Hierarchical Prompt Learning for Compositional Zero-Shot Recognition

Henan Wang , **Muli Yang*** , **Kun Wei** and **Cheng Deng***
School of Electronic Engineering, Xidian University, Xi'an, China
{nnhhwang, muliyang.xd, weikunsk, chdeng.xd}@gmail.com

Abstract

Compositional Zero-Shot Learning (CZSL) aims to imitate the powerful generalization ability of human beings to recognize novel compositions of known primitive concepts that correspond to a state and an object, *e.g.*, "purple apple". To fully capture the intra- and inter-class correlations between compositional concepts, in this paper, we propose to learn them in a hierarchical manner. Specifically, we set up three hierarchical embedding spaces that respectively model the states, the objects, and their compositions, which serve as three "experts" that can be combined in inference for more accurate predictions. We achieve this based on the recent success of large-scale pretrained Vision-Language Models, *e.g.*, CLIP, which provides a strong initial knowledge of image-text relationships. To better adapt this knowledge to CZSL, we propose to learn three hierarchical prompts by explicitly fixing the unrelated word tokens in the three embedding spaces. Despite its simplicity, our proposed method consistently yields superior performance over current state-of-the-art approaches on three widely-used CZSL benchmarks.

1 Introduction

When someone shows a photo of `purple apple`, even though you may have never seen one, it should be easy for you to immediately recognize it based on the common knowledge of how `purple` and `apple` respectively look like. This ability of *compositional generalization* [Atzmon *et al.*, 2016] is one of the key differences between human beings and other creatures, which underlies the human intelligence by composing and understanding new things based on known knowledge. Therefore, compositional generalization is also deemed as a holy grail in artificial intelligence, attracting decades of devotion from a large number of researchers [Johnson *et al.*, 2017; Hudson and Manning, 2018].

In light of this, Compositional Zero-Shot Learning (CZSL) [Misra *et al.*, 2017] has emerged as a promising test bed of compositional generalization. Specifically, in CZSL, the content in each image sample is regarded as a combination of two primitive concepts — a state and an object, *e.g.*, `red apple` and `purple eggplant`. In inference, the model is expected to recognize unseen combinations of known primitives (*e.g.*, `purple apple`) after trained on other seen combinations (*e.g.*, `red apple` and `purple eggplant`). The challenge of CZSL lies in the entanglement of the two primitive concepts — states and objects are highly dependent on each other in visual contents, making it considerably difficult to generalize to unseen combinations.

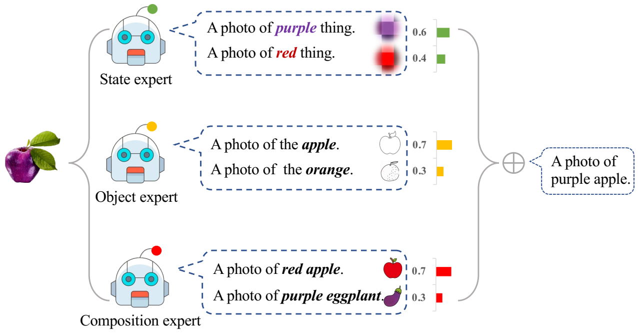


Figure 1: The illustration of our motivation. Each robot represents an expert with different expertise. We can obtain more accurate predictions for compositional concepts by fully leveraging the complementary ability of each expert.

Some existing attempts [Purushwalkam *et al.*, 2019; Li *et al.*, 2020] respectively train two separate classifiers for states and objects, aiming to directly predict each of them based on visual features. Due to the lack of consideration of the aforementioned entanglement, these methods are not able to capture discriminative enough state and object features, resulting in limited recognition performance. In contrast, other methods [Nagarajan and Grauman, 2018; Nan *et al.*, 2019] attempt to project visual features and state-object embeddings into a shared embedding space, aiming to pull closer associated visual and text embeddings using metric learning techniques such as cosine similarity. Despite showing effective in capturing the entanglement between states and objects, however, these methods are prone to over-fit to seen compositions, which may produce ambiguous predictions for unseen compositions in inference, limiting the overall CZSL performance. As a result, most exist-

*Corresponding authors.

ing methods may suffer when handling varying compositional concepts due to the lack of comprehensive modeling of the intriguing state-object relationships.

In this paper, we propose to combine the strengths of existing methods by learning the compositional concepts in a hierarchical manner. Specifically, we set up three hierarchical embedding spaces that respectively model the states, the objects, and their combinations, which serve as three "experts" that can be integrated in inference for more accurate predictions. As shown in Fig. 1, given an image of `purple apple`, the composition expert may prefer to predict it as `red apple` or `purple eggplant` due to the entanglement in {`red`,`apple`} and {`purple`,`eggplant`} since these samples are more common in training. In this regard, the state expert can be of help thanks to its object-agnostic expertise — it can easily identify the `purple` concept in the given image, which is beneficial to correcting the bias in pair expert predictions. Likewise, the object expert can also help a more accurate object prediction; with their union, each expert plays a complementary role to derive a more accurate answer. There are also cases when the state and the object experts cannot produce faithful predictions due to the ambiguity in compositional concepts. In these cases, the pair expert can instead help to discern the compositional concepts as a whole. We show in experiments that the integration of the three experts always makes the best of the overall CZSL performance.

We achieve this based on the recent success of large Vision-Language Models (VLMs), *e.g.*, CLIP [Radford *et al.*, 2021], which provides a strong initial knowledge of image-text relationships. To better adapt this knowledge to CZSL, we propose to learn three hierarchical prompts in the three embedding spaces based on our above motivation, by explicitly fixing the unrelated word tokens. For example, in the state embedding space, we learn the prompt with the template of `A photo of [state] thing` in which `[state]` is filled with the state label while other word tokens are kept fixed. After training, the learned hierarchical prompts are ready to use in the three embedding spaces, which can be integrated to achieve more accurate CZSL performance.

In summary, our main contributions are

1. We construct three hierarchical embedding spaces at different levels, which serve as three "experts" who play complementary and comprehensive roles to help a better CZSL performance.
2. We propose to learn three hierarchical prompts by explicitly fixing the unrelated word tokens in the three embedding spaces to adapt the strong knowledge of pretrained VLMs.
3. Extensive experiments demonstrate the effectiveness of our method, which shows clear superiority over the state-of-the-art approaches in both closed-world and open-world settings.

2 Related Work

Compositional Zero-Shot Learning. The task of Compositional Zero-Shot Learning (CZSL) [Misra *et al.*, 2017; Purushwalkam *et al.*, 2019; Wei *et al.*, 2019; Yang *et al.*, 2020; Huynh and Elhamifar, 2020; Xu *et al.*, 2021] is to recognize images of novel primitive compositions that are absent during the training stage, which is a special case of Zero-Shot Learning (ZSL) task [Wei *et al.*, 2020; Yang *et al.*, 2021; Li *et al.*, 2021].

Some recent works utilize two separate classifiers to recognize states and objects respectively, and then combine them for the final predictions [Li *et al.*, 2020; Saini *et al.*, 2022; Karthik *et al.*, 2022; Li *et al.*, 2022]. However, these works consider simple primitives as independent probability distributions, ignoring the inherent dependence among states, objects and compositions. Existing studies learn a joint embedding space using separate encoders for states and objects, and then combine them with a linear layer or a multilayer perceptron to achieve the alignment with the images. CGE [Naeem *et al.*, 2021] establishes the dependencies between states and objects using a Graph Convolutional Neural (GCN) network. Moreover, in the open-world setting, there exist a huge amount of unreasonable compositions from arbitrary combinations of states and objects. Some works address this challenge by using external knowledge to filter out infeasible compositions [Mancini *et al.*, 2021; Karthik *et al.*, 2022; Mancini *et al.*, 2022]. Besides, because of the ability of large-scale pretrained Vision-Language Models (VLMs) for representing arbitrary classes as natural language prompts in their flexible text encoders, [Nayak *et al.*, 2023] first attempted to learn soft state and object tokens in a language prompt, while overlooking the individual roles of states and objects. Inspired by [Yang *et al.*, 2022], our proposed method combines the advantages of the previous methods by learning more comprehensive representations for compositional concepts in a hierarchical manner.

Prompt Learning. In the past few years, the "pretraining and fine-tuning" paradigm [Peters *et al.*, 2018; Dong *et al.*, 2019; Yang *et al.*, 2019] plays an important role in Natural Language Processing (NLP) tasks. As models become larger, storing and serving a tuned copy of the model for each downstream task becomes impractical. Nowadays, the "pretraining and fine-tuning" procedure is replaced by the one we dub "pretraining, prompting, and predicting". In prompt learning [Bommasani *et al.*, 2021; Sanh *et al.*, 2022; Vu *et al.*, 2022; Zhou *et al.*, 2022; Bach *et al.*, 2022], downstream tasks are reformulated to adapt to the form of original pretrained models with the help of a textual prompt. Prompting renders the pretrained model and downstream tasks closer by tuning only the input with a specific template, which is different from fine-tuning which requires to update the model parameters.

Benefiting from the multi-modal knowledge of VLMs which are pretrained on large-scale datasets, prompt learning achieves outstanding performance across a wide range of tasks, under the zero-shot and few-shot settings [Qin and Eisner, 2021; Radford *et al.*, 2021]. Also, because of the successful applications of CLIP [Radford *et al.*, 2021], prompt learning is also of great interest in computer vision. Unlike the discrete text prompts used by GPT-3 [Brown *et al.*, 2020], which is time-consuming and impractical to find the best choice, soft prompting can achieve this through back-propagation

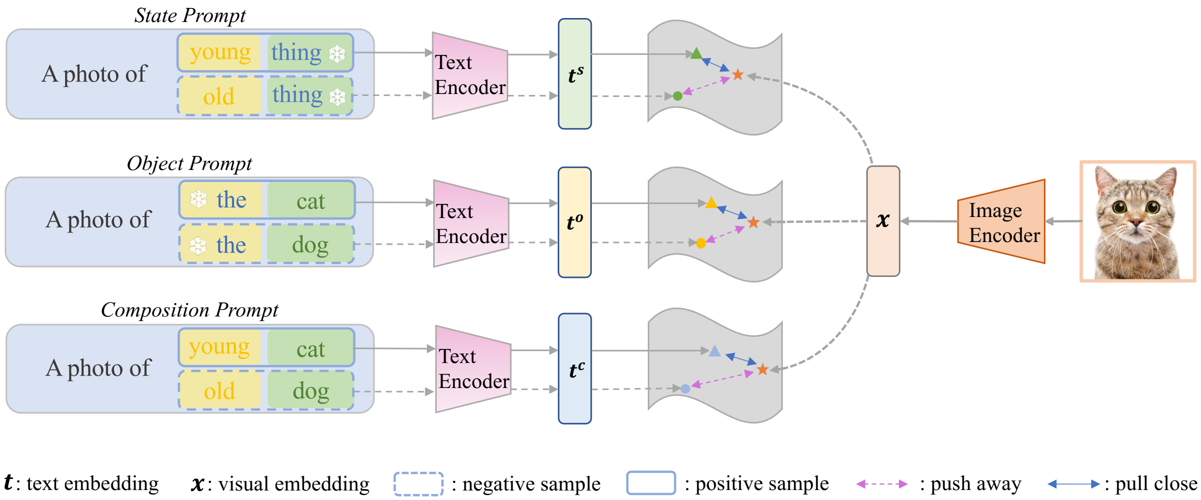


Figure 2: Overview of our framework. We learn three hierarchical prompts with different fixed word tokens for each compositional concept. We embed the three prompts using the text encoder from a large-scale pretrained Vision-Language Model (VLM), and pull close the visual embedding from the associated image which is also encoded using the same VLM.

and shows its surprising performance without fine-tuning the entire models. Context Optimization (CoOp) [Zhou *et al.*, 2022] demonstrated that CLIP's performance is susceptible to prompts and soft prompting can improve performance in the image recognition task. In light of this, we learn three hierarchical soft prompts by explicitly fixing the unrelated word tokens in the three embedding spaces to adapt the strong knowledge of pretrained VLMs.

3 Approach

In Compositional Zero-Shot Learning (CZSL), each image is composed of two primitive concepts, *i.e.*, a state and an object, and the model needs to be trained on a set of seen state-object combinations and tested on another set of unseen combinations. In this paper, we propose Hierarchical Prompt Learning (HPL), *i.e.*, learning hierarchical prompts for compositional concepts in different levels. We start with the problem definition of CZSL, followed by the detailed formulation of our proposed HPL, the overall framework of which is shown in Fig. 2.

3.1 Problem Definition

We use $\mathcal{S} = \{s_0, s_1, \ldots, s_n\}$ and $\mathcal{O} = \{o_0, o_1, \ldots, o_m\}$ to denote the sets of states and objects respectively. Let the composition set $\mathcal{C}$ be the Cartesian product of the state set and object set, *i.e.*, $\mathcal{C} = \mathcal{S} \times \mathcal{O} = \{(s,o) \mid s \in \mathcal{S}, o \in \mathcal{O}\}$. We define two disjoint subsets of the composition set $\mathcal{C}$, namely $\mathcal{C}_{\text{seen}} \subset \mathcal{C}$ and $\mathcal{C}_{\text{unseen}} \subset \mathcal{C}$, where $\mathcal{C}_{\text{seen}}$, $\mathcal{C}_{\text{unseen}}$ represent the seen set and the unseen set, respectively, and $\mathcal{C}_{\text{seen}} \cap \mathcal{C}_{\text{unseen}} = \emptyset$.

At training time, the seen set $\mathcal{C}_{\text{seen}}$ is used for training, and the training data is defined as $\mathcal{T} = \{(x_i, c_i) \mid x \subset \mathcal{X}_{\text{seen}}, c \subset \mathcal{C}_{\text{seen}}\}$, where $\mathcal{X}_{\text{seen}}$ is a set of seen images. Each corresponding label is a tuple $c = (s, o)$ of a state $s \in \mathcal{S}$ and an object $o \in \mathcal{O}$. During inference, the unseen set $\mathcal{C}_{\text{unseen}}$ is used for testing, and the testing data can be denoted as $\mathcal{N} = \{(x_i, c_i) \mid x \subset \mathcal{X}_{\text{unseen}}, c \subset \mathcal{C}_{\text{unseen}}\}$. Following [Purushwalkam *et al.*, 2019; Chao *et al.*, 2016], we study the generalized CZSL where the test set also includes images from seen compositional labels. Other than that, we follow [Mancini *et al.*, 2021] to adopt two evaluation protocols, namely the closed-world and the open-world settings. In the close-world setting, the test set is $\mathcal{C}_{\text{test}} = \mathcal{C}_{\text{seen}} \cup \mathcal{C}_{\text{unseen}}$, while in the open-world setting, the label space contains all possible combinations, *i.e.*, $\mathcal{C}_{\text{test}} = \mathcal{C}$.

3.2 Hierarchical Embedding Spaces

In CZSL, based on above problem definition, the prediction $\hat{y}$ for a given image x is made by calculating

$$\hat{y} = (\hat{s}, \hat{o}) = \underset{(s,o)\in\mathcal{C}}{\arg\max}\; p(x|s,o)\,, \tag{1}$$

in which the model is required to learn to fit the likelihood $p(x|s,o)$. To this end, a straightforward way is to learn a shared embedding space where the associated visual features and state-object compositions are mapped close to each other. It is also feasible to learn $p(x|s)$ and $p(x|o)$ separately and combine them in inference to output state-object predictions. However, as discussed in Sec. 1, solely learning a shared embedding space for visual features and state-object compositions may lead to spurious correlations between certain states and objects, while separately modeling them, on the other hand, ignores their inherent correlation. We propose to address this dilemma by learning with three hierarchical embedding spaces that respectively capture the state, the object, and the paired concepts, which can serve as three experts with complementary expertise. After training, we can combine the three experts for inference:

$$\hat{y} = (\hat{s}, \hat{o}) = \underset{(s,o)\in\mathcal{C}}{\arg\max}\; p(x|s)\, p(x|o)\, p(x|s,o)\,. \tag{2}$$

3.3 Hierarchical Prompt Learning

We aim to better adapt large-scale pretrained Vision-Language Models (VLMs) to the task of CZSL; more specif-

			Train Set		Val Set		Test Set	
Dataset	State	Object	Pairs	Images	Pairs (S/U)	Images	Pairs (S/U)	Images
MIT-States	115	245	1,262	30,338	300/300	10,420	400/400	12,995
UT-Zappos	16	12	83	22,998	15/15	3,214	18/18	2,914
C-GQA	413	674	5,592	26,920	1,252/1,040	7,280	888/923	5,098

Table 1: Dataset details with respect to state/object numbers, pair/image numbers in seen/unseen (S/U) splits, and in val/test sets.

ically, we propose to teach them to respectively recognize the states, the objects, and the state-object compositions in a hierarchical manner. To achieve that, as shown in Fig. 2, we represent each compositional concept using three language prompts as the input to the VLM's text encoder. These prompts are all in the form of `a photo of [state] [object]`, with fixed prefix `a photo of` which is a common practice in prompt learning literature [Vu *et al.*, 2022; Zhou *et al.*, 2022]. The tokens `[state]` and `[object]` are to be filled with the compositional concept of each training sample, such as `young cat` in Fig. 2. To further encourage different expertise, we define the *state prompt* by replacing `[object]` into a fixed word `thing`, enabling the model to solely focus on the state representation. Likewise, we define the *object prompt* by replacing `[state]` into a fixed word `the`. Finally, the *composition prompt* is constructed with the original prompt template untouched.

In practice, we initialize each prompt using the pretrained word embeddings in the VLM:

$$t^i_{s,o} = \left(\boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{\theta}^i_s, \boldsymbol{\theta}^i_o\right) \ , \tag{3}$$

where $\boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_2$ correspond to the word embeddings of the fixed prefix "`a photo of`", and $\boldsymbol{\theta}^i_s, \boldsymbol{\theta}^i_o$ are learnable parameters for state and object word. Specifically, state, object, and composition prompts are initialized as $t^s_{s,o}$, $t^o_{s,o}$, and $t^c_{s,o}$ with learnable parameters $(\boldsymbol{\theta}^s_s, \boldsymbol{\theta}^s_o)$, $(\boldsymbol{\theta}^o_s, \boldsymbol{\theta}^o_o)$, $(\boldsymbol{\theta}^c_s, \boldsymbol{\theta}^c_o)$ corresponding to `[state] thing`, `the [object]`, and `[state] [object]`, respectively.

Accordingly, the three prompts are passed into the VLM's text encoder to obtain the ℓ_2-normalized text representations:

$$\boldsymbol{t}^i_{s,o} = e_t\left(t^i_{s,o}\right) / \left\| e_t\left(t^i_{s,o}\right)\right\| , \tag{4}$$

where $t^i_{s,o}$ is one of $t^s_{s,o}$, $t^o_{s,o}$, and $t^c_{s,o}$. Likewise, we can get the ℓ_2-normalized visual representation of a given image x as

$$\boldsymbol{x} = e_v\left(x\right) / \left\| e_v\left(x\right)\right\| . \tag{5}$$

Now we can optimize the learnable parameters $(\boldsymbol{\theta}^i_s, \boldsymbol{\theta}^i_o)$ by minimizing the cross-entropy loss on the training set $\mathcal{T}$ with seen compositions from $\mathcal{C}_{\text{seen}}$:

$$\mathcal{L}^i = -\frac{1}{|\mathcal{T}|} \sum_{(x,(s,o))\in\mathcal{T}} \log \frac{\exp\left(\boldsymbol{x}\cdot \boldsymbol{t}^i_{s,o}/\tau\right)}{\sum_{(s',o')\in\mathcal{C}_{\text{seen}}} \exp\left(\boldsymbol{x}\cdot \boldsymbol{t}^i_{s',o'}/\tau\right)} , \tag{6}$$

where τ is the temperature parameter.

3.4 Training and Inference

Our overall training loss is the linear combination of the three terms defined in Eq. (6):

$$\mathcal{L} = \mathcal{L}^s + \mathcal{L}^o + \mathcal{L}^c . \tag{7}$$

During training we only update the learnable parameters $(\boldsymbol{\theta}^i_s, \boldsymbol{\theta}^i_o)$ of the three prompts while keeping other model parameters fixed.

In inference, we predict the compositional concepts of a given image x by calculating

$$(\hat{s}, \hat{o}) = \underset{(s,o)\in\mathcal{C}}{\arg\max}\ \alpha\left(\boldsymbol{x}\cdot\boldsymbol{t}^s_{s,o} + \boldsymbol{x}\cdot\boldsymbol{t}^o_{s,o}\right) + (1-\alpha)\boldsymbol{x}\cdot\boldsymbol{t}^c_{s,o} , \tag{8}$$

in which $\alpha \in [0, 1]$ is a linear combination coefficient that controls the contribution of each prompt.

4 Experiments

4.1 Experimental Setup

Datasets. Our method is evaluated on three Compositional Zero-Shot Learning (CZSL) benchmark datasets, *i.e.*, MIT-States [Isola *et al.*, 2015], UT-Zappos [Yu and Grauman, 2014] and C-GQA [Naeem *et al.*, 2021]. We follow the standard split in previous works [Purushwalkam *et al.*, 2019], and the detailed information of each dataset is summarized in Tab. 1.

MIT-States is a challenging dataset containing 53,753 everyday images, *e.g.*, "young cat" and "old dog". It is annotated to a variety of classes with 115 state classes and 245 object classes. MIT-States has 1,962 compositions in total under the closed-world setting, in which 1,262 compositions are seen in training and 700 compositions are unseen.

UT-Zappos is a fine-grained dataset containing 50,025 images, primarily of various types of shoes, *e.g.*, "canvas slippers" and "rubber sandals", with 12 object classes and 16 state classes, yielding 116 plausible compositions, 83 of which are seen compositions and the rest 33 compositions are unseen.

C-GQA is a compositional version of Stanford GQA dataset [Hudson and Manning, 2019], contains 39,298 images in total, including 5,592 seen compositions and 1,932 unseen compositions. It contains 413 state classes and 674 object classes.

Evaluation Metrics. We evaluate the performance according to prediction accuracy for recognizing seen and unseen compositions. Following the setting in [Purushwalkam *et al.*, 2019], we compute the accuracy in two situations: 1) Seen, testing only on seen compositions; 2) Unseen, testing only on unseen compositions. Based on these, we can compute 3) Harmonic Mean (HM) of the two metrics, which balances the performance between seen and unseen accuracy. Eventually, we compute 4) Area Under the Curve (AUC) to quantify the overall performance of both seen and unseen accuracy at different operating points with respect to the bias. Following [Chao *et al.*, 2016], we utilize a calibration bias to trade

	Val AUC			Test AUC			Test		
Method	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Seen	Unseen	HM
RedWine [Misra *et al.*, 2017]	2.9	7.3	11.8	2.4	5.7	9.3	20.7	17.9	11.6
LE+ [Misra *et al.*, 2017]	3.0	7.6	12.2	2.0	5.6	9.4	15.0	20.1	10.7
AoP [Nagarajan and Grauman, 2018]	2.5	6.2	10.1	1.6	4.7	7.6	14.3	17.4	9.9
TMN [Purushwalkam *et al.*, 2019]	3.5	8.1	12.4	2.9	7.1	11.5	20.2	20.1	13.0
SymNet [Li *et al.*, 2020]	4.3	9.8	14.8	3.0	7.6	12.3	4.4	25.2	16.1
Causal [Atzmon *et al.*, 2020]	1.7	4.0	5.9	1.5	3.4	5.3	17.5	11.8	9.5
CGE [Naeem *et al.*, 2021]	6.8	–	–	5.1	–	–	28.7	25.3	17.2
CompCos [Mancini *et al.*, 2021]	5.9	13.4	19.8	4.5	10.9	16.5	25.3	24.6	16.4
ProtoProp [Ruis *et al.*, 2021]	4.1	9.5	14.4	2.7	7.0	11.3	19.2	20.4	12.6
Co-CGE [Mancini *et al.*, 2022]	–	–	–	6.6	–	–	32.1	28.3	20.0
CLIP [Radford *et al.*, 2021]	13.0	26.2	34.4	11.0	23.0	31.7	30.0	46.0	26.1
CoOp [Zhou *et al.*, 2022]	16.3	31.4	41.5	15.0	28.4	37.2	36.4	49.3	31.7
CSP [Nayak *et al.*, 2023]	21.7	39.1	48.7	19.4	35.3	45.1	47.2	49.6	36.3
HPL (Ours)	**22.7**	**40.0**	**50.1**	**20.2**	**35.8**	**45.9**	**47.5**	**50.6**	**37.3**

Table 2: Comparison with state-of-the-art baselines on *MIT-States* in the closed-world setting. Results are reported in seen/unseen composition recognition accuracy (%). Best and second best results are highlighted in each column.

	Val AUC			Test AUC			Test		
Method	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Seen	Unseen	HM
RedWine [Misra *et al.*, 2017]	30.4	52.2	63.5	27.1	54.6	68.8	57.3	62.3	41.0
LE+ [Misra *et al.*, 2017]	26.4	49.0	66.1	25.7	52.1	67.8	53.0	61.9	40.6
AoP [Nagarajan and Grauman, 2018]	21.5	44.2	61.6	25.9	51.3	67.6	59.8	54.2	40.8
TMN [Purushwalkam *et al.*, 2019]	36.8	57.1	69.2	29.3	55.3	69.8	58.7	60.0	45.0
SymNet [Li *et al.*, 2020]	25.9	–	–	23.4	–	–	49.8	57.4	40.4
Causal [Atzmon *et al.*, 2020]	21.0	43.4	58.3	24.3	47.1	62.0	59.1	51.8	40.5
CGE [Naeem *et al.*, 2021]	38.7	–	–	26.4	–	–	56.8	63.6	41.2
CompCos [Mancini *et al.*, 2021]	38.6	60.1	71.8	28.7	55.9	72.5	59.8	62.5	43.1
ProtoProp [Ruis *et al.*, 2021]	31.6	52.0	65.5	23.2	47.3	63.2	54.1	54.7	38.8
Co-CGE [Mancini *et al.*, 2022]	–	–	–	33.9	–	–	62.3	66.3	48.1
CLIP [Radford *et al.*, 2021]	6.3	22.2	37.5	4.9	14.0	23.2	15.6	49.0	15.5
CoOp [Zhou *et al.*, 2022]	41.1	67.3	76.9	25.7	54.9	72.6	61.8	59.6	39.1
CSP [Nayak *et al.*, 2023]	42.2	69.4	80.7	32.5	62.8	**78.6**	**64.0**	65.8	46.2
HPL (Ours)	**45.4**	**71.8**	**81.9**	**35.0**	**64.2**	78.4	63.0	**68.8**	**48.2**

Table 3: Comparison with state-of-the-art baselines on *UT-Zappos* in the closed-world setting. Results are reported in seen/unseen composition recognition accuracy (%). Best and second best results are highlighted in each column.

off between the prediction scores of seen and unseen pairs. As the calibration bias varies, we can draw a seen-unseen accuracy curve where the AUC metric can be computed. We also follow [Mancini *et al.*, 2021] to conduct evaluations in both closed-world and open-world settings.

Implementation Details. We follow CSP [Nayak *et al.*, 2023] to use the pretrained CLIP [Radford *et al.*, 2021] as our backbone. Specifically, the image encoder and text encoder are directly inherited from the pretrained CLIP ViT-L/14 model's vision transformer (ViT) and language transformer. Our model is implemented with PyTorch [Paszke *et al.*, 2019] and optimized by Adam [Kingma and Ba, 2014] optimizer with the learning rate set to $5e-05$, $5e-04$, $5e-05$ for MIT-State, UT-Zappos, C-GQA, respectively. The weight decay is respectively $1e-05$, $1e-05$, $5e-05$ for the datasets mentioned above. The batch size is set to 128 for all three datasets. All of our experiments were conducted on an NVIDIA RTX A6000 GPU.

4.2 Comparison with State of the Arts

We compare our proposed HPL with the state-of-the-art methods on the three CZSL benchmark datasets. The results are shown in Tabs. 2 to 5. According to the used backbone, we divide the compared methods into traditional ResNet-based methods and VLM-based methods, which are indicated in the tables.

The experiments are conducted in both closed-world and open-world settings. The results on the closed-world setting are shown in Tabs. 2 to 4, respectively for MIT-States, UT-Zappos and C-GQA, from which we can observe that our method consistently outperforms all baselines in terms of both HM and AUC metrics. We reach the highest AUC of 20.2% on MIT-States, 35.0% on UT-Zappos and 7.2% on C-GQA. Besides, we improve the harmonic mean on all the

	Val AUC			Test AUC			Test		
Method	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Seen	Unseen	HM
LE+ [Misra *et al.*, 2017]	3.0	–	–	2.0	–	–	15.0	20.1	10.7
AoP [Nagarajan and Grauman, 2018]	2.5	–	–	1.6	–	–	14.3	17.4	9.9
TMN [Purushwalkam *et al.*, 2019]	3.5	–	–	2.9	–	–	20.2	20.1	13.0
SymNet [Li *et al.*, 2020]	4.3	–	–	3.0	–	–	24.4	25.2	16.1
CGE [Naeem *et al.*, 2021]	**6.8**	–	–	5.1	–	–	28.7	25.3	17.2
CompCos [Mancini *et al.*, 2021]	5.9	–	–	4.5	–	–	25.3	24.6	16.4
Co-CGE [Mancini *et al.*, 2022]	–	–	–	4.1	–	–	**33.3**	14.9	15.5
CLIP [Radford *et al.*, 2021]	0.7	2.0	3.2	1.4	3.5	5.3	7.6	5.0	8.6
CoOp [Zhou *et al.*, 2022]	3.4	7.3	10.9	4.4	9.1	13.1	21.4	25.2	17.2
CSP [Nayak *et al.*, 2023]	5.1	10.3	14.3	6.1	12.3	17.5	28.5	26.9	20.0
HPL (Ours)	5.8	**11.6**	**16.3**	**7.2**	**13.2**	**18.2**	30.8	**28.4**	**22.4**

Table 4: Comparison with state-of-the-art baselines on *C-GQA* in the closed-world setting. Results are reported in seen/unseen composition recognition accuracy (%). Best and second best results are highlighted in each column.

	MIT-States				UT-Zappos				C-GQA			
Method	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
LE+ [Misra *et al.*, 2017]	14.2	2.5	2.7	0.3	60.4	36.5	30.5	16.3	19.2	0.7	1.0	0.08
AoP [Nagarajan and Grauman, 2018]	16.6	5.7	4.7	0.7	50.9	34.2	29.4	13.7	–	–	–	–
TMN [Purushwalkam *et al.*, 2019]	12.6	0.9	1.2	0.1	55.9	18.1	21.7	8.4	–	–	–	–
SymNet [Li *et al.*, 2020]	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CGE [Naeem *et al.*, 2021]	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	**32.7**	1.8	2.9	0.47
CompCos [Mancini *et al.*, 2021]	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	–	–	–	–
Co-CGE [Mancini *et al.*, 2022]	30.3	11.2	10.7	2.3	61.2	45.8	**40.8**	23.3	32.1	3.0	4.8	0.78
CLIP [Radford *et al.*, 2021]	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.27
CoOP [Zhou *et al.*, 2022]	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.70
CSP [Nayak *et al.*, 2023]	46.3	15.7	17.4	5.7	**64.1**	44.1	38.9	22.7	28.7	5.2	6.9	1.20
HPL (Ours)	**46.4**	**18.9**	**19.8**	**6.9**	63.4	**48.1**	40.2	**24.6**	30.1	**5.8**	**7.5**	**1.37**

Table 5: Comparison with state-of-the-art baselines in the open-world setting. Results are reported in seen/unseen composition recognition accuracy (%). Best and second best results are highlighted in each column.

datasets compared with other existing methods. The seen and unseen accuracy on these datasets are also the best. It is worth noting that the seen accuracy of CSP and Co-CGE for UT-Zappos and C-GQA is higher, but their AUC and HM are lower. This is probably because they may have encountered an over-fitting during training, so that their unseen accuracy and harmonic mean are relatively lower during testing. Similarly, on the C-GQA dataset, CGE over-fits in the validation set, so it cannot generalize well to the test set, which shows a lower harmonic mean. Tab. 5 shows the results in the open-world setting, which also demonstrates that our method achieves superior results in this challenging setting. Experimental results on the three challenging datasets demonstrate that our method can effectively improve the performance of the model in CZSL.

4.3 Ablation Study

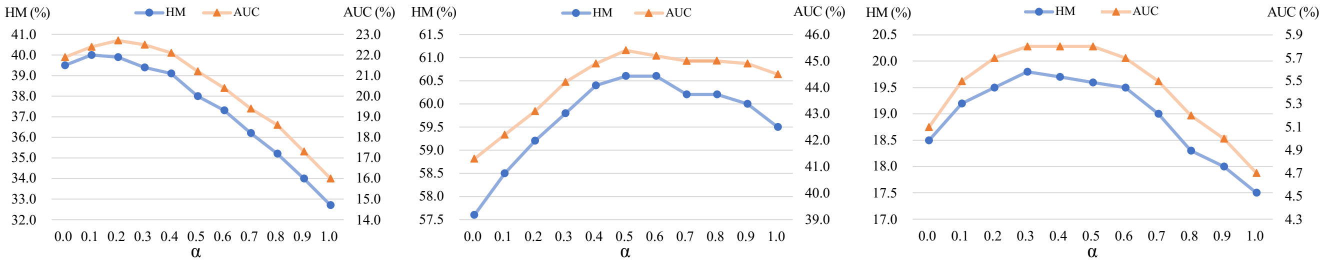
To evaluate the effectiveness of each proposed prompt, we ablate our method on the validation set of the three benchmark datasets. The results are summarized in Tab. 6, Tab. 7 and Fig. 3.

Is Each Prompt Necessary? We test the effects of respectively removing state prompt, object prompt and composition prompt, corresponding to rows 1-3 in Tab. 6. Compared with using full prompt combinations, no matter which one is removed, unfortunately, the accuracy will decrease, which means that all three prompts play positive roles to the compositional recognition ability. We can also observe that, for MIT-States and C-GQA, the composition prompt seems to be more important for recognition, while the combination of state and object prompts is better for UT-Zappos. This is mainly because of the distribution difference between the three datasets: the images of MIT-States and C-GQA present significant visual diversity, while UT-Zappos contains only same-orientated shoe images in white background. The number of state and object in UT-Zappos are much less than that in the other two datasets, and also, with less ambiguity, so that the state and object prompts will have better chance to capture disentangled state and object representations. Broadly speaking, even though either the combination of state prompt with object prompt or the composition prompt can be used to recognize images, one can play a positive role for the recognition performance to each other. The fourth row in Tab. 6 can prove the superiority of combining all three prompts.

How Three Prompts Cooperate? We explore the effect of α in Eq. (8) to see how three prompts can be combined to

	MIT-State				UT-Zappos				C-GQA			
Method	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
w/o *state prompt*	46.9	55.1	39.9	22.3	63.4	70.0	57.8	41.0	27.9	23.6	18.9	5.3
w/o *object prompt*	46.9	54.9	39.4	22.2	65.8	70.4	52.4	39.5	28.6	24.0	19.3	5.5
w/o *comp. prompt*	40.2	48.5	32.7	16.0	66.0	73.9	59.5	44.5	27.3	22.4	17.5	4.7
Full (HPL)	**47.2**	**55.6**	**39.9**	**22.7**	**66.0**	**74.9**	**60.6**	**45.4**	**29.2**	**24.6**	**19.8**	**5.8**

Table 6: Ablation study. Results are reported in seen/unseen composition recognition accuracy (%) in the validation sets of MIT-States, UT-Zappos and C-GQA. Best results are highlight in each column.



(a) MIT-States (b) UT-Zappos (c) C-GQA

Figure 3: The effect of α in Eq. (8). Results are reported in seen/unseen composition recognition accuracy (%) in the validation sets.

Datasets →	MIT-States		UT-Zappos		C-GQA	
Tokens ↓	HM	AUC	HM	AUC	HM	AUC
`[the]` + `[thing]` (ori.)	37.3	20.2	48.2	35.0	22.4	7.2
`[a]` + `[thing]`	37.2	20.0	48.0	34.8	22.0	7.0
`[the]` + `[stuff]`	37.1	20.1	48.3	35.0	22.2	7.1
`[a]` + `[stuff]`	37.2	20.2	48.4	35.3	22.3	7.0
`[?]` + `[thing]`	37.5	20.4	48.1	34.9	22.5	7.4
`[the]` + `[?]`	37.3	20.2	48.0	34.9	22.2	7.1
`[?]` + `[?]`	37.4	20.4	48.5	35.3	22.4	7.4

Table 7: Ablation studies on different tokens in the prompts.

boost the CZSL performance. Concretely, α controls the decision about how we assign prediction strengths for different prompts. Note that we equally treat state and object prompts since they should be firstly combined to produce a compositional prediction. Haven said that, carefully tuning their proportion will arguably lead to better accuracy, but we intend to make it simple by equally treating these two prompts.

We report in Fig. 3 the results by changing $\alpha \in [0, 1]$ with a 0.1 interval. The ends of the lines in each figure are two extreme cases: when $\alpha = 0$, only composition prompt is used in inference; when $\alpha = 1$, we only use the sum of state and object prompt. As can be observed, in MIT-States and C-GQA, the recognition accuracy reaches its peak when α equals 0.2 and 0.3 respectively; in UT-Zappos, the value is 0.5. A possible explanation is that UT-Zappos contains most of the similar appearance combinations, such that the side information provided by the state/object can be of more help to compositional generalization. In contrast, the composition prompt in MIT-State and C-GQA provides more precise decision than the state/object prompts since the images in these two datasets are already diverse. But what can be certained is that either solely using the composition prompt ($\alpha = 0$) or the sum of state and object prompts ($\alpha = 1$) cannot make the best of the final results; by combining all three prompts with a proper α, we can achieve the overall best CZSL performance.

Token-Agnostic Setting. In the whole training stage, we fix partial tokens in different prompts to allow complementary and fine-grained concept learning. To achieve that, these tokens are kept invariant to different state-object compositions, while their embeddings are totally learnable and optimized to be close to the visual features in the state and the object embedding spaces, respectively. In other words, our default token choices, *i.e.*, `the` and `thing`, are only used for initializing the learnable token embeddings.

To verify how different token choices affect our model's performance, Tab. 7 shows the results of ablation studies in the token-agnostic setting, where `the` and `thing` in origin prompts are replaced with `a`, `stuff`, and even random variables (denoted as "`[?]`"). We can see from the results in rows 2–7 that different tokens result in negligible performance discrepancy since they are optimized with the same objective. Overall, the performance of our method is agnostic to the choice of tokens, showing its wide applicability.

5 Conclusion

In this paper, we propose to address Compositional Zero-Shot Learning (CZSL) by learning hierarchical prompts based on large-scale pretrained Vision-Language Models (VLMs). Our motivation comes from the need for a comprehensive modeling for the states, the objects, and the compositions. Specifically, we learn three hierarchical prompts with different fixed word tokens, which can be regarded as three experts with each own expertise. For a given image in inference, we calculate the prediction by combining the strengths of all the three prompts, and demonstrate consistent improvements over using either one (or two) of them. Extensive experiments on three challenging benchmarks also validate its superiority against state-of-the-art competitors.

Acknowledgments

Our work was supported in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, and Grant 62071361, in part by Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, in part by the Fundamental Research Funds for the Central Universities ZDRC2102, and in part by Joint Funds of Ministry of Education of China under Grant 8091B022149. We would like to thank the anonymous reviewers for their valuable comments.

References

[Atzmon *et al.*, 2016] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

[Atzmon *et al.*, 2020] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *Conference on Neural Information Processing Systems*, pages 1462–1473, 2020.

[Bach *et al.*, 2022] Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*, 2022.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*, pages 1877–1901, 2020.

[Chao *et al.*, 2016] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pages 52–68, 2016.

[Dong *et al.*, 2019] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Conference on Neural Information Processing Systems*, 2019.

[Hudson and Manning, 2018] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, pages 1–20, 2018.

[Hudson and Manning, 2019] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[Huynh and Elhamifar, 2020] Dat Huynh and Ehsan Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. In *Conference on Neural Information Processing Systems*, pages 19849–19860, 2020.

[Isola *et al.*, 2015] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1383–1391, 2015.

[Johnson *et al.*, 2017] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

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

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[Li *et al.*, 2021] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. In *AAAI Conference on Artificial Intelligence*, pages 1966–1974, 2021.

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

[Mancini *et al.*, 2021] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5222–5230, 2021.

[Mancini *et al.*, 2022] 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*, pages 1–15, 2022.

[Misra *et al.*, 2017] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017.

[Naeem *et al.*, 2021] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning.

In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021.

[Nagarajan and Grauman, 2018] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *European Conference on Computer Vision*, pages 169–185, 2018.

[Nan *et al.*, 2019] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *AAAI Conference on Artificial Intelligence*, pages 8811–8818, 2019.

[Nayak *et al.*, 2023] Nihal V. Nayak, Peilin Yu, and Stephen Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations*, pages 1–21, 2023.

[Paszke *et al.*, 2019] 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. In *Conference on Neural Information Processing Systems*, pages 8024–8035, 2019.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Meeting of the Association for Computational Linguistics*, pages 2227–2237, 2018.

[Purushwalkam *et al.*, 2019] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *IEEE International Conference on Computer Vision*, pages 3593–3602, 2019.

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

[Radford *et al.*, 2021] 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, 2021.

[Ruis *et al.*, 2021] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. In *Conference on Neural Information Processing Systems*, pages 10641–10653, 2021.

[Saini *et al.*, 2022] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022.

[Sanh *et al.*, 2022] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, pages 1–216, 2022.

[Vu *et al.*, 2022] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *Meeting of the Association for Computational Linguistics*, pages 5039–5059, 2022.

[Wei *et al.*, 2019] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *IEEE International Conference on Computer Vision*, pages 3741–3749, 2019.

[Wei *et al.*, 2020] Kun Wei, Cheng Deng, Xu Yang, et al. Lifelong zero-shot learning. In *International Joint Conference on Artificial Intelligence*, pages 551–557, 2020.

[Xu *et al.*, 2021] Ziwei Xu, Guangzhi Wang, Yongkang Wong, and Mohan S Kankanhalli. Relation-aware compositional zero-shot learning for attribute-object pair recognition. *IEEE Transactions on Multimedia*, 24(8):3652–3664, 2021.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Conference on Neural Information Processing Systems*, 2019.

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

[Yang *et al.*, 2021] Yanhua Yang, Xiaozhe Zhang, Muli Yang, and Cheng Deng. Adaptive bias-aware feature generation for generalized zero-shot learning. *IEEE Transactions on Multimedia*, 25(11):280–290, 2021.

[Yang *et al.*, 2022] Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, pages 1–11, 2022.

[Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[Zhou *et al.*, 2022] 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.