Prefix Conditioning Unifies Language and Label Supervision

Kuniaki Saito[1,2]*, Kihyuk Sohn[3], Xiang Zhang[2], Chun-Liang Li[2],
Chen-Yu Lee[2], Kate Saenko[1,4], Tomas Pfister[2]
{keisaito, saenko}@bu.edu
{kihyuks,fancyzhx,chunliang,chenyulee,tpfister}@google.com
[1]Boston University, [2]Google Cloud AI Research, [3]Google Research, [4]MIT-IBM Watson AI Lab

Abstract

Pretraining visual models on web-scale image-caption datasets has recently emerged as a powerful alternative to traditional pretraining on image classification data. Image-caption datasets are more "open-domain", containing broader scene types and vocabulary words, and result in models that have strong performance in few- and zero-shot recognition tasks. However large-scale classification datasets can provide fine-grained categories with a balanced label distribution. In this work, we study a pretraining strategy that uses both classification and caption datasets to unite their complementary benefits. First, we show that naively unifying the datasets results in sub-optimal performance in downstream zero-shot recognition tasks, as the model is affected by dataset bias: the coverage of image domains and vocabulary words is different in each dataset. We address this problem with novel Prefix Conditioning, a simple yet effective method that helps disentangle dataset biases from visual concepts. This is done by introducing prefix tokens that inform the language encoder of the input data type (e.g., classification vs caption) at training time. Our approach allows the language encoder to learn from both datasets while also tailoring feature extraction to each dataset. Prefix conditioning is generic and can be easily integrated into existing VL pretraining objectives, such as CLIP or UniCL. In experiments, we show that it improves zero-shot image recognition and robustness to image-level distribution shift.

1. Introduction

Supervised classification datasets (e.g., ImageNet [7]) have traditionally been used to pretrain image representations for use in downstream tasks. However, web-scale image-caption datasets have recently emerged as a powerful pretraining alternative [13, 20, 31]. Such datasets are more "open-domain", containing a wider variety of scene types and vocabularies than traditional classification datasets, which are biased towards specific categories in their fixed label sets. Consequently, models trained on web-scale image-caption datasets have shown stronger generalization in novel tasks [4, 31] and demonstrated remarkable performance on few and zero-shot image classification tasks [31]. Nevertheless, classification datasets are still useful for pre-training as they have a more balanced coverage of categories, including rare and fine-grained categories, and a better focus on the labeled objects in each image.

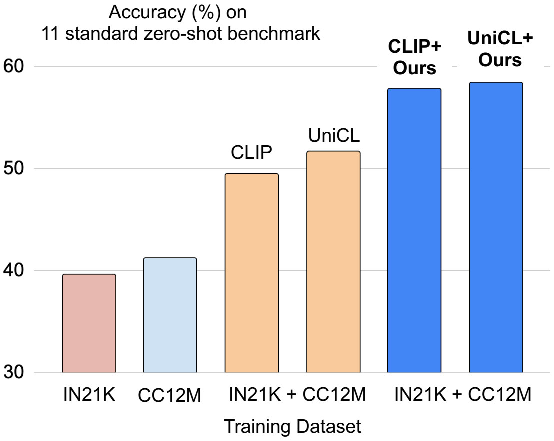


Figure 1. We propose *Prefix Conditioning* to unify image-caption (e.g., CC12M [5]) and image classification datasets (e.g., ImageNet21K (IN21K) [7]) for training better zero-shot models. Prefix conditioning improves zero-shot recognition performance by more than 6% on average when training on ImageNet21K and CC12M.

Recent works [43, 45] therefore propose to combine caption and classification datasets for pre-training. [43] convert classification labels to "label-prompts" by inserting the label into a template sentence, e.g., "a photo of a <label>."[1]

[*] Work done during internship at Google Cloud AI Research.

[1] We use the term *prompt* to indicate a template sentence filled with a class name.

Although training on the caption and label-prompt data achieves promising results, it does not fully resolve distribution differences between the open-domain caption data and the classification data. In particular, it produces a language embedding entangled with the classification dataset "bias". We note that classification datasets tend to be biased in at least two ways: 1) the images mostly contain single objects from restricted domains, and 2) the vocabulary is limited and lacks the linguistic flexibility required for zero-shot learning. Therefore, the class embedding of "a photo of a dog" optimized for ImageNet may really mean *a photo of a dog from ImageNet* instead, which is biased to ImageNet and does not generalize well to other datasets. We empirically show that such dataset biases negatively affect unified pretraining by reducing the generalization of learned representations and thus jeopardizing zero-shot performance.

To recognize diverse concepts in the open domain, the language model needs to disentangle the dataset bias from the visual concepts and extract language embeddings generalizable to the open domain, e.g., the language embedding representing *a photo of a dog from an open-domain dataset, such as image-caption dataset*, instead of *a photo of a dog from ImageNet*. Given this intuition, we propose to learn dataset-specific language embeddings, while sharing knowledge from both datasets during training. We achieve this by a simple yet effective approach we call *Prefix Conditioning*. The idea is to learn a dataset-specific text token (*prefix*) for each dataset so that the bias of the dataset can be absorbed into this token, and in return the remaining text tokens can focus on learning visual concepts. Specifically, we prepend a different token for each dataset (e.g., image classification or caption dataset) to the text input token sequence during pre-training.

The idea is in part inspired by prefix or prompt tuning [18, 21, 46], which showed that learnable tokens prepended to the input token sequences of the pre-trained language models are able to learn task-specific knowledge and thus can be used to solve downstream tasks by combining the knowledge of pre-trained large language models and task-specific prefix tokens. Our approach differs from prompt tuning in two ways: 1) the proposed prefix conditioning is designed to unify image-caption and classification datasets by disentangling the dataset bias, which is a unique distinction to prompt-tuning works, 2) our approach is applied for VL *pre-training* while the standard prompt tuning is used in fine-tuning.

In experiments, the proposed simple technique achieves superior performance on zero-shot evaluation if we use the prefix of the caption dataset to get the language embedding at test time as shown in Fig. 1. Meanwhile, inserting the prefix of the classification dataset leads to better performance on classification data. We also observe a drastic performance improvement when combining our prefix conditioning with the UniCL [43] objective because of their complementarity. Our contributions are summarized as follows:

- We propose novel Prefix Conditioning *at pre-training time* to unify image-label and image-caption supervision. It is the first mechanism to use prefixes to condition the source of the dataset during vision language contrastive pre-training, rather than post pre-training.
- This simple approach improves zero-shot recognition performance by more than 6% on average in experiments on ImageNet21K [7] and CC12M [5].
- Our comprehensive ablation study shows that prefix conditioning enables the model to switch its approach to extracting language features, e.g., attend to different words.

2. Related Work

Vision-Language Contrastive Learning. Zero-shot recognition is conventionally solved by learning the relationship between visual representations and word embeddings of the class names [1, 9, 12, 26, 38, 40, 41]. Vision-language contrastive learning models, such as CLIP [31], pre-train a model with a large-scale image-caption data (400M) and achieve a remarkable improvement in zero-shot recognition. ALIGN [13] demonstrated the effect of scaling up the size of image-caption data. Various techniques have been proposed to improve the data efficiency given a relatively small amount of image-caption data (order of 10M). ALBEF [20] employs model distillation and masked language modeling. DeCLIP [22], SLIP [29] and TCL [42] harness self-supervised contrastive learning. FILIP [44] uses token-to-token contrastive learning rather than the global contrastive learning used in CLIP. BLIP [19] generates pseudo captions to diversify the language modality for each image. Unlike these works that handle only caption-style supervision, we focus on the use of label supervision in vision-language pre-training. Our approach brings orthogonal improvement to the aforementioned works as they seek to improve training on image-caption data.

UniCL [43] and K-Lite [34] unite the image-caption and image-label supervision by converting labels into text with pre-defined template sentences. UniCL leverages a supervised contrastive loss [15] for image-label pairs. K-Lite [34] utilizes external knowledge from WordNet [28] and Wikitionary [27]. The input noun is augmented with the class hierarchy and definition to enrich the supervision. Our method is complementary to these approaches since both UniCL and K-Lite do not consider the domain shift between datasets. In experiments, we observe a significant performance boost when UniCL is combined with the prefix conditioning.

Learning with Prompts. Prompt tuning is a popular technique to adapt a large language model to a specific task with few training data and low computational

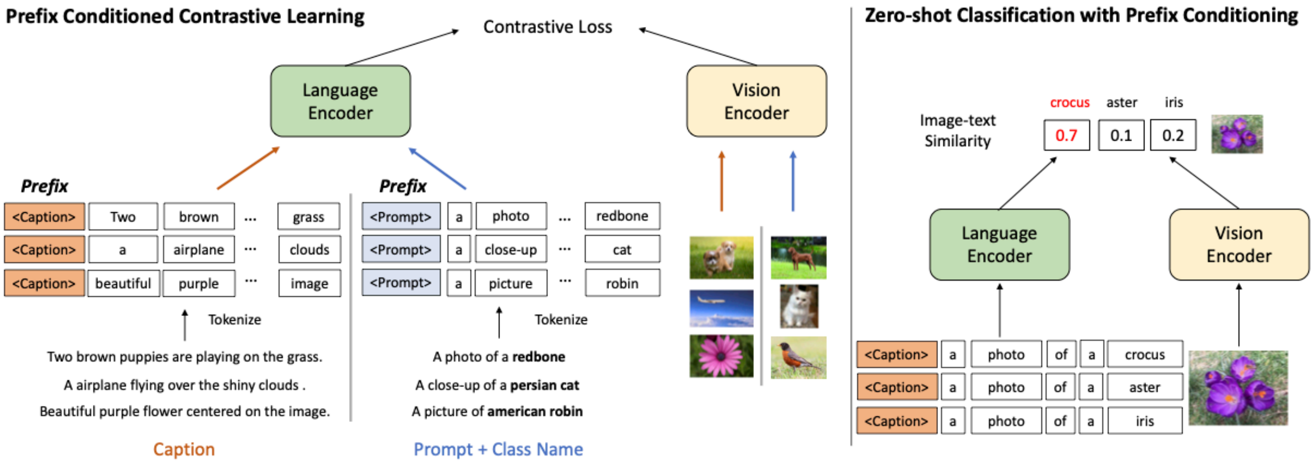


Figure 2. **Left**: Prefix conditioning at training time. Dataset-specific token is added to the input tokens with a contrastive learning objective applied. **Right**: Prefix conditioning at test time. Given a class name, we construct a class prompt with pre-defined templates and add a token used to condition real caption during training considering that image-caption dataset covers much wider range of image domains and vocabulary words than image classification dataset.

cost [10,18,21,23,30]. To avoid tuning all parameters of the model and using hand-crafted prompts, prefix embeddings are added to the training input and are the only parameters optimized during fine-tuning. The prefix embedding can be viewed as the knowledge of the downstream task. In this paper, since the target task is the zero-shot classification, the bias of the language embedding needs to be from the dataset covering a wide range of domains rather than a specific domain. Therefore, we choose to use the prefix embedding learned for image-caption dataset during test time. This technique is also effective in adapting a pre-trained vision-language model [46, 47] to few-shot classification by tuning the prompts of the language encoder to adapt to a downstream task. Additionally, prompt-tuning is effective in adapting a pre-trained vision model to a target task [14]. While these works aim to tailor a large pre-trained model to a specific downstream task with a small amount of data or low computational cost, our goal here is to condition a model with the prefix during the pre-training stage by distinguishing between the image label and image caption data. This allows a model to effectively share the knowledge obtained from two different types of data sources.

Dataset bias in image recognition. A large-scale image recognition dataset such as ImageNet [7] is known to be biased towards a specific image domain. Therefore, a model trained on such a dataset shows vulnerability to the distribution shift, e.g., shift in object pose [3] and style of the images [37]. Nevertheless, [16, 39] show that adapting only a linear layer on the pre-trained models can improve performance on the downstream tasks with distribution shifts. This indicates the importance of having a good classifier on top of image encoders, such as linear classifiers generated by language encoders with preconditioning in our work. [8] propose a method for domain generalization. They condition image recognition models with the domain embedding, which discriminates the input image domains, and demonstrate the importance of the domain-specific image classifier. Our prefix conditioning can be seen as an attempt to de-bias the linear classifier to obtain a domain-specific classifier and adapt it from the classification to the captioning domain. Also, [2, 17] approach the dataset bias in image classification by de-biasing image representations. By contrast, we tackle the problem in the framework of vision-language learning, disentangle the dataset bias in the language embedding and utilize the classifier obtained by the caption domain. We note that while captioning datasets can also have data biases, they tend to be more open-domain than existing classification datasets.

3. Method

In this section, we introduce the Prefix Conditioning technique for pretraining a deep learning model on both image-caption and image-label (classification) data. In Sec. 3.1, we discuss our problem setting and the background of contrastive learning with image-caption data. In Sec. 3.2, we explain the details of our training approach, and in Sec. 3.3 our inference procedure.

3.1. Preliminaries

Setup. Suppose we have access to two datasets: (i) an image label dataset $\mathcal{S}_L = \{(\boldsymbol{x}_n, \boldsymbol{t}_n^P, y_n)\}_{n=1}^{N_L}$, where $\boldsymbol{x} \in \mathcal{X}$ is the image and $\boldsymbol{t}^P \in \mathcal{P}$ is a prompt-style language description based on its class label $y \in \mathcal{Y}$, and (ii) a dataset of image-caption pairs $\mathcal{S}_C = \{(\boldsymbol{x}_n, \boldsymbol{t}_n^C)\}_{n=1}^{N_C}$, where $\boldsymbol{t}^C \in \mathcal{T}$ is a caption. We assume that $\boldsymbol{t}$ is the tokenized language description. For each image $\boldsymbol{x}$, an image encoder model $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ extracts a visual representation $\tilde{\boldsymbol{v}} \in \mathbb{R}^{d \times 1}$:

$\tilde{\boldsymbol{v}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$. For each caption or prompt $\boldsymbol{t} \in \mathcal{T}$, a text encoder $f_{\boldsymbol{\phi}}$ parameterized by $\boldsymbol{\phi}$ extracts a language representation $\tilde{\boldsymbol{u}} \in \mathbb{R}^{d\times 1} : \tilde{\boldsymbol{u}} = f_{\boldsymbol{\phi}}(\boldsymbol{t})$.

Contrastive Loss. CLIP [31] is designed to find representations that match an image to its paired caption while separating unpaired ones. For i-th image $\boldsymbol{x}_i$ and j-th language description $\boldsymbol{t}_j$ in a batch $\mathcal{B}$, their features are normalized using $\boldsymbol{v}_i = \frac{\tilde{\boldsymbol{v}}_i}{\|\tilde{\boldsymbol{v}}_i\|}$ and $\boldsymbol{u}_j = \frac{\tilde{\boldsymbol{u}}_j}{\|\tilde{\boldsymbol{u}}_j\|}$. Finally, CLIP optimizes the symmetric multi-class N-pair loss [35]:

$$\min_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}} \quad \mathcal{L}_{con} = \mathcal{L}_{t2i} + \mathcal{L}_{i2t}, \tag{1}$$

which includes two contrastive terms (a temperature hyperparameter τ controls the strength of penalties on hard negative samples):

$$\mathcal{L}_{t2i} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j)}, \tag{2}$$

$$\mathcal{L}_{i2t} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \frac{\exp(\tau \boldsymbol{v}_i^T \boldsymbol{u}_i)}{\sum_{j\in\mathcal{B}} \exp(\tau \boldsymbol{v}_i^T \boldsymbol{u}_j)}. \tag{3}$$

UniCL [43] composes each mini-batch with samples from both $\mathcal{S}_L$ and $\mathcal{S}_C$. Then, for pairs from $\mathcal{S}_L$, they regard all samples from the same class as positive pairs while a sample from $\mathcal{S}_C$ has a unique pair. Except for the number of positive pairs, no special treatment is given to differentiate between the image-caption and image-label data.

3.2. Prefix Conditioned Contrastive Learning

Fig. 2 describes the overview of our approach. We aim to enable the language encoder to learn embedding strategies conditioned on the type of input dataset. The conditioning can then be used to manipulate the bias at inference time.

Prefix-tuning [10, 18, 21, 23, 30] shares the intuition that the prefix tokens are responsible for switching the context of a language model from the pre-trained task to the downstream task. These approaches leverage the prefix to tailor a model to a single task during fine-tuning and construct different prefixes for different natural language tasks [18]. In our problem setting, there is no task distinction between the image-caption and image-prompt matching since both are formulated as contrastive learning. However, we focus on the fact that the two datasets have different biases in the image distributions and vocabulary words. The label-prompt sentences are embedded closer to the image classification data, even though we may want to use them to match a new label to an image from the open-domain image distribution during zero-shot classification.

To solve this problem, we propose to inform the model of the type of dataset at the input level to switch the feature extraction. Specifically, to make the model aware of the dataset type, prefix-conditioning prepends a prefix token to an input sentence to obtain $\bar{\boldsymbol{t}}^P = [\text{PREFIX}^P; \boldsymbol{t}^P]$, $\bar{\boldsymbol{t}}^C = [\text{PREFIX}^C; \boldsymbol{t}^C]$. The brackets indicate the concatenation of two lists of discrete tokens; PREFIX^P and PREFIX^C denote a prompt-style and caption-style token respectively. In this way, we prepend the token to learn the dataset-specific bias, which enables us to disentangle the bias in language representations and utilize the embedding learned on the image-caption dataset at test time *even without an input caption*.

In prompt-tuning, the number of prefix tokens can affect the performance of the model [18, 21, 46]. However, we do not see the performance difference by the number of prefix tokens. This is probably because adding one token is enough to distinguish the domain of input sentences. To avoid significantly increasing the training cost, we set the number of prefixes to one in all experiments. Then, the language representations for each data source are extracted as $\tilde{\boldsymbol{u}}^P, \tilde{\boldsymbol{u}}^C = f_{\boldsymbol{\phi}}(\bar{\boldsymbol{t}}^P), f_{\boldsymbol{\phi}}(\bar{\boldsymbol{t}}^C)$. This input design is independent from the training objectives, and therefore we can easily apply the technique to optimize Eq. 1 or UniCL's loss.

Data Sampling. [6] argue that the data sampling matters when learning from multiple data sources in a contrastive learning framework, as the model may learn to distinguish the samples by exploiting the dataset bias. As such, we need to take data sampling into consideration in our problem setting as we learn from two different data sources. One option is a debiased sampling [6], which constructs each mini batch to contain samples from a single data source. Alternatively, as done in UniCL [43], we can compose each minibatch with samples from both data sources (image-caption and image-label) with equal probability. In experiments, we choose the debiased sampling, but find that the choice of sampling does not significantly affect the performance.

3.3. Inference with Prefix Conditioning

During inference (the right side of Fig. 2), an input image is classified as one of K classes by embedding the corresponding label-prompts and choosing the one most similar to the image embedding. Following [31], we obtain class prompts by filling the default prompt templates with class names, and add a prefix. Considering the wider coverage of domains in the image-caption dataset, the caption-style prefix conditioning may work better to classify novel downstream data. In our experiments, we empirically find that the caption-style prefix indeed outperforms the prompt-style prefix with a large margin in zero-shot recognition while prompt prefix performs better on the image classification dataset used to train the model. We provide a detailed analysis of different conditioning in Section 4.3.

4. Experiments

The goal of experiments is twofold: comparing our approach with baselines in zero-shot recognition, and analyzing the behavior of prefix conditioning. We describe the experimental setup in Sec. 4.1, show the main results in

Training Data			Objective	Prefix Conditioning	Metric	
Classification	Caption	Size			IN-1K	Zero-shot 11 datasets
–	CC-3M	3M	CLIP		18.1	28.7
–	CC-12M	12M	CLIP		33.4	41.2
ImageNet-1K	–	1M	CLIP		72.1	20.2
ImageNet-21K	–	12M	CLIP		47.1	39.6
ImageNet-1K	CC-12M	13M	CLIP		68.7	43.3
ImageNet-1K	CC-12M	13M	CLIP	✓	**71.5**	**45.5**
ImageNet-1K	CC-12M	13M	UniCL		68.8	43.1
ImageNet-1K	CC-12M	13M	UniCL	✓	**71.7**	**44.5**
ImageNet-21K	CC-12M	25M	CLIP		56.8	49.5
ImageNet-21K	CC-12M	25M	CLIP	✓	**67.3**	**57.8**
ImageNet-21K	CC-12M	25M	UniCL		58.2	51.7
ImageNet-21K	CC-12M	25M	UniCL	✓	**66.5**	**58.4**
ImageNet-21K w/o IN-1K	CC-12M	24M	CLIP		29.1	46.9
ImageNet-21K w/o IN-1K	CC-12M	24M	CLIP	✓	**47.8**	**56.4**

Table 1. Performance comparison among different training datasets and training objectives. Note that we use caption prefix to obtain these results. The proposed prefix conditioning shows improved zero-shot recognition accuracy across models trained with different combinations of image-classification and image-caption datasets and training objectives.

Sec. 4.2, and analyze the properties of prefix-conditioning in Sec. 4.3.

4.1. Setup

Training Datasets. We conduct experiments on the setting where we have a large source of image-caption and image-label datasets. Following UniCL [43], we utilize CC3M [33] and CC12M [5] as image-caption data. For the image classification dataset, we utilize ImageNet21K and ImageNet1K [7]. While ImageNet1k contains 1,000 classes, ImageNet21K has more than 20,000 categories that include fine-grained and general objects. To observe the behavior in diverse image classification data, we also run experiments on ImageNet21K while excluding the classes of ImageNet1K. Details are explained in each section.

Training. We use the same prompt strategy and 80 prompt templates as used in CLIP [31]. During training, we randomly sample one prompt template and fill it with the class names, followed by a tokenization step before feeding into the text encoder. We average language embeddings extracted from all 80 templates in validation. We use the same language encoder as CLIP [31] and Swin-Tiny transformer [24] as the vision encoder following UniCL [43]. All models are optimized with AdamW [25] where the learning rate is set to 0.001, and weight decay to 0.1. All models are trained with a batch size of 1024. Considering the amount of training data, we train the models for 15 and 50 epochs in the experiments on ImageNet21K and ImageNet1K respectively.[2] For all training, we used a cosine learning rate schedule with a warm-up of 10,000 iterations.

Baselines. We train CLIP [31] and UniCL [43] as our baselines. For comparison, we present results on CLIP trained only on image-caption or image classification data, as well as CLIP and UniCL trained on both image-caption and IN21K data. Unless otherwise stated, CLIP and UniCL are trained with equal sampling (ES) strategy as in [43], while our prefix conditioning model is trained with debiased sampling (DS) [6]. We provide an analysis of the sampling in Sec. 4.2 and find that DS itself does not have a noticeable advantage over ES.

Evaluation. We evaluate the learned representations on supervised and zero-shot image classification on ImageNet1K[3] and on 11 datasets chosen from the ones used in CLIP [31] including object classification (e.g., CIFAR10, CIFAR100), fine-grained classification (e.g., Oxford-IIIT Pets, Oxford Flowers 102, and Food-101), and aerial images (e.g., EuroSAT and Resisc45). Although our main focus is at the zero-shot generalization, we also provide an analysis of a linear-probe evaluation of the image encoder.

[2] When training a model on two different datasets, e.g., IN21K and CC12M, we count the epochs based on how many samples are used from the image classification dataset. For instance, in UniCL, each mini-batch consists of approximately the same number of samples from IN21K and CC12M. Then, to train a model for 15 epochs, we train for $N/1024 \times 2 \times 15$ iterations, where N indicates the number of samples in IN21K.

[3] While we follow the same zero-shot evaluation protocol when evaluating on ImageNet1K, we note that it is zero-shot only where we explicitly exclude ImageNet1K from the training, last two rows of Table 1

Train Prefix	Sampling	IN-1K	Cal	CF100	CF10	ESTAT	Food	Flower	Pets	Patch	R45	VOC	DTD	AVG
	ES	56.8	70.2	55.0	79.4	21.1	46.0	60.3	57.2	51.2	24.8	57.7	21.4	49.5
✓	ES	**65.4**	**81.2**	**62.6**	**88.9**	**30.4**	**51.7**	**61.8**	**71.9**	50.0	**28.2**	**78.1**	**27.7**	**57.5**
	DS	58.7	65.9	55.0	85.7	22.8	40.8	55.7	60.2	50.0	20.6	45.2	23.8	47.8
✓	DS	**67.3**	**79.7**	**63.8**	**87.9**	**31.5**	**53.4**	**58.8**	**69.6**	**50.6**	**31.5**	**80.5**	**28.4**	**57.8**

Table 2. Ablation study for sampling in IN21K + CC12M. Equal sampling (ES) composes a mini-batch with roughly equal number of samples from two datasets. Debiased sampling (DS) samples a mini-batch of either IN21K or CC12M with equal probability.

Train Data	Prefix Conditioning	IN-1K	Cifar10	Cifar100	Caltech	Food	Pet	Patch	VOC	DTD
ImageNet-21K		71.5	94.3	79.1	83.5	79.1	86.3	82.3	88.9	61.3
ImageNet-21K + CC12M		69.2	93.0	76.4	82.4	78.4	82.2	81.4	88.7	61.4
ImageNet-21K + CC12M	✓	69.4	93.5	77.3	83.2	78.8	83.6	82.0	88.8	62.5

Table 3. Linear evaluation accuracy on models trained with and without prefix conditioning. Prefix conditioning slightly improves the performance upon a model without it (second row vs. last row).

4.2. Main Results

We describe our main results in Table 1, followed by the analysis of prefix conditioning in Sec. 4.3.

There are three observations. First, the improvements upon a model trained only with image-caption or image-label data are obvious in almost all cases. As the previous work indicates [43], the effectiveness of combining two types of supervision is clear from these results.

Second, in all cases, our prefix conditioning significantly improves performance on both ImageNet-1K (supervised recognition) and 11 zero-shot recognition tasks. When training on ImageNet-21K, the conditioning improves the baseline by more than 8% in ImageNet-1K and more than 6% in zero-shot recognition on average. In training with ImageNet-1K, the margin from the baseline is smaller than training with ImageNet-21K, probably because the size of ImageNet-1K (1M) is much smaller than that of ImageNet-21K (12M). Also, prefix conditioning is effective in both UniCL and CLIP objectives. Due to its simplicity, our approach can be easily integrated with various objectives.

Finally, our method is less affected by ablating a part of categories. The classes of ImageNet-1K are excluded from ImageNet-21K in the last two rows of Table 1. Therefore, both approaches significantly drop performance on ImageNet-1K, whose task now becomes true zero-shot recognition, compared to other settings. Even in this setting, prefix conditioning maintains high accuracy and outperforms a CLIP baseline model by a large margin.

Sampling Method. We analyze the data sampling scheme to construct a mini-batch in Table 2. We apply debiased sampling (DS) in our method, namely, sampling one data source with equal probability and getting a mini-batch of it. The other option is mixing two data sources with equal probability (ES). The table indicates that prefix conditioning works well with ES sampling and the sampling strategy itself is not advantageous. Ablating prefix conditioning during training clearly drops the performance in both sampling strategies, and the performance is worse than ES on average in zero-shot results (49.5 vs. 47.8). ES sampling should allow the model to differentiate sentences by using the prepended prefix. Interestingly, this result implies that differentiating sentences by prefix information does not much degrade the performance. The distinguished sentences enable the model to associate images from different datasets. Since images of two datasets are different with respect to the categories and the locations of objects in images, distinguishing the two kinds of images may not harm generalizability of the representations.

Linear-probe Evaluation. We evaluate the linear-probe performance in Table 3 to see the quality of learned image representations. Although the accuracy is better than the model trained without prefix conditioning (second line), the improvements are not substantial. This result indicates that the zero-shot performance gain obtained by our method is not due to the image representations. We investigate the learned language and image features in the next subsection.

4.3. Analysis of Prefix-Conditioning

We present a detailed analysis of prefix conditioning. We first study how different prefixes impact the zero-shot recognition performance and analyze their behaviors by looking into the attention weights of the language transformer encoder. We also demonstrate improved robustness with respect to the image-level domain shift. Unless otherwise stated, we employ a model trained with CLIP objective on ImageNet-21K and CC12M in this analysis. Finally, this section concludes that prefix conditioning enables the language encoder to switch its role during training, which eases learning from different types of datasets, e.g., image classification and image caption dataset.

Data	Test-time Prefix	IN-1K	Cal	C100	C10	ESTAT	Food	Flower	Pets	Patch	R45	VOC	DTD	AVG
IN-1K + CC12M	N/A	68.7	68.7	38.4	69.5	24.4	31.9	13.3	66.6	50.2	25.4	65.6	22.3	43.3
	Prompt	**75.4**	71.7	35.5	63.9	24.2	20.0	8.1	72.2	50.4	24.2	61.1	15.3	40.6
	Caption	71.5	**75.1**	**39.4**	**70.5**	**26.7**	**33.9**	**13.9**	**72.3**	**50.5**	**25.8**	**67.8**	**25.4**	**45.5**
IN-21K + CC12M	N/A	56.8	70.2	55.0	79.4	21.1	46.0	60.3	57.2	51.2	24.8	57.7	21.4	49.5
	Prompt	**71.4**	76.5	59.0	86.0	20.1	45.7	**62.3**	69.1	**52.4**	26.3	76.8	21.4	54.1
	Caption	67.3	**79.7**	**63.8**	**87.9**	**31.5**	**53.4**	58.8	**69.6**	50.6	**31.5**	**80.5**	**28.4**	**57.8**
IN-21K w/o 1K + CC12M	N/A	29.1	67.4	45.9	80.0	28.6	40.8	56.9	39.2	50.2	21.9	64.9	19.8	46.9
	Prompt	40.8	74.9	61.0	84.6	31.2	48.1	58.7	45.2	**51.2**	23.5	67.5	21.4	51.6
	Caption	**47.8**	**81.9**	**63.3**	**87.3**	**32.4**	**52.9**	**62.8**	**57.0**	50.6	**25.6**	**80.1**	**26.2**	**56.4**

Table 4. Ablation study for test-time prefix conditioning. Note that the difference between two results come from the prefix used in test time and we use the same model for this evaluation. A model trained without conditioning is shown at the top of each block.

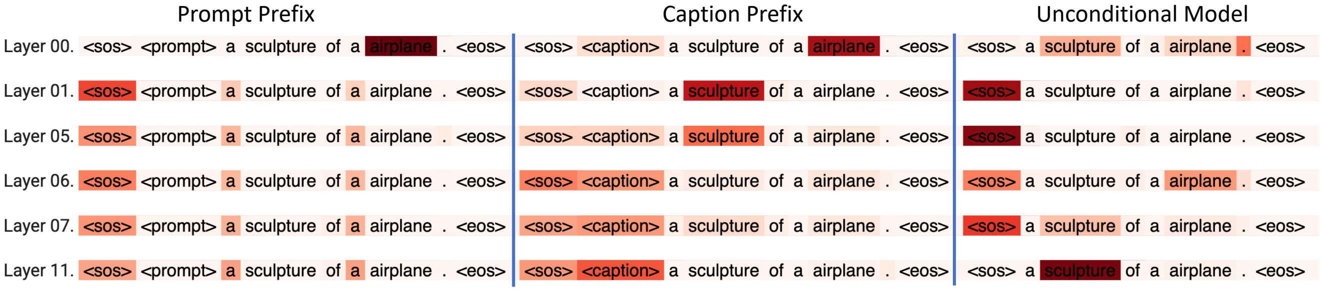


Figure 3. An example of attention weights for an end token. Best viewed in color. The sentence shown here is one of class prompts in the VOC 2007 dataset. Different rows show the weights of different transformer layers. With a prompt prefix (leftmost), the model focuses on a class name (*airplane*) while caption prefix (middle) allows a model to pay attention to another noun, *sculpture*. By prefix conditioning, the attention of the model changes as intended.

Test Time Prefix. We analyze the role of the prefix token in Table 4, where the table describes the comparison in the choice of test time prefix conditioning. As explained in Sec. 3, the choice of prefix during test time should change the behavior of the model since the prefix should tailor the language encoder for classification-style or caption-style feature extraction. Except for the IN-1K results of a model trained with the entire IN21K or IN-1K, conditioning with the caption prefix shows much better results. The superiority of the caption prefix is noticeable in several datasets. This means caption prefix works better if the target comes from outside the image classification data, indicating that the class-prompt prefix conditioning makes the model tailored for the image classification dataset. Class-prompts prefix works better to categorize IN-1K data because the prefix is trained to specialize in classifying it. Note that caption-style prefix performs better than prompt-style prefix in IN-1K for a model trained with IN21K excluding IN1k classes. This indicates that the caption-style prefix works better when the vocabulary of the class name comes from outside the image classification data since the caption data covers much more diverse words.

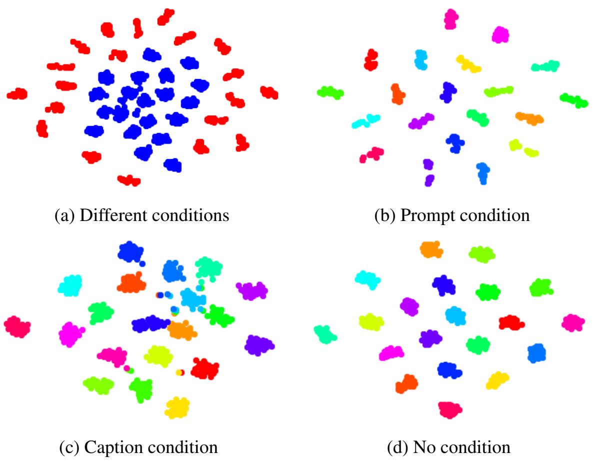
Prefix controls attention. Fig. 3 visualizes the attention weights for an end token in different prefix conditions and models. The input sentence, *a sculpture of an airplane*, is one of the class-prompts. When a prompt prefix (leftmost) is employed, the language model pays attention to the class name at the first layer, it does not focus on the noun in other layers. The only noun the encoder focuses on is *airplane*. By contrast, the model attends to both *sculpture* and *airplane* in the case of the caption prefix and unconditional model. Note that this behavior does not mean that the prompt-prefix performs better in zero-shot recognition as shown in experiments due to the effect of the bias in image classification dataset.

While we visualize only one example in the main text due to the space limit and defer more examples to the appendix, this highlights a general trend that the prompt prefix guides the language encoder to focus on a single word (e.g., class name), whereas the caption prefix makes the model attend to multiple words. In other words, prefix conditioning allows the language encoder to “switch gears” to represent sentences from different datasets (i.e., image-classification vs image-caption). On the other hand, the baseline model without prefix conditioning attends to multiple words (e.g., Fig. 3 rightmost) even though the input sentence is a class prompt. This indicates that it is hard to switch the gears without explicitly informing of the type of dataset.

Language Feature Visualization. Fig. 4 visualizes extracted language features conditioned with different pre-

Train Data	Prefix Conditioning	Test-Time Prefix	IN	IN-V2	IN-R	IN-S
ImageNet-1K		N/A	72.1	59.3	19.9	17.8
ImageNet-1K + CC12M		N/A	68.7	57.4	27.7	27.8
ImageNet-1K + CC12M	✓	Caption	71.5	60.2	**31.8**	**30.7**
ImageNet-1K + CC12M	✓	Prompt	**75.4**	**63.3**	29.2	27.9
ImageNet-21K		N/A	47.1	41.1	20.1	16.1
ImageNet-21K + CC12M		N/A	56.8	48.6	29.4	30.6
ImageNet-21K + CC12M	✓	Caption	67.3	57.5	**35.2**	**34.6**
ImageNet-21K + CC12M	✓	Prompt	**71.4**	**61.1**	32.1	32.2

Table 5. Evaluation on the robustness to the image-level domain shift. Prefix conditioned training achieves better robustness, and caption-prefix outperforms prompt-prefix in the images distinct from those used in training (IN-R and IN-S).


(a) Different conditions

(b) Prompt condition

(c) Caption condition

(d) No condition

Figure 4. T-SNE [36] visualization of the class-prompt features of 20 classes of VOC 2007 with different prefix conditions. (a): Language embeddings with prompt (red) and caption (blue) prefixes, respectively. (b)(c)(d): Different colors indicate language embeddings of different classes.

fixes. As seen in Fig. 4a, language features extracted with caption-prefix (blue) and prompt-prefix (red) are clearly separated. In addition, prompt-prefix (Fig. 4b) has lower intra-class and higher inter-class variance, whereas caption-prefix (Fig. 4c) shows higher intra-class variance across prompts. Interestingly, results in Table 4 suggest that the caption-prefix conditioned language features result in a better zero-shot recognition performance than those conditioned on the prompt-prefix. Although the prompt-prefix mode extracts discriminative language embeddings, the embeddings do not perform well on the zero-shot recognition because the embeddings contain significant bias from image-classification dataset.

Robustness in image domain shift. Test samples can be unseen with respect to image classification data in two ways (or combinations of two): 1) The image is similar to the training distribution, but the class name is different from the seen image classification labels. 2) Although the class label is the same, the image data comes from a different distribution. Datasets evaluated in the zero-shot recognition include both two cases since the vocabularies and image are from different domains. To understand them, we analyze the test-time prefix by using ImageNet-1K and evaluate the performance on image-level domain shift using variants of ImageNet, i.e., ImageNet-V2 [32], ImageNet-R [11], and ImageNet-S [37]. Table 5 describes the results of ablating prefix-conditioned training and the test-time prefix. The prefix-conditioned training outperforms all baselines. This reveals that the prefix-conditioned training achieves class embeddings that are generalizable across image domains. The prompt-style prefix performs the best in IN, IN-V2, both of which have image styles similar to ImageNet. By contrast, the caption-style prefix performs the best in IN-R and IN-S, which has art-style and sketch-style images respectively. Thus, the caption-style prefix generates more generalizable class embeddings for the domain dissimilar from the ImageNet training data. This observation is consistent with the results in the paragraph *Test time Prefix*.

5. Conclusion

In this paper, we explore a simple yet effective mechanism for unified pre-training on image-caption and image classification data. We propose to learn prefix tokens at training time to condition the language encoder to switch the input source. Specifying the prefix allows the model to switch the manner of feature extraction and can control which visual domain the embedding is projected to. This approach boosts the performance of zero-shot recognition accuracy of the contrastive learning models. Our analysis suggests that the trained language encoder provides robustness to the image-level domain shift. Although we limit our scope to unifying image-caption and image-label supervision, incorporating other supervision such as object detection or semantic segmentation is an interesting next step.

Acknowledgment. We thank Zizhao Zhang for their helpful feedback on the manuscript. This work was supported in part by DARPA LwLL.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2927–2936, 2015. 2

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 3

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 3

[4] Stephanie CY Chan, Adam Santoro, Andrew K Lampinen, Jane X Wang, Aaditya Singh, Pierre H Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent few-shot learning in transformers. *arXiv preprint arXiv:2205.05055*, 2022. 1

[5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1, 2, 5

[6] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, and Osamu Yoshie. Zerovl: A strong baseline for aligning vision-language representations with limited resources. *arXiv preprint arXiv:2112.09331*, 2021. 4, 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2, 3, 5

[8] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14340–14349, 2021. 3

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 2

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 3, 4

[11] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 8

[12] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems*, 27, 2014. 2

[13] 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. 1, 2

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 3

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2

[16] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 3

[17] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021. 3

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2, 3, 4

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2

[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. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2, 3, 4

[22] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 3, 4

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 5

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[26] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2441–2448, 2014. 2

[27] Christian M Meyer and Iryna Gurevych. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012. 2

[28] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[29] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 2

[30] Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021. 3, 4

[31] 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, 4, 5

[32] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 8

[33] 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, 2018. 5

[34] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022. 2

[35] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016. 4

[36] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[37] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 3, 8

[38] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018. 2

[39] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 3

[40] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 69–77, 2016. 2

[41] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4582–4591, 2017. 2

[42] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. *arXiv preprint arXiv:2202.10401*, 2022. 2

[43] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *arXiv preprint arXiv:2204.03610*, 2022. 1, 2, 4, 5, 6

[44] 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. *arXiv preprint arXiv:2111.07783*, 2021. 2

[45] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 2, 3, 4

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*, 2022. 3