and DeepFashion retrieval tasks are summarized in Table 3. On DocMNIST, our approach contributes to 7.8 points of improvement over CLIP-FT-Img, which achieves the next highest score. We also achieve 3.5 points of improvement on DeepFashion over the next highest score. Again, we note that the improvements on DocMNIST are larger than DeepFashion.

Table 4 shows results on the CheXpert 5x200 task. Our approach contributes to 3.8 points of improvement over GLoRIA, which achieves the next highest score. Two prior methods in Table 4 are one-to-one VLMs; we surpass these models by 6.9 and 8.3 points respectively.

6.3. Evaluating Region-Attribute Mappings

In this section, we demonstrate that the downstream performance improvements observed in Section 6.2 result from the improved quality of our region-attribute mappings. We evaluate the accuracy of region-attribute mappings on a test set associated with each of our pretraining datasets. In Table 5, we compare our approach to a random baseline as well as VLM-ZS, which refers to a one-to-one VLM (CLIP for DocMNIST, DeepFashion, and COCO and ConVirt for MIMIC) applied in a zero-shot manner. The VLM-ZS approach has been used in prior work to map regions to attributes [43]. Our results demonstrate that our approach outperforms baselines by up to 25.8 F1 points, suggesting that our mappings are higher quality than previous approaches and are contributing to improvements in representation quality. In Figure 4, we provide examples of region-attribute pairs generated by ViLLA on the COCO dataset.

7. Conclusion

In this work, we first demonstrate with evaluations on DocMNIST that as the complexity of the training dataset

Method	DocMNIST	DeepFashion	MIMIC	COCO
Random	27.4	32.0	35.4	38.4
VLM-ZS	42.6	54.8	61.8	72.9
ViLLA	**68.4**	**74.1**	**74.5**	**77.8**

Table 5. *Mapping quality*: We compare region-attribute pairs generated by ViLLA with previous approaches [43]. VLM-ZS refers to a pretrained one-to-one VLM (CLIP or ConVIRT) applied in a zero-shot manner [28, 42]. We report F1 scores.

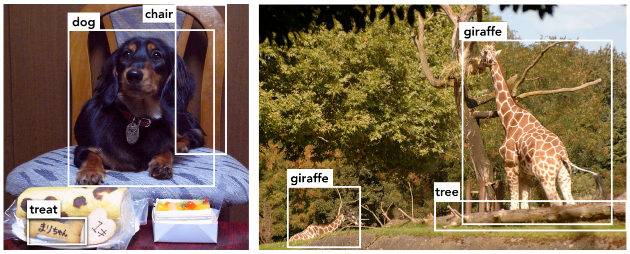


Figure 4. Examples of region-attribute pairs generated by ViLLA.

increases, standard one-to-one VLMs struggle to capture fine-grained region-attribute relationships. To address this issue, we introduce ViLLA. We demonstrate through evaluations with multiple real-world datasets that ViLLA can effectively capture region-attribute relationships, even when training datasets exhibit high pairwise complexity. Limitations of our work include: (a) our evaluations are currently limited to image-text datasets, and (b) our evaluation of region-attribute mapping accuracy is limited on datasets like MIMIC-CXR that do not include ground-truth annotations. Future directions include extending our approach to real-world datasets with other data modalities (such as audio, video, time-series) and conducting user studies to better evaluate the quality of region-attribute mappings on datasets that lack ground-truth annotations.

Acknowledgments

MV is supported by graduate fellowship awards from the Department of Defense (NDSEG) and the Knight-Hennessy Scholars program at Stanford University. SH is supported by the Fannie and John Hertz Foundation, the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1656518, and as a Texas Instruments Fellow under the Stanford Graduate Fellowship in Science and Engineering. AC is supported by NIH grants R01 AR077604, R01 EB002524, R01 AR079431, and P41 EB027060; NIH contracts 75N92020C00008 and 75N92020C00021. CL is supported by NIH grants R01 HL155410, R01 HL157235, by AHRQ grant R18HS026886, by the Gordon and Betty Moore Foundation, and by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021.

References

[1] Amit Alfassy, Assaf Arbelle, Oshri Halimi, Sivan Harary, Roei Herzig, Eli Schwartz, Rameswar Panda, Michele Dolfi, Christoph Auer, Kate Saenko, PeterW. J. Staar, Rogerio Feris, and Leonid Karlinsky. Feta: Towards specializing foundation models for expert task applications, 2022. 1, 3

[2] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 1–21, Cham, 2022. Springer Nature Switzerland. 3, 8

[3] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020. 1

[4] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation, 2022. 1

[5] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhaes, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1), Nov. 2022. 3

[6] Jean-Benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Jared A. Dunnmon, Pierre Chambon, Juan Manuel Zambrano, Akshay Chaudhari, and Curtis P. Langlotz. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, May 2022. 8

[7] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 4

[8] Sabri Eyuboglu, Maya Varma, Khaled Saab, Jean-Benoit Delbrouck, Christopher Lee-Messer, Jared Dunnmon, James Zou, and Christopher Ré. Domino: Discovering systematic errors with cross-modal embeddings. International Conference on Learning Representations (ICLR), 2022. 1

[9] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. Neural Information Processing Systems (NeurIps), 2022. 1, 3

[10] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and

PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220, Jun 2000. 6

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. 2022. 3

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 7

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2017. 5

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[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. 3, 7, 8

[16] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. 7

[17] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021. 1, 3

[18] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 6

[19] Alistair E W 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):1–8, 2019. 1, 6

[20] Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image retrieval from contextual descriptions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. 3

[21] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training, 2022. 3

[22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, abs/2112.03857, 2022. 3

[23] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *International Conference on Learning Representations (ICLR)*, 2023. 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 7

[25] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 6

[26] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023. 3

[27] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V. Le. Combined scaling for zero-shot transfer learning. *European Conference on Computer Vision*, abs/2111.10050, 2022. 1, 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 1, 2, 3, 4, 5, 6, 7, 9

[29] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection, 2022. 1, 3

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 6

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc., 2015. 7

[32] Khaled Saab, Sarah Hooper, Mayee Chen, Michael Zhang, Daniel Rubin, and Christopher Re. Reducing reliance on spurious features in medical image classification with spatial specificity. *Machine Learning for Healthcare*, 2022. 1

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine

Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 1

[34] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. 1, 2

[35] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection, 2020. 3

[36] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher, 2021. 3

[37] 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, 2022. 2, 3, 5

[38] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14388–14397, 2021. 3, 7

[39] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *International Conference on Machine Learning*, 2022. 3

[40] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing, 2023. 3

[41] Xujie Zhang, Yu Sha, Michael C. Kampffmeyer, Zhenyu Xie, Zequn Jie, Chengwen Huang, Jianqing Peng, and Xiaodan Liang. ARMANI: Part-level garment-text alignment for unified cross-modal fashion design. In *Proceedings of the 30th ACM International Conference on Multimedia*. ACM Conference on Multimedia, oct 2022. 3

[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. *Machine Learning for Healthcare*, abs/2010.00747, 2022. 1, 3, 5, 6, 8, 9

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16793–16803, June 2022. 1, 3, 5, 7, 8, 9