CAILA: Concept-Aware Intra-Layer Adapters for Compositional Zero-Shot Learning

Zhaoheng Zheng Haidong Zhu Ram Nevatia
Viterbi School of Engineering
University of Southern California
{zhaoheng.zheng, haidongz, nevatia}@usc.edu

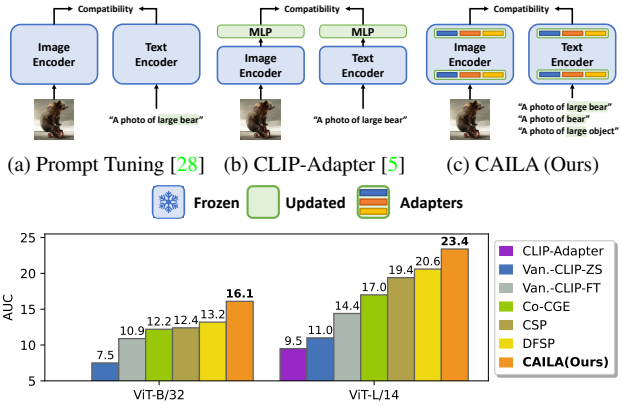
Abstract

In this paper, we study the problem of Compositional Zero-Shot Learning (CZSL), which is to recognize novel attribute-object combinations with pre-existing concepts. Recent researchers focus on applying large-scale Vision-Language Pre-trained (VLP) models like CLIP with strong generalization ability. However, these methods treat the pre-trained model as a black box and focus on pre- and post-CLIP operations, which do not inherently mine the semantic concept between the layers inside CLIP. We propose to dive deep into the architecture and insert adapters, a parameter-efficient technique proven to be effective among large language models, into each CLIP encoder layer. We further equip adapters with concept awareness so that concept-specific features of "object", "attribute", and "composition" can be extracted. We assess our method on four popular CZSL datasets, MIT-States, C-GQA, UT-Zappos, and VAW-CZSL, which shows state-of-the-art performance compared to existing methods on all of them.

1. Introduction

When facing a novel concept such as a *large castle*, humans can deconstruct individual components (*large* and *castle*) from familiar concepts (*large bear, old castle*) to comprehend the new composition. Such task of recognizing new attribute-object compositions based on a set of observed pairs is Compositional Zero-Shot Learning (CZSL) [24], a sine qua non for an intelligent entity. However, the inherent challenge in CZSL lies in the capacity to identify unobserved novel compositions without compromising the recognition of previously observed combinations. Conventional approaches [1, 19, 20, 23–27, 31, 35, 39, 40, 43] often suffer from training biases. Even though recent methods employ large-scale Vision-Language Pre-training (VLP) models with strong generalization ability, e.g., CLIP [32], to accommodate this issue, they simply treat VLP models as frozen black box encoders and fail to exploit the potential of VLP models. Thus, here, we explore how to more effectively extract and utilize the knowledge embedded in pre-trained vision-language models for the recognition of novel attribute-object compositions.

Code will be available at github.com/zhaohengz/CAILA



(a) Prompt Tuning [28] (b) CLIP-Adapter [5] (c) CAILA (Ours)

(d) Comparisons of AUC on MIT-States [9] with different backbones

Figure 1. Illustrations of CAILA and previous CLIP-based baselines. CAILA has adapters integrated into both CLIP encoders and thus better transfers the knowledge from CLIP to CZSL, resulting in significant performance boosts compared with other CLIP-based baseline methods. "Van.-CLIP" refers to models with vanilla CLIP architecture. `Prompts` highlighted in green are set to be learnable parameters.

More specifically, to adapt VLP models for CZSL, some researchers apply prompt-tuning [28, 47, 48] or fine-tune the model with extra adaptation layers [5] on the top of CLIP. However, prompt-tuning methods, depicted in Figure 1(a), only learn trainable prompts, while CLIP-Adapter, shown in Figure 1(b), only adds external modules outside CLIP. Both strategies abstain from altering the fundamental CLIP encoder, consequently retaining CLIP as a static black box. Nayak *et al.* [28] have shown that exhaustively fine-tuning CLIP falls short of attaining practicable performance. Thus,

we argue that properly optimizing features across layers through a task-specific design is critical to effectively harnessing the knowledge embedded in CLIP. A feasible CLIP-based CZSL should: i) have task-specific designs for CZSL; ii) be capable of extracting concept-specific features related to compositions and individual primitives.

Hence, we propose **CAILA**, **C**oncept-**A**ware **I**ntra-**L**ayer **A**dapters, that satisfy the given prerequisites and substantiate its superiority, as shown in Fig. 1(d), compared with other CLIP-based methods. Fig. 1(c) highlights the difference between CAILA and other VLP-based methods. Instead of prompt tuning or fully fine-tuning, we adopt adapters [7] to transfer knowledge from VLP models while avoiding strong training biases.

Moreover, given that adapters are low-overhead components, it is feasible to employ a variety of adapters to extract concept-wise representations. More specifically, CAILA integrates a group of adapters into each layer of both encoders; each group possesses concept-specific components to extract knowledge corresponding to particular concepts, including attributes, objects, and compositions. To merge features extracted by various concept-aware adapters, we propose the **M**ixture-of-**A**dapters (MoA) mechanism for both vision and yrcy encoder. In addition, the property that CAILA can extract concept-specific features allows us to further propose Primitive Concept Shift, which generates additional vision embeddings by combining the attribute feature from one image and the object feature from another for a more comprehensive understanding.

We evaluate our approach on three popular CZSL datasets: MIT-States [9], C-GQA [25] and UT-Zappos [44, 45], under both closed world and open world settings. We also report the performance of CAILA in closed world on VAW-CZSL [35], a newly released benchmark. Our experiments show that, in both scenarios, our model beats the state-of-the-arts over all benchmarks following the generalized evaluation protocol [31], by significant margins.

To summarize, our contributions are as follows: (i) We propose CAILA, which is the first model exploring CZSL-oriented designs with CLIP models to balance model capacity and training bias robustness; (ii) we design the Mixture-of-Adapter (MoA) mechanism to fuse the knowledge from concept-aware adapters and improve the generalizability; (iii) we further enrich the training data and exploit the power of CAILA through Primitive Concept Shifts; (iv) we conduct extensive experiments in exploring the optimal setup for CAILA on CZSL. Quantitative experiments show that our model outperforms the SOTA by significant margins in both closed world and open world, on all benchmarks.

2. Related Works

Zero-Shot Learning (ZSL). Unlike conventional fully-supervised learning, ZSL requires models to learn from side information without observing any visual training samples [16]. The side information comes from multiple non-visual resources such as attributes [16], word embeddings [36, 38] , and text descriptions [33]. Notably, Zhang *et al.* [46] propose to learn a deep embedding model bridging the seen and the unseen, while [2, 42, 49] investigate generative models that produce features for novel categories. Moreover, [11, 38] integrate Graph Convolution Networks (GCN) [15] to better generalize over unseen categories.

Compostional Zero-Shot Learning (CZSL). Previous CZSL approaches are built with pre-trained image encoders, *e.g.* ResNet and separate word embeddings, *e.g.* GloVe [30]. More specifically, Li *et al.* [20] investigate the symmetrical property between objects and attributes, while Atzmon *et al.* [1] study the casual influence between the two. Moreover, Li *et al.* [19] construct a Siamese network with contrastive learning to learn better object/attribute prototypes. On the other hand, joint representations of compositions can be leveraged in multiple ways. [31] utilizes joint embeddings to control gating functions for the modular network, while [26, 27, 40, 43] treat them as categorical centers in the joint latent space. Furthermore, some approaches [23–25, 34, 39] directly take compositional embeddings as classifier weights, while OADis [35] disentangles attributes and objects in the visual space.

Parameter-Efficient Tuning. Recent research on large scale pre-training models [6, 8, 10, 18, 32] has achieved superior performance on various downstream tasks, compared with regular approaches. Various works [7, 12, 37] show that tuning adapters [7] on the language side yields comparable results with fully fine-tuned variants, while Chen *et al.* [3] investigate the adaptation of image encoders on dense prediction tasks. For CZSL, a few models [28, 48] leverage the knowledge of CLIP through prompt tuning [17] , while Gao *et al.* [5] attach a post-processor to CLIP for knowledge transfer. Though these methods show strong performance on CZSL against regular models, they treat the CLIP model as a black box and keep it completely frozen. In CAILA, we open up the CLIP black box by integrating intra-layer adapters to both image and text encoders.

3. Approach

The problem of CZSL can be formulated as follows. We denote the training set by $\mathcal{T} = \{(x, y)|x \in \mathcal{X}, y \in \mathcal{Y}_s\}$, where $\mathcal{X}$ contains images represented in the RGB color space and $\mathcal{Y}_s$ is a set of seen composition labels which are available during the training phase. Each label $y = (a, o)$ is a pair of attribute $a \in \mathcal{A}$ and object category $o \in \mathcal{O}$. When testing, CZSL expects models to predict a set of unseen compositions $\mathcal{Y}_u$ that is mutually exclusive with training labels $\mathcal{Y}_s$: $\mathcal{Y}_s \cup \mathcal{Y}_u = \varnothing$. Note that $\mathcal{Y}_s$ and $\mathcal{Y}_u$ share the same set of $\mathcal{A}, \mathcal{O}$, while CZSL assumes that each $a \in \mathcal{A}$ or $o \in \mathcal{O}$ exists in the training set and only the composition

$(a, o) \in \mathcal{Y}_u$ is novel. Following [25, 31, 41], we focus on generalized CZSL, where the test set contains both seen and unseen labels, formally denoted by $\mathcal{Y}_{test} = \mathcal{Y}_s \cup \mathcal{Y}_u$.

Most recent works [1,25,31] study the generalized CZSL problem under the *closed world* setting, where $\mathcal{Y}_{test}$ is a subset of the complete composition set $\mathcal{Y} : \mathcal{A} \times \mathcal{O}$. The *closed world* setting assumes that $\mathcal{Y}_u$ are known during testing and thus greatly reduce the size of the search space. On the contrary, Mancini *et al.* [22] argue that such constraint should not be applied to the search space and introduce the *open world* setting, where models are required to search over the complete set of compositions, formally $\mathcal{Y}_s \cup \mathcal{Y}_u = \mathcal{Y}$. In this paper, we investigate the problem in both *closed world* and *open world*.

3.1. Compatibility Estimation Pipeline

As different attributes can lead to significant appearance shifts even inside the same object category, performing attribute and object predictions separately may be ineffective. Hence, we model attribute-object compositions jointly and learn a combined estimation function to measure the compatibility of input image x and query composition (a, o). In addition, we let the model estimate attribute and object compatibilities as auxiliary sub-tasks during training.

The estimation of composition compatibility is represented as $\mathcal{C}(x, a, o) : \mathcal{X} \times \mathcal{A} \times \mathcal{O} \rightarrow \mathbb{R}$. It contains two components: The image feature extractor $\mathcal{F}_C : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^d$ and the text embedding generator $\mathcal{G} : \mathcal{A} \times \mathcal{O} \rightarrow \mathbb{R}^d$. Note that d denotes the number of channels that each representation has. Given an image x and a composition (a, o), the compatibility score is defined as the dot product of $\mathcal{F}_C(x)$ and $\mathcal{G}(a, o)$, formally

$$\mathcal{C}(x, a, o) = \mathcal{F}_C(x) \cdot \mathcal{G}(a, o). \tag{1}$$

Furthermore, as CZSL requires models to recognize novel pairs composed of known attributes and objects, it is important for a model to possess the capability of primitive feature extraction that is disentangled with training compositions. Thus, we make our model extract features corresponding to primitives and estimate the compatibility between vision features and text representations during training. Similar to Eqn. 1, we have

$$\mathcal{C}(x, a) = \mathcal{F}_A(x) \cdot \mathcal{G}_A(a),\ \mathcal{C}(x, o) = \mathcal{F}_O(x) \cdot \mathcal{G}_O(o). \tag{2}$$

All three compatibility scores contribute independently to the loss function, while $\mathcal{C}(x, a, o)$ is leveraged during inference. More specifically, our framework learns separate representations through CAILA discussed in Sec. 3.2 and conducts knowledge fusion through **M**ixture-of-**A**dapters (MoA), which will be covered in Sec. 3.3.

Following [32], we create a prompt template similar to `"a photo of [CLASS]"` for each compatibility estimation sub-task. For composition compatibility, we feed the text encoder with `"a photo of [ATTRIBUTE] [OBJECT]"`; We use `"a photo of [ATTRIBUTE] object"` and `"a photo of [OBJECT]"` for attribute and object compatibilities, respectively. Similar to [28], we only make `[CLASS]` prompts trainable. For both encoders $\mathcal{F}$ and $\mathcal{G}$, we take the output hidden state of the **`[CLS]`** token as the representation.

3.2. Concept-Aware Intra-Layer Adapters

Though CLIP-based CZSL approaches [5, 28, 48] have achieved significant improvements compared with earlier methods [22, 24, 25, 28, 31], the CLIP encoder is considered as a black box and no modifications are made to improve its generalizability. Thus, we propose to improve CLIP-based CZSL models in both modalities with CAILA , Concept-Aware Intra-Layer Adapters.

As shown in Fig. 2 (a)(b), we take the CLIP image encoder as $\mathcal{F}$ and the text encoder as $\mathcal{G}$, while adding concept awareness to both encoders when estimating compatibilities of different concepts. Fig. 2 (c) demonstrates how adapters are integrated into a regular transformer encoding block. For each encoding block, we add adapters behind the frozen self-attention layer and the feed-forward network. More specifically, given the input hidden state $\mathbf{h}$ of an adapter, we compute the latent feature $\mathbf{z}$ by the downsampling operator f_{Down}, followed by the activation function σ. The output $\mathbf{h}'$ of an adapter is obtained by upscaling $\mathbf{z}$ and summing it with $\mathbf{h}$ through the skip connection. Formally, we have

$$\mathbf{z} = \sigma(f_{Down}(\mathbf{h})), \quad \mathbf{h}' = f_{Up}(\mathbf{z}) + \mathbf{h}, \tag{3}$$

where both f_{Down} and f_{Up} are fully-connected layers.

To extract concept-specific features, at each depth level, we create three encoding blocks corresponding to attribute, object, and composition, respectively. As in Fig. 2(c), encoding blocks of at the same level share the same weights except for the adapter layers. Inputs from both modalities are processed by encoders equipped with different types of encoding blocks and features related to each of the three concepts are produced. During training, vision-language compatibility scores for "attribute", "object" and "compositions" are estimated. More specifically, encoders referred in Fig. 2(a) and (b) are the same ones; There are not extra side encoders for auxiliary sub-tasks.

3.3. MoA: Mixture of Adapters

To aggregate the knowledge extracted by adapters corresponding to attributes, objects, and compositions, we propose Mixture-of-Adapters mechanisms for both the vision side and language side of the encoder.

On the vision side, we perform a two-stage feature extraction. As shown in Fig. 2 (a), for the first $N_V - M$ layers, we extract features related to the attribute ($\mathbf{h}_A$) and the object ($\mathbf{h}_O$) through corresponding encoding blocks, which

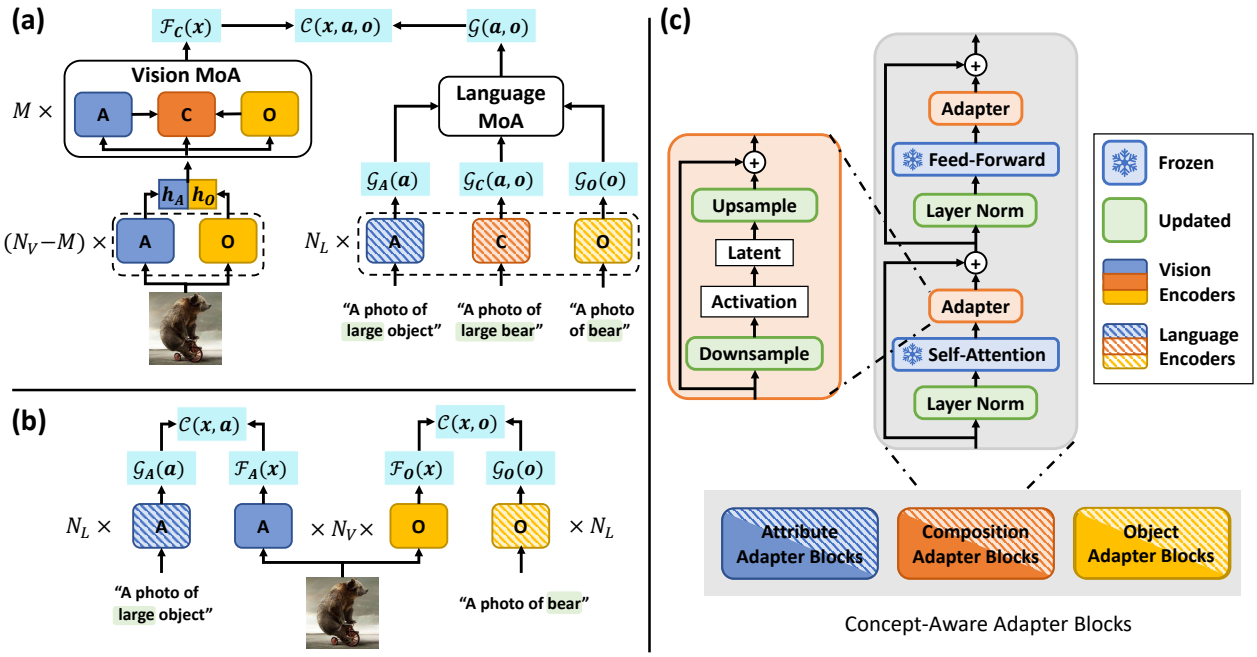


Figure 2. An overview of CAILA: (a) The main composition compatibility estimation pipeline; (b) Auxiliary sub-tasks on primitive compatibility during training; (c) The structure of CAILA layers. Our model extracts concept-specific features by learning different adapters and fuses them through the Mixture-of-Adapters (MoA) mechanism. Note that for each layer of encoders in (a) and (b), the weights of encoding blocks of the same concept are shared. N_V, N_L and M indicate numbers of layers.

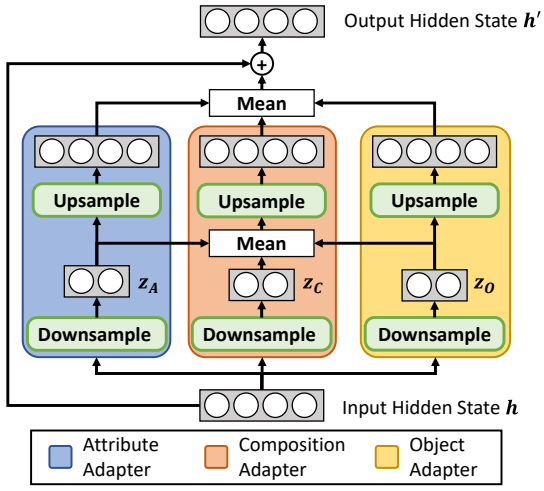


Figure 3. Details of our vision Mixture-of-Adapter. Latent features of each adapter, $\mathbf{z_A}, \mathbf{z_O}, \mathbf{z_C}$, are mixed, and further processed by the upsampling function to generate $\mathbf{h'_C}$. $\mathbf{h'_C}$ is joined with $\mathbf{h'_A}, \mathbf{h'_O}$ and input feature $\mathbf{h}$ for output.

are further concatenated and processed by the trailing M ternary MoA layers. An example of the vision MoA layer is shown in Fig. 3. Given the hidden state h, we extract latent features $\mathbf{z_A}$, $\mathbf{z_O}$ and $\mathbf{z_C}$ from the adapters. We then combine all three features and create $\mathbf{z'_C}$, followed by f_{Up}:

$$\mathbf{z'_C} = \text{Avg}\big[\mathbf{z_A}, \mathbf{z_O}, \mathbf{z_C}\big], \quad \mathbf{h'_C} = f_{Up}(\mathbf{z'_C}). \tag{4}$$

We further combine $\mathbf{h'_C}$ with outputs of attribute and object adapters, $\mathbf{h'_A}$ and $\mathbf{h'_O}$, to create the output:

$$\mathbf{h'} = \text{Avg}\big[\mathbf{h'_A}, \mathbf{h'_O}, \mathbf{h'_C}\big] + \mathbf{h}. \tag{5}$$

The output of the last mixture layer is L2-normalized and adopted as $\mathcal{F}_C(x)$ for compatibility estimation. Ablation study on this module is discussed in Sec. 4.3.

Unlike the vision side, where attributes and objects are deeply entangled within the same input image. On the language side, we can create disentangled language inputs through different prompt templates for attributes and objects separately. Thus, we adopt a simple mixture strategy for language adapters. We compute the compositional embedding through N_L encoding blocks for the composition and combine it with primitive language embeddings:

$$\mathcal{G}(a, o) = \text{Avg}\big[\mathcal{G}_A(a), \mathcal{G}_O(o), \mathcal{G}_C(a, o)\big]. \tag{6}$$

3.4. Primitive Concept Shift on Image Embeddings

Due to the limited diversity of training data, current CZSL models often suffer from training biases. As discussed in Sec. 3.3, in addition to the composition-related feature, CAILA extracts attribute- and object-oriented features during the first stage of $\mathcal{F}_C$. That motivates us to leverage these primitive-specific features to create additional embeddings for certain compositions. As it leads to changes in

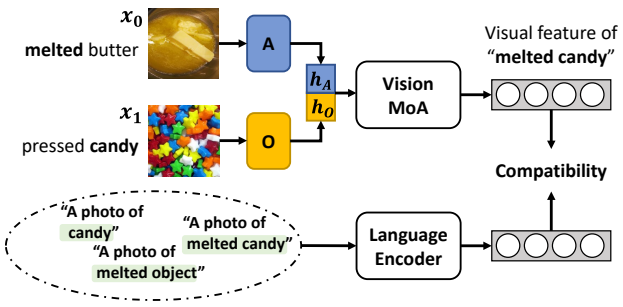


Figure 4. Illustrations of concept shift. We perform concept shift by combining the attribute (**melted**) feature from one image with the object (**candy**) feature to create an additional composition (**melted candy**) feature. Newly generated features are shuffled with regular samples during training.

labels of original images, *e.g.* from `melted butter` to `melted candy`, we call it *primitive concept shift*.

Fig. 4 demonstrates the process of concept shift: Given one sample x_0 of `melted butter` and one sample x_1 of `pressed candy`, we create a new sample of `melted candy` in the feature space, by combining the attribute-oriented feature $\mathbf{h}_A$ of x_0 and the object-oriented feature $\mathbf{h}_O$ of x_1. The newly combined feature is further processed by vision MoA layers described in Sec. 3.3, leading to an embedding representing `melted candy`. Such change can be viewed as an "object shift" from `melted butter` or an "attribute shift" from `pressed candy`. Thus, we name this process "primitive concept shift". In practice, we randomly pick a proportion of samples for shifting and ensure that the new label after shifting still lies in the training set. We discuss the effectiveness of the shifting in Sec .4.3.

Although there are previous explorations [19,40] in generating novel features in the latent space, our method is essentially novel from two aspects: i) Wei *et al.* [40] generate features directly from word embeddings, while our method leverages disentangled vision features that have richer and more diverse knowledge; ii) Li *et al.* [19] uses generated features to augment primitive vision encoders, while ours augments the entire model through CAILA for both compositions and individual primitives.

3.5. Training and Testing

Objective. We optimize our model with a main loss on attribute-object compositions and auxiliary losses on attributes and objects. As our model only has access to seen compositions Y_s, we create our training objective upon Y_s and ignore other compositions during training. More specifically, given an image x, we compute the compatibility score $\mathcal{C}(x,a,o), \mathcal{C}(x,a)$ and $\mathcal{C}(x,o)$ for all $(a,o) \in Y_s$. We then jointly optimize $\mathcal{F}$ and $\mathcal{G}$ by the cross-entropy loss with temperature:

$$\mathcal{L} = \frac{-1}{|\mathcal{T}|}\sum_i \Bigg\{ \log \frac{e^{[\mathcal{C}(x_i,a_i,o_i)/\tau_C]}}{\sum_j e^{[\mathcal{C}(x_i,a_j,o_j)/\tau_C]}} + \log \frac{e^{[\mathcal{C}(x_i,a_i)/\tau_A]}}{\sum_j e^{[\mathcal{C}(x_i,a_j)/\tau_A]}} + \log \frac{e^{[\mathcal{C}(x_i,o_i)/\tau_O]}}{\sum_j e^{[\mathcal{C}(x_i,o_j)/\tau_O]}} \Bigg\}. \tag{7}$$

Intuitively, the cross-entropy loss will force the model to produce a higher compatibility score when (x,a,o) matches and lower the score when a non-label composition occurs.

Inference. The generalized CZSL task requires models to perform recognition over a joint set of seen and unseen compositions. Thus, for each test sample x, we estimate the compatibility score between x and every candidate (a,o) inside the search space $\mathcal{Y}_s \cup \mathcal{Y}_u$. We predict the image x as the composition that has the highest compatibility score:

$$\hat{y} = \underset{(a,o)\in\mathcal{Y}_s\cup\mathcal{Y}_u}{\arg\max}\ \mathcal{C}(x,a,o) \tag{8}$$

We apply the prediction protocol to all benchmarks.

4. Experiments

4.1. Experiment Settings

Datasets. We evaluate CAILA on four popular datasets: MIT-States [9], C-GQA [25], UT-Zappos [44, 45] and VAW-CZSL [35]. For splits, we follow [25] for C-GQA, [35] for VAW-CZSL, and [31] for MIT-States/UT-Zappos. Statistically, the numbers of images in train/val/test are 29k/10k/10k for MIT-States, 23k/3k/3k for UT-Zappos, 26k/7k/5k for C-GQA, and 72k/10k/10k for VAW-CZSL.

Scenarios. We perform evaluation of CZSL models on both *closed* and *open* world scenarios and denote them as ● and ○, respectively. Regarding the *closed world* setting, we follow [1, 25, 31] and conduct CZSL with a limited search space. We further run models in the *open* world scenario, proposed by Mancini *et al.* [22], to assess the scalability of CZSL models. It is worth noting that C-GQA becomes much more challenging under the *open world* setting, as the size of the search space drastically increases from 2k to nearly 400k. We also notice similar space expansions on MIT-States, while the number of possible compositions does not increase much on UT-Zappos.

Evaluation Metrics. Our evaluation follows the generalized CZSL protocol adopted by [1, 22, 25, 31]. [31, 41] argue that it is unreasonable to evaluate only Y_u as significant biases enter during training and model selection. They suggest computing both seen and unseen accuracy with various bias values added to unseen categories and taking the Area Under the Curve (AUC) as the core metric. We select our models with the best AUC on *val* sets and report performance on *test* sets.

	Closed World Model	● MIT-States AUC (↑)	HM (↑)	S (↑)	U (↑)	● C-GQA AUC (↑)	HM (↑)	S (↑)	U (↑)	● UT-Zappos AUC (↑)	HM (↑)	S (↑)	U (↑)
Without CLIP	CompCos [22]	4.5	16.4	25.3	24.6	2.6	12.4	28.1	11.2	28.7	43.1	59.8	62.5
	ProtoProp [34]	-	-	-	-	3.7	15.1	26.4	18.1	34.7	50.2	62.1	65.7
	OADis [35]	5.9	18.9	31.1	25.6	-	-	-	-	30.0	44.4	59.5	65.5
	SCEN [19]	5.3	18.4	29.9	25.2	2.9	12.4	28.9	12.1	32.0	47.8	63.5	63.1
	CGE [25]	6.5	21.4	32.8	28.0	4.2	15.5	33.5	16.0	33.5	**60.5**	64.5	71.5
	Co-CGE [23]	6.6	20.0	32.1	28.3	4.1	14.4	33.3	14.9	33.9	48.1	62.3	66.3
	CAPE [14]	6.7	20.4	32.1	28.0	4.6	16.3	33.0	16.4	35.2	49.5	62.3	68.5
With CLIP	CLIP-ZS [32]	11.0	26.1	30.2	46.0	1.4	8.6	7.5	25.0	5.0	15.6	15.8	49.1
	CoOp [48]	13.5	29.8	34.4	47.6	4.4	17.1	26.8	20.5	18.8	34.6	52.1	49.3
	Co-CGE† [23]	17.0	33.1	46.7	45.9	5.7	18.9	34.1	21.2	36.3	49.7	63.4	71.3
	CSP [28]	19.4	36.3	46.6	49.9	6.2	20.5	28.8	26.8	33.0	46.6	64.2	66.2
	DFSP [21]	20.6	37.3	46.9	52.0	10.5	27.1	38.2	32.9	36.0	47.2	66.7	71.7
	CAILA (Ours)	**23.4**	**39.9**	**51.0**	**53.9**	**14.8**	**32.7**	**43.9**	**38.5**	**44.1**	57.0	**67.8**	**74.0**

Table 1. Quantitative results on generalized CZSL in *closed world*, all numbers are reported in percentage. S and U refer to best seen and unseen accuracy on the accuracy curve. CLIP-ZS refers to the vanilla CLIP model without fine-tuning. All CLIP-based models are run with ViT-L/14 and we conduct extensive experiments in Tab. 4. †We run Co-CGE with similar CLIP features and report our best number of the model. Models published before CGE are omitted as their performances are inferior to current baselines.

	Closed World Model	● VAW-CZSL AUC (↑)	HM (↑)	S (↑)	U (↑)
Without CLIP	CompCos [22]	5.6	14.2	23.9	18.0
	OADis [35]	6.1	15.2	24.9	18.7
	CGE [25]	5.1	13.0	23.4	16.8
With CLIP	CLIP-ZS [32]	2.6	11.9	12.8	27.8
	CSP [28]	8.5	23.3	31.9	33.6
	DFSP [21]	14.1	31.1	40.1	40.9
	CAILA (Ours)	**17.2**	**34.6**	**41.6**	**49.2**

Table 2. Quantitative results on generalized CZSL of VAW-CZSL in closed world, all numbers are reported in percentage.

Furthermore, best-seen accuracy and best-unseen accuracy are calculated when other candidates are filtered out by specific bias terms. We also report best *Harmonic Mean* (HM), defined as $(2 * seen * unseen)/(seen + unseen)$.

Implementation Details: We build our model on the PyTorch [29] framework. As for optimization, we use Adam optimizer with a weight decay of $5e-5$. The learning rate is set to $2e-5$. The batch size is set to 32 for all three datasets. The temperature τ_C, τ_A, τ_O is set to 0.01, 0.0005 and 0.0005, respectively. Most of the experiments are run on two NVIDIA A100 GPUs. We the number of vision MoA layers M to 6 by default. For the downsampling function f_{Down}, we set the reduction factor to 4. Ablation studies on these settings can be found in Sec 4.3.

4.2. Quantitative Results

In this section, we present quantitative results in detail under both *closed world* and *open world* settings. Such results verify the effectiveness of our method, which surpasses the current SOTA on most metrics, in both scenarios.

Closed World Results. Performance of the *closed world* scenario are reported in Tab. 1 and 2. On MIT-States, results show that CAILA overcomes the label noise and achieves SOTA. More specifically, on AUC, we observe a 2.8% improvement, from 20.6% to 23.4%. Furthermore, regarding HM, CAILA achieves 39.9%, outperforming all baselines. When it comes to best seen and unseen accuracy, our model improves by ∼4% and ∼2%, respectively.

Our results on C-GQA further verify the advantage of CAILA, especially when the number of unseen compositions is larger. On AUC, our model achieves a 4.3% improvement, 40% of the previous SOTA, from 10.5% to 14.8%. HM is also improved by 5.6%. Moreover, improvements of best seen and unseen accuracy are 5.7% and 5.6%.

UT-Zappos has much fewer attributes and object categories, compared with its counterparts, and is thus much easier, as the gap between various methods is smaller. But it is noticeable that our model, CAILA , outperforms all other baselines, with a 7.2% improvement on the AUC metric.

Moreover, on the recently released benchmark, VAW-CZSL, CAILA is able to achieve noticeable improvements against baseline models, particularly the newly published method, DFSP [21]. CAILA improves the AUC by 3.1% while boosting the harmonic mean by 3.5%.

Open World Results. We further conduct experiments under the *open world* setting to evaluate the robustness of CAILA . Results are shown in Tab. 3. Noticeably, *open world* is much harder than *closed world*, as performance on all benchmarks drops drastically, while CAILA achieves SOTA on most metrics in this scenario without any filtering techniques adopted in the previous papers [22, 23, 28].

On MIT-States, our approach greatly beats SOTA on all metrics, particularly the AUC. Our model improves AUC from 6.8% to 8.2% and achieves a 21.6% harmonic

	Open World Model	○ MIT-States				○ C-GQA				○ UT-Zappos			
		AUC (↑)	HM (↑)	S (↑)	U (↑)	AUC (↑)	HM (↑)	S (↑)	U (↑)	AUC (↑)	HM (↑)	S (↑)	U (↑)
Without CLIP	CompCos [22]	0.8	5.8	21.4	7.0	0.43	3.3	26.7	2.2	18.5	34.5	53.3	44.6
	CGE [25]	1.0	6.0	32.4	5.1	0.47	2.9	32.7	1.8	23.1	39.0	61.7	47.7
	KG-SP [13]	1.3	7.4	28.4	7.5	0.78	4.7	31.5	2.9	26.5	42.3	61.8	52.1
	Co-CGE^{CW} [23]	1.1	6.4	31.1	5.8	0.53	3.4	32.1	2.0	23.1	40.3	62.0	44.3
	Co-CGE^{open} [23]	2.3	10.7	30.3	11.2	0.78	4.8	32.1	3.0	23.3	40.8	61.2	45.8
With CLIP	CLIP-ZS [32]	3.0	12.8	30.1	14.3	0.27	4.0	7.5	4.6	2.2	11.2	15.7	20.6
	CoOp (a) [48]	4.7	16.1	36.8	16.5	0.73	5.7	20.9	4.5	19.5	35.6	61.8	39.3
	CoOp (b) [48]	2.8	12.3	34.6	9.3	0.70	5.5	21.0	4.6	13.2	28.9	52.1	31.5
	Co-CGE† [23]	5.6	17.7	38.1	20.0	0.91	5.3	33.2	3.9	28.4	45.3	59.9	56.2
	CSP [28]	5.7	17.4	46.3	15.7	1.20	6.9	28.7	5.2	22.7	38.9	64.1	44.1
	DFSP [21]	6.8	19.3	47.5	18.5	2.40	10.4	38.3	7.2	30.3	44.0	66.8	**60.0**
	CAILA (Ours)	**8.2**	**21.6**	**51.0**	**20.2**	**3.08**	**11.5**	**43.9**	**8.0**	**32.8**	**49.4**	**67.8**	59.7

Table 3. Quantitative results on generalized CZSL in *open world*, all numbers are reported in percentage. S and U refer to best seen and unseen accuracy on the curve. CLIP-ZS refers to the vanilla CLIP model without fine-tuning. All CLIP-based models are run with ViT-L/14. Note that our models tested have identical weights as in Tab. 1. †We run Co-CGE with similar CLIP features and report our best number of the model. Models published before CGE are omitted as their performances are inferior to current baselines.

Image Encoder	Closed World Model	●MIT-States	●C-GQA	●UT-Zappos
ViT B/32	CLIP-ZS* [32]	7.5	1.2	2.4
	CLIP-FT [32]	10.9	7.6	21.1
	Co-CGE† [23]	12.2	5.0	31.2
	CSP* [28]	12.4	5.7	24.2
	DFSP [21]	13.2	-	23.3
	CAILA (Ours)	**16.1**	**10.4**	**39.0**
	Δ	+2.9 *(21.9%)*	+2.8 *(36.8%)*	+7.8 *(25.0%)*
ViT L/14	CLIP-ZS* [32]	11.0	1.4	5.0
	CLIP-FT* [32]	14.4	10.5	4.8
	CoOp* [48]	13.5	4.4	18.8
	CLIP-Adapter* [5]	9.5	3.2	31.5
	Co-CGE† [23]	17.0	5.7	36.3
	CSP* [28]	19.4	6.2	33.0
	DFSP [21]	20.6	10.5	36.0
	CAILA (Ours)	**23.4**	**14.8**	**44.1**
	Δ	+2.8 *(13.6%)*	+4.3 *(41.0%)*	+7.8 *(21.5%)*

Table 4. Comparison of the AUC performance on all three benchmarks among CLIP-based models. ZS and FT stand for zero-shot and fine-tuned. Best results are shown in **bold** and runner-ups are underlined. Δ is calculated between CAILA and the second-best. Numbers with * are acquired from the CSP paper [28]. †We obtain these numbers by running Co-CGE on similar CLIP features.

mean. Moreover, CAILA achieves improves seen accuracy by 3.5% and unseen accuracy by 0.2%.

The performance of CAILA on C-GQA in the *open world* scenario is consistent with the one in *closed world*. More specifically, our model achieves 3.08% AUC, 128% of DFSP [21]. We also observe a ∼10% relative improvement on harmonic mean, from 10.4% to 11.5%. CAILA achieves 5.6% and 0.8% boosts on seen and unseen.

Regarding UT-Zappos, our model also brings in performance gains. It achieves a 49.4% harmonic mean, 4.1% higher than Co-CGE. CAILA also gets the best AUC of 32.8%, at least 2% higher against other baselines.

Comparisons between CLIP-based Methods. We further make head-to-head comparisons between CAILA and other approaches built with CLIP in Tab. 4, with variations on the vision encoder: ViT-B/32 [4] and ViT-L/14. Results verify CAILA 's effectiveness and consistency with different visual backbones. In particular, CAILA achieves >35% relative improvements on C-GQA against other baselines.

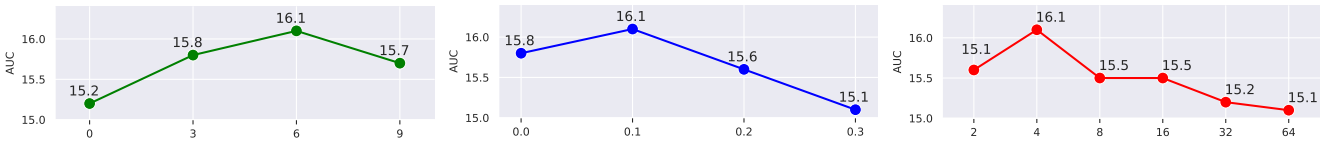
Discussion. Given that CLIP is trained on a web-scale dataset, ensuring fair comparisons between CLIP-based [21,23,28] and CLIP-free methods [13,22,25] can be difficult, particularly as CLIP-based methods significantly outperform CLIP-free ones. We follow the setting in existing CLIP-based methods [21, 23, 28, 32, 48] with a focus on enhancing CLIP-based CZSL. Comparisons between CAILA and fine-tuned CLIP models show that a partially tuned model can beat its fully fine-tuned counterpart by a large margin, justifying that CAILA better suppresses training biases while remaining sharp on knowledge transfer for CZSL, thus is a better way to exploit CLIP knowledge.

4.3. Ablation Studies

We conduct the ablation study with CLIP ViT-B/32 and MIT-States in *closed world*.

Adapter and MoA. We evaluate different adapter/MoA settings on MIT-States and report results in Tab. 5. We observe that compared with CSP [28], adding adapters to either side of encoders can effectively improve the performance while attaching adapters to both sides shows further improvements. Experiments in the last three rows verify that our Mixture-of-Adapters mechanism further improves the performance when it is applied on both sides.

Vision Mixture Strategies. We compare different ways of mixing $\mathbf{z}$ and $\mathbf{h}'$ inside the vision MoA layer as described in Eqn. 4,5. Tab. 6 shows the results of mixing only one of the feature vectors or none at all. The last row corre-



(a) Number of Vision MoA Layers M (b) Concept Shift Ratio (c) Reduction Factor

Figure 5. Ablation studies: (a) The number of vision MoA layer M; (b) The ratio of concept shift; (c) The reduction factor of f_{Down}.

Adapter V	Adapter L	MoA V	MoA L	● MIT-States AUC (↑)	HM (↑)	S (↑)	U (↑)
CSP [28]				12.4	28.6	36.4	42.5
✓				14.0	30.1	41.4	42.0
	✓			13.9	30.5	40.3	42.8
✓	✓			14.4	30.7	42.2	43.2
✓	✓	✓		15.4	31.4	**43.4**	44.5
✓	✓		✓	15.2	31.7	41.6	44.8
✓	✓	✓	✓	**16.1**	**32.9**	43.3	**45.6**

Table 5. Ablation on adapters and MoA modules. V and L refer to Vision and Language, respectively.

Closed World Model	Mixture $\mathbf{z}$	Mixture $\mathbf{h}'$	● MIT-States AUC (↑)	HM (↑)	S (↑)	U (↑)
CAILA (Ours)	✓	✓	**16.1**	**32.9**	**43.3**	**45.6**
	✓		15.8	32.2	43.3	45.2
		✓	15.5	31.7	43.0	45.1
			15.5	32.0	42.7	44.8

Table 6. Ablation on vision MoA strategies.

sponds to averaging $\mathcal{F}_A(x), \mathcal{F}_O(x), \mathcal{F}_C(x)$ without intra-layer mixture, which is similar to the language side MoA. Experiment results demonstrate that mixing both $\mathbf{z}$ and $\mathbf{h}'$ as proposed in Sec. 3.3 yields optimal performance while applying a similar strategy as the language side hurts.

Vision Mixture Functions. We evaluate various mixture functions of vision MoA besides the default mean function, including summation (Sum.), element-wise multiplication (Mul.), and concatenation (Concat.). We add one linear layer after "Concat" to align the feature dimension with upcoming operations. Results in Tab. 7 show that the "Mean" operation performs the best. We also notice that the variation with "Sum." performs worse, possibly because summation greatly changes the magnitude of the feature vector.

Learnable Prompts. We perform experiments to study the effect of learnable prompts in our framework. Results reported in Tab. 8 show that our model remains competitive with prompt embeddings fixed. Such behavior justifies that performance gains of CAILA come from designs that have been discussed in the Approach section.

CAILA Setups. We explore different aspects of our setup and show the results in Fig. 5. Fig. 5(a) demonstrates that CAILA performs better with MoA layers and achieves the best performance with 6 MoA layers on the vision side,

Closed World Model	Mix. Fn.	● MIT-States AUC (↑)	HM (↑)	S (↑)	U (↑)
CAILA (Ours)	Mean	**16.1**	**32.9**	**43.3**	**45.6**
	Sum.	14.6	30.7	42.8	42.1
	Mul.	15.8	32.2	43.3	45.0
	Concat.	15.2	31.9	41.8	44.8

Table 7. Ablation on vision MoA mixture functions.

Closed World Model	● MIT-States AUC (↑)	HM (↑)	S (↑)	U (↑)
CAILA(Ours)	**16.1**	**32.9**	43.3	**45.6**
w/o Learnable Prompts	15.8	32.1	**43.5**	44.6
DFSP [21]	13.2	29.4	36.7	43.4
CSP [28]	12.4	28.6	36.4	42.5

Table 8. Ablation study on learnable prompts.

which is also better than the single-stage MoA when M=0; Fig. 5(b) indicates that replacing 10% of a batch with post-shift features can increase the AUC while adding more shift reduces it; In Fig. 5(c), we find that the optimal reduction factor for the latent feature $\mathbf{z}$ is 4, while using higher reduction factors does not affect the performance significantly and can be considered for efficiency reasons.

5. Conclusion

In this paper, we explore the problem of how to leverage large-scale Vision-Language Pre-trained (VLP) models, particularly CLIP, more effectively for compositional zero-shot learning. Unlike previous methods which treat CLIP as a black box, we propose to slightly modify the architecture and attach Concept-Aware Intra-Layer Adapters (CAILA) to each layer of the CLIP encoder to enhance the knowledge transfer from CLIP to CZSL. Moreover, we design the mixture-of-adapters mechanism to further improve the generalizability of the model. Quantitative evaluations demonstrate that CAILA achieves significant improvements on all three common benchmarks. Due the lack of unfeasible pair filter, CAILA's performance drops from closed world to open world, when the number of possible pairs greatly increases, though. We also provide comprehensive discussions on deciding the optimal setup.

Acknowledgment

This research was supported, in part, by the Office of Naval Research under grant #N00014-21-1-2802.

References

[1] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *arXiv preprint arXiv:2006.14610*, 2020. 1, 2, 3, 5

[2] Shiming Chen, Wenjie Wang, Beihao Xia, Qinmu Peng, Xinge You, Feng Zheng, and Ling Shao. Free: Feature refinement for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 122–131, 2021. 2

[3] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 7

[5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1, 2, 3, 7

[6] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14105–14115, 2022. 2

[7] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[8] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2

[9] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 1, 2, 5

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

[11] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P Xing. Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11487–11496, 2019. 2

[12] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035, 2021. 2

[13] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2022. 7

[14] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Learning attention propagation for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3828–3837, 2023. 6

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2

[16] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013. 2

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

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

[19] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9335, 2022. 1, 2, 5, 6

[20] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325, 2020. 1, 2

[21] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023. 6, 7, 8

[22] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5222–5230, 2021. 3, 5, 6, 7

[23] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 6, 7

[24] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1792–1801, 2017. 1, 2, 3

[25] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 1, 2, 3, 5, 6, 7

[26] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 1, 2

[27] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8811–8818, 2019. 1, 2

[28] Nihal V. Nayak, Peilin Yu, and Stephen Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 6, 7, 8

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[31] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019. 1, 2, 3, 5

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7

[33] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016. 2

[34] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6

[35] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022. 1, 2, 5, 6

[36] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *arXiv preprint arXiv:1301.3666*, 2013. 2

[37] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 2

[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] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2019. 1, 2

[40] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3741–3749, 2019. 1, 2, 5

[41] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 3, 5

[42] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018. 2

[43] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10256, 2020. 1, 2

[44] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *Computer Vision and Pattern Recognition (CVPR)*, Jun 2014. 2, 5

[45] A. Yu and K. Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 5

[46] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017. 2

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 6, 7

[49] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 1004–1013, 2018. 2