Troika: Multi-Path Cross-Modal Traction for Compositional Zero-Shot Learning

Siteng Huang[1,3]*, Biao Gong[2], Yutong Feng[2], Min Zhang[3], Yiliang Lv[2], Donglin Wang[3]†
[1]Zhejiang University [2]Alibaba Group
[3]Machine Intelligence Lab (MiLAB), AI Division, School of Engineering, Westlake University
{siteng.huang, a.biao.gong}@gmail.com
{zhangmin, wangdonglin}@westlake.edu.cn,, {fengyutong.fyt, yiliang.lyl}@alibaba-inc.com

Abstract

Recent compositional zero-shot learning (CZSL) methods adapt pre-trained vision-language models (VLMs) by constructing trainable prompts only for composed state-object pairs. Relying on learning the joint representation of seen compositions, these methods ignore the explicit modeling of the state and object, thus limiting the exploitation of pre-trained knowledge and generalization to unseen compositions. With a particular focus on the universality of the solution, in this work, we propose a novel paradigm for CZSL models that establishes three identification branches (i.e., Multi-Path) to jointly model the state, object, and composition. The presented ***Troika*** *is an outstanding implementation that aligns the branch-specific prompt representations with decomposed visual features. To calibrate the bias between semantically similar multi-modal representations, we further devise a Cross-Modal Traction module into* ***Troika*** *that shifts the prompt representation towards the current visual content. We conduct extensive experiments on three popular benchmarks, where our method significantly outperforms existing methods in both closed-world and open-world settings. The code will be available at* https://github.com/bighuang624/Troika.

1. Introduction

As for the study of human-like compositional generalization ability, compositional zero-shot learning (CZSL) [20, 26, 32] studies to recognize unseen *compositions* at test time, while states and objects (*i.e.*, *primitives*) are presented in seen *compositions* during training. Rather than learning to associate images with such state-object compositions from scratch [25, 27], recent efforts [23, 29, 37] focus on adapting pre-trained vision-language models (VLMs), *e.g.*, CLIP [33]. Since CZSL datasets provide only compositional labels (*e.g.*, "*red*"+"*wine*") instead of complete sentences (*e.g.*, "*A glass of wine placed on the table*") used in pre-training, to bridge the gap without fine-tuning the entire model, prompts [18, 22] are constructed by appending a simple prefix like "*a photo of*" before the composed state-object labels. Prior methods [23, 29, 37] have replaced the fixed prompt tokens with learnable tokens that are directly optimized during fine-tuning. By designing prompt tuning solutions for compositions, existing methods can efficiently contrive a cross-modal alignment between images and compositional labels, thereby unlocking the potential compositional generalization capability of VLMs.

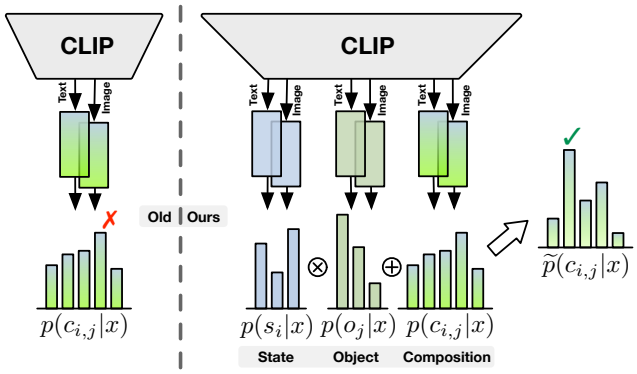

Figure 1. **Graphical comparison of the existing paradigm and the proposed *Multi-Path* paradigm.**

However, due to the lack of independent and explicit *primitive modeling*, these methods have suffered from two challenges: **(1)** the full leveraging of pre-trained knowledge fails, since a large amount of cross-modal information is not tied to the compositions, but related to the single primitive. **(2)** the difficulty of generalizing to unseen compositions is increased, since the model easily over-rely on a limited number of seen compositions. To overcome the issues with a particular focus on the universality of the solution, we propose a novel ***Multi-Path*** paradigm for CZSL with VLMs. As shown in Fig. 1, our paradigm emphasizes the joint modeling of the state, object, and composition without redundant assumptions about the specific implementations. Different from previous methods that depend only on

*Work done during internship at Alibaba Group.
†Corresponding author.

the estimated composition probability, *Multi-Path* paradigm integrates the predictions of all semantic components for the final decision. Following the accessible paradigm, we present an outstanding implementation ***Troika***[1]. On the language side, *Troika* constructs branch-specific prompts that inject learnable priors for describing the context of specific target classes. And on the vision side, while introducing parameter-efficient adaptation, *Troika* decomposes primitive visual features for individual recognition.

Moreover, for calibrating the bias between semantically similar multi-modal representations, we further devise a ***Cross-Modal Traction*** module into *Troika*. The motivation is that compared to diverse visual presentations, learning only a fixed prompt representation is intuitively insufficient to match all corresponding images from different domains (Fig. 2). By selecting and integrating the most semantically relevant visual features, the module pulls the originally static prompt representation towards the visual content. As shown in Tab. 6, while the basic *Troika* has already achieved state-of-the-art (SOTA), the incorporation of the *Cross-Modal Traction* module leads to a significant improvement. Follow-up researches are also free to try other traction ways as the module and *Troika* are decoupled.

Three popular benchmark datasets MIT-States [11], UT-Zappos [38], and C-GQA [27] are used for comparisons. Experiments show that on the closed-world setting, *Troika* exceeds the current state-of-the-art methods by up to **+7.4%** HM and **+5.7%** AUC. And on the more challenging open-world setting, *Troika* still surpasses the best CLIP-based method by up to **+3.8%** HM and **+2.7%** AUC. We also conduct abundant ablations to verify the effectiveness of all component elements of *Troika*. In summary, the main contributions of our work are four-fold:

- We propose a novel *Multi-Path* paradigm for CZSL with VLMs, which explicitly constructs vision-language alignments for the state, object, and composition. The paradigm is flexible enough to derive new approaches.
- Based on the paradigm, we implement a model named *Troika* that effectively aligns the branch-specific prompt representations and decomposed visual features.
- We further design a *Cross-Modal Traction* module for *Troika* that adaptively adjusts the prompt representation depending on the visual content.
- We conduct extensive experiments on three CZSL benchmark datasets to show that *Troika* achieves the SOTA performance on both closed-world and open-world settings.

2. Related Work

Compositional Zero-Shot Learning (CZSL). Aiming to recognize unseen state-object pairs at test time while each

[1]"Troika" is a traditional harness driving combination, using three horses abreast, usually pulling a sleigh.

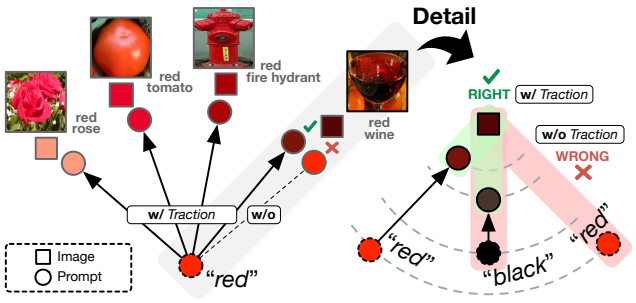

Figure 2. **An example of the *Cross-Modal Traction* module.** The commonly learned prompt of "*red*" may be further away (compared to "*black*") from individual images with the same concept, and the module reduces such mismatches by adaptively pulling the prompt representation towards the current visual content.

semantic primitive exists in training samples, early CZSL efforts can be broadly divided into two lines. One line of works [1, 25–27, 32] learns the combined state-object semantic representation for both seen and unseen compositions with a transformation function, *e.g.*, a multi-layer perceptron (MLP) [26] or a graph convolutional network [27]. Another line of works [14, 19, 20, 24, 41] learns two individual classifiers to identify state and object separately from the image features. While both lines build the connection between visual features and compositional labels from scratch, recent works focus on transferring the encyclopedic knowledge from pre-trained VLMs. CSP [29] first adapts the CLIP model [33] by replacing the classes in textual prompts with trainable state and object tokens. PromptCompVL [37] creates a fully learnable soft prompt including the prefix, state, and object. The latest DFSP [23] proposes a cross-modal decomposed fusion module to learn more expressive image features. In this work, by proposing the *Multi-Path* paradigm for CZSL with VLMs, we highlight the importance of jointly and explicitly modeling the state, object, and composition.

Parameter-Efficient Transfer Learning (PETL). PETL refers to updating only a small number of pre-trained or additional parameters during fine-tuning [4, 6], which reduces the training and storage burdens. As a popular PETL technique, *prompt tuning* [18, 21, 22] optimizes learnable tokens inserted into the input token sequence while freezing the backbone. CLIP-based CZSL methods [23, 29, 37] continue the vein by tuning both the inserted prefix and the primitive vocabulary tokens on downstream semantics. While following the *Multi-Path* paradigm to establish three independent branches, our *Troika* constructs branch-specific prompts with individual prefixes and a shared primitive vocabulary. Fig. 3 illustrates the differences in prompt design between *Troika* and existing methods, and experimental results in Sec. 5.3 show that our design benefits from modeling prior knowledge specified with semantic roles.

In this work, we also attempt to introduce PETL techniques to the vision side, implemented as Adapter [9]. Since

	CLIP [33]				CoOp [43]			
	S	U	HM	AUC	S	U	HM	AUC
w/o MP	15.8	49.1	15.6	5.0	52.1	49.3	34.6	18.8
w/ MP	**24.3**	**49.6**	**21.9**	**8.2**	**62.5**	**58.1**	**41.9**	**28.3**

Table 1. **Improvements to the baselines introduced by the *Multi-Path* paradigm on the UT-Zappos dataset.**

only tuning the text side while ignoring the image encoder is naturally considered insufficient, the studies of adapting both encoders for multi-modal tasks including image recognition [16, 40] and video-text retrieval [10, 13] have recently emerged. On these tasks, completely adapting both encoders has been proved to mutually promote the alignment of vision-language modalities. For the first time, we empirically demonstrate that it is equally valid for CZSL.

3. Rethinking the Paradigm

In this section, we first formalize the CZSL task and the visual feature extraction of CLIP [33] (Sec. 3.1). Then, we introduce how previous works adapt CLIP by adjusting prompts for image-composition alignments (Sec. 3.2). Finally, we present a *Multi-Path* paradigm to guide the construction of VLM-based pipelines (Sec. 3.3).

3.1. Preliminaries

CZSL Task Formulation. Given the state set $\mathcal{S} = \{s_1, s_2, \dots, s_{|\mathcal{S}|}\}$ and object set $\mathcal{O} = \{o_1, o_2, \dots, o_{|\mathcal{O}|}\}$ as the primitive concepts, where $|\cdot|$ denotes the number of elements in the set, the compositional label space $\mathcal{C}$ is defined as their Cartesian product, *i.e.*, $\mathcal{C} = \mathcal{S} \times \mathcal{O}$. And the set of the seen and unseen compositions, denoted as $\mathcal{C}^{se}$ and $\mathcal{C}^{us}$, are the two disjoint subsets of $\mathcal{C}$, *i.e.*, $\mathcal{C}^{se} \cap \mathcal{C}^{us} = \emptyset$. To learn a model that assigns compositional labels from the target set $\mathcal{C}^{tgt}$ to the input images, a training set $\mathcal{T} = \{(x_i, c_i) | x \in \mathcal{X}, c \in \mathcal{C}^{se}\}$ is provided, where $\mathcal{X}$ denotes the image space. In the closed-world setting, the target set is defined as $\mathcal{C}^{tgt} = \mathcal{C}^{se} \cup \mathcal{C}^{us}$, where only the known composition space is considered. And in the open-world setting, the target set is all possible permutations of the state-object compositions, *i.e.*, $\mathcal{C}^{tgt} = \mathcal{C}$.

Visual Feature Extraction. Given the input image $x \in \mathbb{R}^{H \times W \times C}$, the image encoder E_v, implemented with ViT [5], first splits it into $N^p = HW/P^2$ non-overlapping patches, where (P, P) is the resolution of each patch. The patches are projected to form a sequence of patch tokens together with a pre-trained `[CLS]` token, where the pre-trained position embeddings are also added to preserve positional information. Then, the encoder E_v updates the token sequence $\mathbf{X} \in \mathbb{R}^{(N^p+1) \times d_v^{in}}$ with self-attention-based blocks, where d_v^{in} is the dimension of each visual token. Finally, a single linear layer g^{proj} with parameters $\mathbf{W}_g \in \mathbb{R}^{d_v^{in} \times d}$ projects the output `[CLS]` token, where d is the dimension of the cross-modal latent space. And the projected token $\mathbf{x}^{\text{CLS}} \in \mathbb{R}^d$ serves as the image representation.

Method	Prefix	Vocabulary	Prompt
CLIP	*a photo of*	*red tomato*	$\mathbf{P}_{\text{red,tomato}}$
CoOp	$\mathbf{p}_1, \dots, \mathbf{p}_m$	*red tomato*	$\mathbf{P}_{\text{red,tomato}}$
CSP	*a photo of*	$\mathbf{v}^s_{\text{red}}\,\mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}_{\text{red,tomato}}$
DFSP	$\mathbf{p}_1, \dots, \mathbf{p}_m$	$\mathbf{v}^s_{\text{red}}\,\mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}_{\text{red,tomato}}$
Troika (Ours)	$\mathbf{p}^s_1, \dots, \mathbf{p}^s_m$	$\mathbf{v}^s_{\text{red}}$	$\mathbf{P}^s_{\text{red}}$
	$\mathbf{p}^o_1, \dots, \mathbf{p}^o_m$	$\mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}^o_{\text{tomato}}$
	$\mathbf{p}^c_1, \dots, \mathbf{p}^c_m$	$\mathbf{v}^s_{\text{red}}\,\mathbf{v}^s_{\text{tomato}}$	$\mathbf{P}^c_{\text{red,tomato}}$

Figure 3. **Graphical comparison of prompts constructed by prior methods and *Troika*.** Red tokens are trainable.

3.2. A Revisit of Existing Framework

Taking a closer look into existing CLIP-based works [23, 29, 37], we found that all of them build a single cross-modal alignment for inference, which yields the recognition probability $p(c_{i,j}|x)$ given the input image x and the candidate pair $c_{i,j} = \langle s_i, o_j \rangle$. Since the frozen CLIP backbone has provided a well-established vision-language alignment, an essential step for these methods is to construct appropriate prompts for compositional labels. Commonly, initializing with the pre-trained embeddings from CLIP, a new primitive *vocabulary* $\mathbf{V} = [\mathbf{V}^{\mathcal{S}}, \mathbf{V}^{\mathcal{O}}] \in \mathbb{R}^{(|\mathcal{S}|+|\mathcal{O}|) \times d_t^{in}}$ is first built for all states and objects, where d_t^{in} is the dimension of each vocabulary token. Then, a natural language *prefix* such as "*a photo of*" is transformed into tokens with the pre-trained embeddings. Different from the prompt format adopted in the inference of CLIP, the CZSL methods [29] append the prefix tokens to the state-object composition instead of the class placeholder, acquiring the prompt $\mathbf{P}_{i,j} = [\mathbf{p}_1, \dots, \mathbf{p}_m, \mathbf{v}^s_i, \mathbf{v}^o_j]$ for the pair $c_{i,j}$, where $\{\mathbf{p}_1, \dots, \mathbf{p}_m\} \in \mathbb{R}^{m \times d_t^{in}}$ are the prefix tokens, and $\mathbf{v}^s_i, \mathbf{v}^o_j$ are the vocabulary tokens of s_i and o_j. By feeding $\mathbf{P}_{i,j}$ into the text encoder E_t, the prompt representation $\mathbf{t}_{i,j} \in \mathbb{R}^d$ is acquired to compute $p(c_{i,j}|x)$ by cosine similarity with image representation $\mathbf{x}^{\text{CLS}}$. While earlier works [29, 37] simply turn the primitive vocabulary or prefix tokens from fixed to trainable, DFSP [23] further decomposes the prompt representations into state and object ones to provide more supervision for training. However, the inference still relies on a single path for estimating the composition probability.

3.3. Generalizing in Multi-Path Paradigm

As already discussed in the introduction, existing methods still suffer from limitations in knowledge transfer and generalization. We believe that the issue stems from the applied single-path paradigm, and thus present a novel *Multi-Path* paradigm. An intuitive comparison of the existing and proposed paradigms is illustrated in Fig. 1. Crucially, the *Multi-Path* paradigm requires a recognition branch for each of the three semantic components, *i.e.*, state, object, and composition. These branches are essentially cross-modal alignments that independently unearth specific knowledge from large-scale vision-language pre-training.

Specifically, during training, three branches can collec-

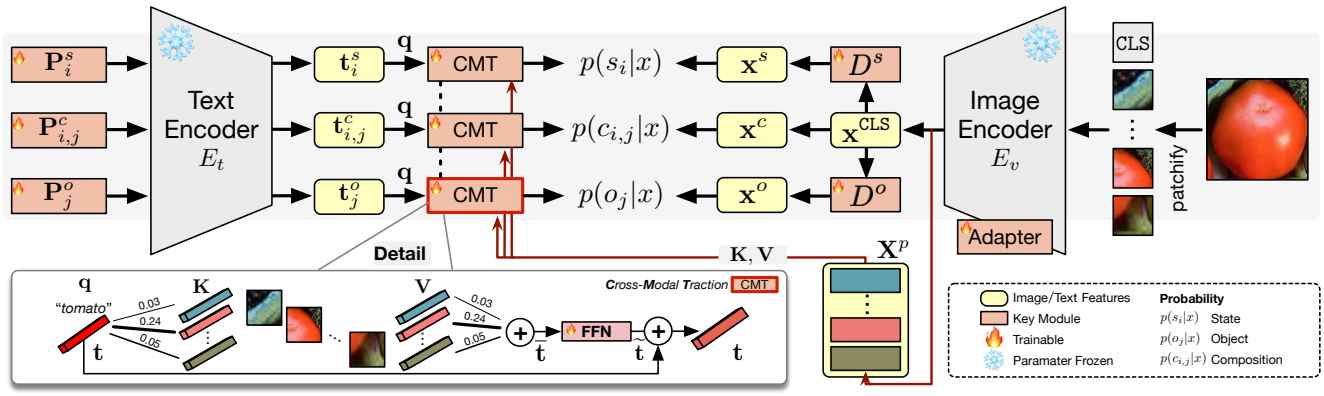

Figure 4. **Overview of the proposed *Troika*.**

tively optimize the parameters in a multi-task learning [42] manner. And for inference, the prediction results of state and object can be incorporated to assist the composition branch. Formally, the integrated composition probability $\widetilde{p}(c_{i,j}|x)$ is defined as

$$\widetilde{p}(c_{i,j}|x) = p(c_{i,j}|x) + p(s_i|x) \cdot p(o_j|x), \tag{1}$$

where the joint distribution of composition probabilities, when attribute and object predictions are considered independent of each other, is treated as a bias correction for the direct prediction of compositions. And the most likely composition can be predicted as

$$\hat{c} = \underset{c_{i,j} \in \mathcal{C}^{tgt}}{\arg\max} \left(\widetilde{p}(c_{i,j}|x) \right), \tag{2}$$

thereby mitigating the excessive bias towards seen compositions and promoting a more robust recognition system. Without further implementation constraints, the flexibility of the *Multi-Path* paradigm allows for the free derivation of new methods based on powerful VLMs. To demonstrate the usability and effectiveness of the proposed paradigm, we follow it to rapidly extend the two popular baseline models, CLIP [33] and CoOp [43], and illustrate the significant improvements in Tab. 1.

4. Troika: An Efficient Implementation

To adhere to the paradigm for more efficient solutions, the multi-modal feature extraction across branches must be carefully designed. In this section, we first detail how the proposed *Troika* develops a range of instantiations. Then, we introduce the plug-and-play *Cross-Modal Traction* module for *Troika*. An overview of *Troika* is illustrated in Fig. 4.

4.1. Instantiations

Learning Prompt Representations. Since the composition prompt can only elicit information related to seen compositions from VLMs, *Troika* respectively conducts prompts for the state, object, and composition branch. Compared to the DFSP [23] approach that decomposes the text features into state and object ones, individual prompts for different semantic components activate the backbone from input, thus maximizing the exploitation of pre-trained knowledge. Moreover, as the semantic roles of the target classes on each branch are different, it is natural to introduce different priors through special contexts. Therefore, while maintaining the same primitive vocabulary as a cue of semantic compositionality, we employ an independent prompt prefix for each branch of *Troika*. For each state-object pair $c_{i,j} = \langle s_i, o_j \rangle$, the state prompt $\mathbf{P}_i^s$, object prompt $\mathbf{P}_j^o$, and composition prompt $\mathbf{P}_{i,j}^c$ can be constructed as

$$\mathbf{P}_i^s = [\mathbf{p}_1^s, \ldots, \mathbf{p}_m^s, \mathbf{v}_i^s], \tag{3}$$

$$\mathbf{P}_j^o = [\mathbf{p}_1^o, \ldots, \mathbf{p}_m^o, \mathbf{v}_j^o], \tag{4}$$

$$\mathbf{P}_{i,j}^c = [\mathbf{p}_1^c, \ldots, \mathbf{p}_m^c, \mathbf{v}_i^s, \mathbf{v}_j^o], \tag{5}$$

where $\{\mathbf{p}_1^s, \ldots, \mathbf{p}_m^s\}$, $\{\mathbf{p}_1^o, \ldots, \mathbf{p}_m^o\}$, and $\{\mathbf{p}_1^c, \ldots, \mathbf{p}_m^c\}$ are the learnable state prefix, object prefix, and composition prefix, respectively. These fully trainable prompts are then fed into the text encoder E_t to obtain the prompt representation for each branch, formulated as

$$\mathbf{t}_i^s = E_t(\mathbf{P}_i^s), \quad \mathbf{t}_j^o = E_t(\mathbf{P}_j^o), \quad \mathbf{t}_{i,j}^c = E_t(\mathbf{P}_{i,j}^c). \tag{6}$$

Learning Visual Representations. While existing methods [23, 29, 37] directly apply the frozen image encoder, we first trial several easy-to-implement and effective PETL techniques. Based on the experimental results, we finally introduce Adapter [9] to adapt the image encoder without updating its original parameters. These small neural modules (adapters) can also be freely replaced by more complicated techniques to pursue further enhancements. Then, to establish cross-modal alignment on each branch, specific visual features for the composition, state, and object should be extracted. Still treating the image representation $\mathbf{x}^{\text{CLS}}$ as the composition visual representation $\mathbf{x}^c$, we introduce the state disentangler D^s and object disentangler D^o to decompose the state and object visual features $\mathbf{x}^s$ and $\mathbf{x}^o$ as

$$\mathbf{x}^s = D^s(\mathbf{x}^{\text{CLS}}), \quad \mathbf{x}^o = D^o(\mathbf{x}^{\text{CLS}}), \quad \mathbf{x}^c = \mathbf{x}^{\text{CLS}}, \tag{7}$$

where D^s and D^o are implemented with two individual MLPs. This simple structure provides the necessary non-linear mapping to resolve primitive-specific features from entangled global features.

Training. Given the prompt representations and visual features specified with different branches, the probability of assigning labels of the state s_i, object o_j, and composition $c_{i,j}$ to the image can be computed separately as

$$p(s_i|x) = \frac{\exp(\mathbf{x}^s \cdot \mathbf{t}_i^s/\tau)}{\sum_{k=1}^{|\mathcal{S}|} \exp(\mathbf{x}^s \cdot \mathbf{t}_k^s/\tau)}, \tag{8}$$

$$p(o_j|x) = \frac{\exp(\mathbf{x}^o \cdot \mathbf{t}_j^o/\tau)}{\sum_{k=1}^{|\mathcal{O}|} \exp(\mathbf{x}^o \cdot \mathbf{t}_k^o/\tau)}, \tag{9}$$

$$p(c_{i,j}|x) = \frac{\exp(\mathbf{x}^c \cdot \mathbf{t}_{i,j}^c/\tau)}{\sum_{k=1}^{|\mathcal{C}^{tgt}|} \exp(\mathbf{x}^c \cdot \mathbf{t}_k^c/\tau)}, \tag{10}$$

where $\tau \in \mathbb{R}$ is the pre-trained temperature parameter from CLIP. In each branch, the cross-entropy loss encourages the model to explicitly recognize the corresponding semantic role, described as

$$\mathcal{L}^s = -\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log p(s|x), \tag{11}$$

$$\mathcal{L}^o = -\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log p(o|x), \tag{12}$$

$$\mathcal{L}^c = -\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \log p(c|x). \tag{13}$$

Therefore, the overall training loss $\mathcal{L}_{all}$ is defined as

$$\mathcal{L}_{all} = \alpha^s \mathcal{L}^s + \alpha^o \mathcal{L}^o + \alpha^c \mathcal{L}^c, \tag{14}$$

where $\alpha^s, \alpha^o, \alpha^c \in \mathbb{R}$ are weighting coefficients to balance the influences of different losses. Note that we have omitted the weight decay here for simplicity. *Troika* adheres to the unified inference of the *Multi-Path* paradigm (*i.e.*, Eqs. (1) and (2)) without making any additional modifications.

4.2. Cross-Modal Traction

Although inheriting the rich cross-modal understanding of VLMs, discrepancies may still exist between semantically similar vision-language representations. Given the same semantic concept, the static and monotonous prompt representation naturally fails to be commonly optimal for all input images that come from a plentiful distribution. This issue becomes more serious in the additional state and object branches, as the visual content of the same primitive changes considerably when paired with different primitives. Therefore, we further develop a *Cross-Modal Traction* module for *Troika*. The module adaptively shifts the prompt representation to accommodate the content diversity and diminish the cross-modal discrepancies. In this process, relevant patch features serve as the guidance to avoid noise from semantic-agnostic sub-regions interfering with the traction.

Specifically, the *Cross-Modal Traction* module is composed of a stack of N blocks, and in each block, we first consider a scaled dot product attention mechanism [35] with the prompt representation attending to all patch tokens. Given the input prompt representation $\mathbf{t}$ that comes from an arbitrary branch, we first acquire the patch tokens $\mathbf{X}^p \in \mathbb{R}^{N^p \times d}$ after projecting them with the linear layer g^{proj}. Then, the query, key and value can be derived as

$$\mathbf{q} = \mathbf{t}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{X}^p\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}^p\mathbf{W}_V, \tag{15}$$

where $\mathbf{W}_q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d^{attn}}$ are the parameter matrices, and d^{attn} is the dimension of the single-head attention. The dot product attention gives relevance weights from $\mathbf{t}$ to each patch token, which are used to aggregate the value-projected patch tokens as

$$\bar{\mathbf{t}} = \text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{K}^\top}{\sqrt{d^{attn}}}\right)\mathbf{V}. \tag{16}$$

In practice, a multi-head design with $h = d/d^{attn}$ parallel attention heads is naturally introduced to diversify representation subspaces. After the attention layer, a feed-forward network FFN, implemented as a MLP, is introduced as

$$\widetilde{\mathbf{t}} = \text{FFN}(\bar{\mathbf{t}}) = \sigma(\bar{\mathbf{t}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{17}$$

where $\mathbf{W}_1, \mathbf{W}_2$ are parameter matrices, $\mathbf{b}_1, \mathbf{b}_2$ are bias terms, and $\sigma(\cdot)$ is a nonlinear activation function. Note that for both the attention and the feed-forward network, we omit the residual connections around them for simplicity. Then, we can update the prompt representation as

$$\mathbf{t} \leftarrow \mathbf{t} + \boldsymbol{\lambda} \cdot \widetilde{\mathbf{t}}, \tag{18}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^d$ is a trainable parameter vector controlling the strength of the cross-modal traction in each dimension. $\widetilde{\mathbf{t}}$ can be viewed as a traction that pulls $\mathbf{t}$ towards the visual content. And the whole module can be seamlessly inserted into each branch before calculating the cross-modal matching probability. In practice, all three branches share the same module to reduce the parameter overhead.

Dataset	$\mathcal{S}$	$\mathcal{O}$	Training $\mathcal{C}^{se}$	Training $\mathcal{X}$	Validation $\mathcal{C}^{se}$	Validation $\mathcal{C}^{us}$	Validation $\mathcal{X}$	Test $\mathcal{C}^{se}$	Test $\mathcal{C}^{us}$	Test $\mathcal{X}$
MIT-States [11]	115	245	1262	30k	300	300	10k	400	400	13k
UT-Zappos [38]	16	12	83	23k	15	15	3k	18	18	3k
C-GQA [27]	413	674	5592	27k	1252	1040	7k	888	923	5k

Table 2. **Statistics of three datasets in our experiments.** The number of elements in each set is reported.

5. Experiments

5.1. Experimental Setup

Datasets. We experiment with three real-world CZSL benchmarks: MIT-States [11], UT-Zappos [38], and C-GQA [27]. We follow the split suggested by previous works [29, 32], and summarize detailed statistics in Tab. 2.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
Closed-world Results												
CLIP [33]	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
CoOp [43]	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
CSP [29]	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
PromptCompVL [37]	48.5	47.2	35.3	18.3	64.4	64.0	46.1	32.2	-	-	-	-
DFSP(i2t) [23]	47.4	52.4	37.2	20.7	64.2	66.4	45.1	32.1	35.6	29.3	24.3	8.7
DFSP(BiF) [23]	47.1	52.8	37.7	20.8	63.3	69.2	47.1	33.5	36.5	32.0	26.2	9.9
DFSP(t2i) [23]	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
***Troika* (Ours)**	**49.0**$\pm$0.4	**53.0**$\pm$0.2	**39.3**$\pm$0.2	**22.1**$\pm$0.1	**66.8**$\pm$1.1	**73.8**$\pm$0.6	**54.6**$\pm$0.5	**41.7**$\pm$0.7	**41.0**$\pm$0.2	**35.7**$\pm$0.3	**29.4**$\pm$0.2	**12.4**$\pm$0.1
Open-world Results												
CLIP [33]	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.27
CoOp [43]	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.70
CSP [29]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
PromptCompVL [37]	48.5	16.0	17.7	6.1	64.6	44.0	37.1	21.6	-	-	-	-
DFSP(i2t) [23]	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4	35.6	6.5	9.0	1.95
DFSP(BiF) [23]	47.1	18.1	19.2	6.7	63.5	57.2	42.7	27.6	36.4	7.6	10.6	2.39
DFSP(t2i) [23]	47.5	18.5	19.3	6.8	**66.8**	60.0	44.0	30.3	38.3	7.2	10.4	2.40
***Troika* (Ours)**	**48.8**$\pm$0.4	**18.7**$\pm$0.1	**20.1**$\pm$0.1	**7.2**$\pm$0.1	66.4$\pm$1.0	**61.2**$\pm$1.0	**47.8**$\pm$1.3	**33.0**$\pm$1.0	**40.8**$\pm$0.2	**7.9**$\pm$0.2	**10.9**$\pm$0.3	**2.70**$\pm$0.1

Table 3. **Main results on three benchmarks.** All methods use a CLIP ViT-L/14 backbone. For our *Troika*, we report the average performance on 5 random seeds with standard error.

Metrics. We follow the evaluation protocol of previous works [27, 32, 41], where a calibration bias trades off between the prediction scores of seen and unseen pairs at test time. While varying the candidate bias from $-\infty$ to $+\infty$, a curve can be drawn with the accuracy of seen and unseen pairs. To quantify the overall performance on both seen and unseen pairs, we compute the area under the curve (**AUC**) and find the point with the best harmonic mean (**HM**) between the seen and unseen accuracy. We also report the best seen accuracy (**S**) by adjusting the bias to $-\infty$, and the best unseen accuracy (**U**) by adjusting the bias to $+\infty$.

Implementation Details. We implement *Troika* with a pre-trained CLIP ViT-L/14 model in PyTorch [30]. The model is trained and evaluated on an NVIDIA A100 GPU. For the open-world evaluation, we follow the post-training calibration method [29] to filter out infeasible compositions. More details can be found in the supplementary material.

5.2. Main Results

For a fair comparison, we primarily compare with CLIP-based methods using the same CLIP ViT-L/14 backbone. In particular, pre-trained CLIP [33], CoOp [43], CSP [29], PromptCompVL [37], and all versions of DFSP [23] are considered. For comparisons with more baselines involved, please refer to the supplementary material.

In Tab. 3, we report both closed-world and open-world results. On the **closed-world** setting, our *Troika* exceeds the previous SOTA methods on MIT-States, UT-Zappos, and C-GQA. Specifically, relative to existing methods, *Troika* improves the HM by +2.0%, +7.4%, +2.3%, and the AUC by +1.5%, +5.7%, +1.9% respectively on three datasets. And *Troika* also achieves the best seen and unseen accuracies on these datasets. On the **open-world** setting, our *Troika* also achieves the SOTA results on the three datasets in terms of almost all metrics. The only exception is that the best seen accuracy of *Troika* is 0.4% lower than DFSP(t2i) [23]. However, *Troika* outperforms the existing methods by +0.8%, +3.8%, +0.5% in terms of the HM, and by +0.4%, +2.7%, +0.3% in terms of the AUC on the three datasets, indicating that our *Troika* achieves a more consistent and comprehensive performance.

Branch			UT-Zappos				C-GQA			
c	s	o	S	U	HM	AUC	S	U	HM	AUC
Training + Inference										
✓	✓	✓	66.8	73.8	**54.6**	**41.7**	**41.0**	**35.7**	**29.4**	**12.4**
✓			66.8	**74.5**	49.9	37.7	39.6	34.3	28.9	11.6
	✓	✓	67.9	69.4	47.0	35.7	35.8	20.2	19.1	5.9
✓	✓		63.4	73.9	51.1	38.6	40.5	34.4	28.8	11.8
✓		✓	67.2	73.1	49.7	37.7	39.5	33.8	28.8	11.5
Inference										
✓	✓	✓	66.8	73.8	**54.6**	**41.7**	**41.0**	**35.7**	**29.4**	**12.4**
✓			**68.2**	72.1	53.0	41.0	39.6	34.0	28.3	11.5
	✓	✓	66.4	68.8	46.5	34.5	36.9	20.7	19.8	6.3
✓	✓		66.3	70.0	53.7	39.9	40.7	33.7	28.8	11.7
✓		✓	**68.2**	72.9	52.0	40.4	40.0	34.4	28.4	11.7

Table 4. **Ablation on the *Multi-Path* paradigm.** The best results are obtained by keeping all three branches in both the training and inference phases.

5.3. Ablation Study

To empirically show the effectiveness of our framework design, we conduct extensive experiments and report the closed-world results for the ablation study.

Ablation on Multi-Path Paradigm. In Tab. 4, we remove one or more specific branches at a time to prove that all branches in the *Multi-Path* paradigm contribute. Specifically, two scenarios are considered: (1) **Training + Inference**, which refers to simultaneously eliminating the effects of the corresponding branches in both training and inference phases, *i.e.*, removing the corresponding losses from Eq. (14) and the corresponding probabilities from Eq. (1).

Prefix	Vocab.	UT-Zappos S	U	HM	AUC	C-GQA S	U	HM	AUC
$c\|s\|o$	cso	**66.8**	**73.8**	**54.6**	**41.7**	**41.0**	**35.7**	**29.4**	**12.4**
cso	cso	**66.8**	72.8	54.0	41.1	39.6	32.9	28.7	11.3
$c\|so$	cso	66.2	72.9	54.5	41.4	39.8	33.7	28.9	11.6
$c\|s\|o$	$c\|s\|o$	67.2	72.5	52.9	40.8	39.9	33.0	28.9	11.5

Table 5. **Ablation on individual prefixes and shared vocabulary.** Branches separated by "|" do not share the corresponding prompt parameters.

Troika	UT-Zappos S	U	HM	AUC	C-GQA S	U	HM	AUC
w/ CMT	**66.8**	**73.8**	**54.6**	**41.7**	**41.0**	**35.7**	**29.4**	**12.4**
w/o CMT	64.4	70.7	51.9	37.8	38.5	33.2	27.9	11.0

Table 6. **Ablation on the *Cross-Modal Traction* module.**

(2) **Inference**, which refers to eliminating the effects of the corresponding branches only during inference, leaving the training loss unchanged. Several observations in the table are worth highlighting: (1) In both scenarios, removing the composition branch results in the greatest drop in performance, illustrating the importance of the branch for learning compositionality. (2) Generally, removing the branches only during inference achieves a better result than removing them in both phases, indicating that all loss items have a positive impact. (3) Keeping all branches in both scenarios leads to the best HM and AUC. Note that some cases on UT-Zappos achieve a higher best seen or unseen accuracy, which only means that they might be better in unrealistic extremes. And their worse HM and AUC suggest that removing branches would introduce instability.

Ablation on Prefix and Vocabulary Design. In Tab. 5, based on the current design of *Troika* (the top row), we first allow all three branches to share the prefix parameters (the second row), as well as allowing only the state and object branches to share the prefix parameters (the third row). We can observe that separating the prefixes of the composition branch and the primitive branches leads to higher HM and AUC. Moreover, maintaining an individual prefix for each branch achieves the best results for all metrics. The results show that it is necessary to inject branch-specific prior knowledge into the prefix parameters. We also attempt to build an individual primitive vocabulary for each branch (the bottom row), which results in a significant drop in performance. We attribute this to a disruption of the semantic dependency modeling. As a conclusion, jointly optimizing the shared vocabulary parameters from multiple paths contributes to the compositional learning.

Ablation on Cross-Modal Traction Module. In Tab. 6, we validate the effectiveness of our *Cross-Modal Traction* module by removing it from *Troika*. We can observe that equipping *Troika* with the *Cross-Modal Traction* module boosts the HM by 2.1% and the AUC by 2.65% in average. This illustrates that by effectively calibrating cross-modal deviations, the adaptive traction improves the accuracy.

To qualitatively evaluate whether the *Cross-Modal Traction* module indeed exploits the semantically similar patch features, we also visualize the attention weights of several test samples from MIT-States in Fig. 5. We can observe that the patches that are closer to the label semantics receive more attention, which also means that they contribute more to the cross-modal traction.

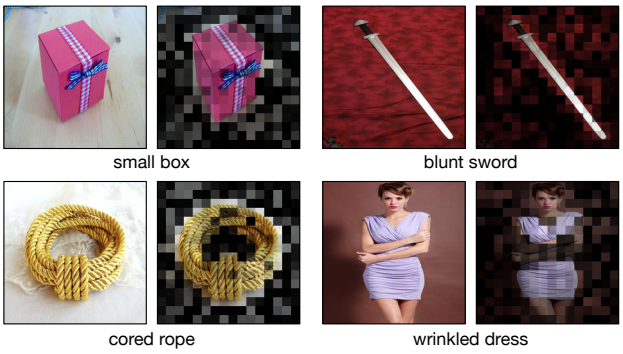

Figure 5. **Visualization analysis of the *Cross-Modal Traction* module.** We show the original image and the visualization result in pairs. The brighter the patch, the greater its role in the traction.

Ablation on Visual Tuning Strategy. In Tab. 7, we compare the following popular tuning strategies for the pre-trained image encoder: (1) **None**: freeze the encoder without updating its parameters. (2) **Full**: fully update all parameters of the encoder. (3) **Bias** [2, 39]: fine-tune only the bias terms. (4) **Proj** [12]: fine-tune only the last linear projection layer g^{proj}. (5) **Partial** [12]: fine-tune only the last block of the Transformer inside the encoder. (6) **Prompt** [12]: fine-tune only the trainable prompt tokens inserted into the token sequence $\mathbf{X}$. (7) **Adapter** [3, 9]: fine-tune only the adapter modules inserted into the Transformer inside the encoder, which is currently applied by *Troika*.

We here highlight some observations from the table: (1) Without adopting any tuning strategies for the image encoder, our approach still outperforms existing CLIP-based methods on most datasets, demonstrating the effectiveness of our proposed innovations including the multi-patch paradigm and *Cross-Modal Traction* mechanism. (2) Although fully fine-tuning the image encoder achieves the best results on C-GQA, it hurts the performance on MIT-States and UT-Zappos to even underperform freezing the encoder. Since C-GQA has much more training classes than the other two datasets, the observation suggests that fully fine-tuning the large pre-trained model easily overfits the training data, which results in poor generalization. We note that this conclusion is consistent with existing studies [29, 34]. (3) All parameter-efficient tuning strategies, including those tune part of the original parameters (Bias, Proj, Partial) and tune additional parameters (Prompt, Adapter), significantly boost the performance compared to freezing the image encoder. (4) Our applied Adapter achieves the most best results while its performance remains in the top two, indicating the superiority of our method design.

Visual Tuning	MIT-States				UT-Zappos				C-GQA			
	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
None	48.3	50.8	37.5	20.6	62.7	70.7	50.3	36.2	34.8	29.5	24.2	8.5
Full	41.7	36.3	28.7	12.2	48.9	57.4	34.4	19.1	**44.5**	**36.5**	**31.8**	**14.1**
Bias [2]	48.6	52.4	38.8	21.7	**66.8**	70.4	51.1	38.1	37.4	32.9	27.0	10.3
Proj [12]	47.9	51.6	38.4	20.9	63.9	71.4	52.3	38.9	35.5	29.1	24.5	8.7
Partial [12]	**49.9**	51.3	38.0	21.4	65.1	70.8	53.9	39.3	38.4	33.3	28.1	11.1
Prompt [12]	48.9	51.3	38.1	21.3	65.0	71.2	51.1	38.0	36.7	30.6	26.1	9.6
Adapter [9]	49.0	**53.0**	**39.3**	**22.1**	**66.8**	**73.8**	**54.6**	**41.7**	41.0	35.7	29.4	12.4

Table 7. **Ablation on visual tuning strategy.** Best results are displayed in **boldface**, and second best results are underlined.

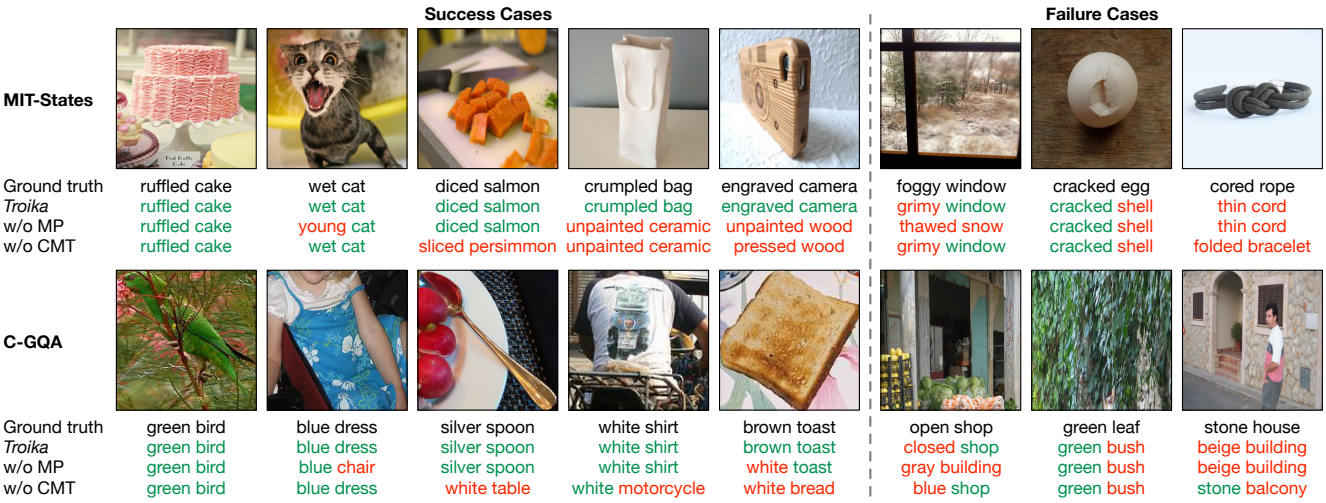

Figure 6. **Qualitative results.** We show top-1 predictions for randomly selected cases from MIT-States (the top row) and C-GQA (the bottom row). The complete *Troika* correctly predicts the examples of five cols on the left, and fails on the examples of three cols on the right. Predictions when removing the *Multi-Path* paradigm (*i.e.*, w/o MP) or the *Cross-Modal Traction* module (*i.e.*, w/o CMT) are also reported. Green denotes the correct prediction and red denotes the wrong prediction.

5.4. Qualitative Results

In Fig. 6, we visualize some qualitative results for both seen and unseen compositions, where the showed cases are randomly sampled from the test set of MIT-States and C-GQA datasets. We report the predictions of the complete *Troika* and the models that remove the *Multi-Path* paradigm or the *Cross-Modal Traction* module. It can be observed that benefiting from both two innovations, *Troika* recognizes the compositions with higher accuracy. Taking the 5th case in the top row as an example, while the incomplete methods may be confused by the color and material presented by the object, the complete *Troika* can focus on details such as shape, surface texture, and even local regions containing lens for comprehensive reasoning. We also show some failure cases, where the entanglement of visual features places extreme demands on combinatorial understanding. However, the proposed *Multi-Path* paradigm enables the complete *Troika* to correctly identify part of the contained primitives. For the cases of complete prediction errors, although different from the provided labels, we find that the predictions can also interpret the content of these images. This indicates the effectiveness of our *Troika* from another perspective beyond the metrics.

6. Conclusion

In this paper, we explore the universal solution of adapting pre-trained VLMs for the downstream CZSL task. We first propose a novel and flexible *Multi-Path* paradigm that requires the simultaneous and explicit modeling of the state, object, and composition. On top of that, follow-up researches can easily generate various new approaches with different multi-modal features. And we develop a method named *Troika*, which implements the paradigm with sophisticated designs on both the language and vision sides. We also present a *Cross-Modal Traction* module to improve *Troika* by calibrating the bias between the static prompt representation and diverse visual content. Both closed-world and open-world results on three benchmarks illustrate the superiority of *Troika*, and extensive ablations also demonstrate the effectiveness of each component. We hope that our work can inspire future research on exploiting foundational VLMs for compositional learning.

Acknowledgement This work was supported by STI 2030—Major Projects (2022ZD0208800), NSFC General Program (Grant No. 62176215). This work was supported by Alibaba Group through Alibaba Research Intern Program.

References

[1] Muhammad Umer Anwaar, Zhihui Pan, and Martin Kleinsteuber. On leveraging variational graph embeddings for open world compositional zero-shot learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 4645–4654, 2022. 2, 3, 4

[2] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. TinyTL: reduce memory, not parameters for efficient on-device learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 11285–11297, 2020. 7, 8

[3] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: adapting vision transformers for scalable visual recognition. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 16664–16678, 2022. 7

[4] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Delta Tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022. 2

[5] 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 *Proceedings of the International Conference on Learning Representations*, 2021. 3

[6] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *Proceedings of the International Conference on Learning Representations*, 2022. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[8] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Unit (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 1

[9] 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 *Proceedings of the International Conference on Machine Learning*, pages 2790–2799, 2019. 2, 4, 7, 8, 1

[10] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. VoP: text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6565–6574, 2023. 3

[11] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1383–1391, 2015. 2, 5

[12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision*, pages 709–727, 2022. 7, 8

[13] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022. 3

[14] 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. 2, 3, 4

[15] 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. 3, 4

[16] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 3

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015. 1

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1, 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. 2, 3, 4

[20] Yonglu 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 11313–11322, 2020. 1, 2, 3, 4

[21] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 2

[22] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 1, 2

[23] Xiaocheng Lu, Ziming Liu, Song Guo, 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. 1, 2, 3, 4, 6

[24] 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. 2, 3, 4

[25] 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, 3, 4

[26] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1160–1169, 2017. 1, 2, 4

[27] 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, 5, 6, 4

[28] Tushar Nagarajan and Kristen Grauman. Attributes as operators: Factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision*, pages 172–190, 2018. 3, 4

[29] Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *Proceedings of the International Conference on Learning Representations*, 2023. 1, 2, 3, 4, 5, 6, 7

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 6

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 1

[32] 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 3592–3601, 2019. 1, 2, 5, 6, 3, 4

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2, 3, 4, 6

[34] Yi-Lin Sung, Varun Nair, and Colin Raffel. Training neural networks with fixed sparse masks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 24193–24205, 2021. 7

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 5

[36] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2023. 4

[37] Guangyue Xu, Parisa Kordjamshidi, and Joyce Chai. Prompting large pre-trained vision-language models for compositional concept learning. *arXiv preprint arXiv:2211.05077*, 2022. 1, 2, 3, 4, 6

[38] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014. 2, 5

[39] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1–9, 2022. 7

[40] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 3

[41] Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. Learning invariant visual representations for compositional zero-shot learning. In *Proceedings of the European Conference on Computer Vision*, pages 339–355, 2022. 2, 6

[42] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34 (12):5586–5609, 2022. 4

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 2337–2348, 2022. 3, 4, 6