GROUNDED COMPOSITIONAL CONCEPT LEARNING

By

Guangyue Xu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science—Doctor of Philosophy

2024

ABSTRACT

Humans learn concepts in a grounded and compositional manner. Such compositional and grounding abilities enable humans to understand an endless variety of scenarios and expressions. Although deep learning models have pushed performance to new limits on many Natural Language Processing and Computer Vision tasks, we still have a lack of knowledge about how these models process compositional structures and their potential to accomplish human-like meaning composition. The goal of this thesis is to advance the current compositional generalization research on both the evaluation and design of the learning models. In this direction, we make the following contributions.

Firstly, we introduce a transductive learning method to utilize the unlabeled data for learning the distribution of both seen and novel compositions. Moreover, we utilize the cross-attention mechanism to align and ground the linguistic concepts into specific regions of the image to tackle the grounding challenge. Unlike traditional learning, we use episodic training where each training item consists of one image and the sampled positive and negative compositional labels. We select the image's compositional label by computing their matching scores Our empirical results show that combining episodic training and transductive learning does help compositional learning.

Secondly, we develop a new prompting technique for compositional learning by considering the interaction between element concepts. In our proposed technique called GIPCOL, we construct a textual input that contains rich compositional information when prompting the foundation vision-language model. We use the CLIP model as the pre-trained backbone vision-language model and improve its compositional zero-shot learning ability with our novel soft-prompting approach. GIPCOL freezes the majority of CLIP's parameters and only learns CLIP's word embedding layer through a graph neural network. By concatenating the learnable soft prompt and the updated word embeddings, GIPCOL achieves better results compared with other prompting-based methods.

Thirdly, since retrieval plays a critical role in human learning, our work studies how retrieval can help compositional learning. We propose MetaReVision which is a new retrieval-enhanced meta-learning model to address the visually grounded compositional concept learning problem. Given an image with a novel compositional concept, MetaReVision first uses a retrieval module

to find relevant items from the training set. Then it constructs an episode for which the retrieved items form the support set and the test item forms the query set. The retrieved support set mimics the primitive concept learning scenario, while the query set encourages the compositional strategy learning by meta-learning's bi-level optimization objective. The experimental results show that such retrieval-enhanced meta-learning framework helps the vision-language model's compositional learning. Moreover, we create two new benchmarks called CompCOCO and CompFlickr for the evaluation of grounded compositional concept learning.

Finally, we evaluate the large generative vision and language models in solving compositional zero-shot learning within the in-context learning framework. We highlight their shortcomings and propose retriever and ranker modules to improve their performance in addressing this challenging problem. These two modules select the most informative in-context examples in their most effective order to guide the backbone generative model. Our approach is novel in the context of grounded compositional learning and our experimental results show improved performance compared to basic in-context learning.

ACKNOWLEDGEMENTS

First and foremost, I am tremendously grateful for my advisor Dr. Parisa Kordjamshidi and Joyce Y. Chai for their continuous support and guidance. They shared with me how to think critically, explore new problems, asking good questions and how to do good research. All of these experiences will have a great influence on my whole life. Besides, their great insights on the domain of large language models and grounded compositional learning have always shed light on problems I have been working on. Without their continuous advice, inspiration and guidance for my PhD. study, this work would have been impossible.

I would also like to thank my dissertation committee members: Dr. Xiaoming Liu and Dr. Taosheng Liu. I greatly appreciate their valuable feedback on every step of my PhD journey.

I'm very happy to have had the opportunity to collaborate with an amazing group of students and researchers: Dr. Shaohua Yang and Dr. Qiaozi Gao provide great suggestions and directions when I start my research career as a PhD student. Thanks to Dr. Sari Saba-Sadiya for his great efforts and enlightening comments. I also appreciate my co-authors on various papers.

I would like to thank all my friends at MSU, who made my time at MSU enjoyable.

Finally, I dedicate this thesis to my family: my parents Pingxian Xu and Jie Zhu, my parents-in-law Junming Gu and Aiju Zhang, my sons Yufeng Xu and Oscar Gu, and my cherished wife Yingjun Gu, for your years of unwavering love and support.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation

Humans acquire language in a compositional and grounded manner. They can understand new scenes and combine known words in novel ways to describe their perceptual world through their compositional and grounding abilities, although these novel compositions may have never been seen before. It would be desirable for intelligent systems to have such compositional generalization ability [Lake et al., 2017]. It is also widely believed that effective semantic representations need to have both compositionality and groundedness as minimum requirements [Carnap, 1988, Baroni and Zamparelli, 2010, Miller and Charles, 1991]. However, recent neural models struggle to generalize outside their training distribution and have difficulties using observed words in a compositional manner, especially in novel situations [Kim and Linzen, 2020]. In recent years, there has been remarkable advancement in large-scale neural network models that can integrate information from both natural language textual and visual data. Despite their impressive progress, the extent to which such large-scale neural network models can effectively encode compositional representations of learned element concepts is still an open question. For instance, correctly identifying a sliced apple when this combination has not been observed by reasoning over its constituents, red and car, is a challenge for such models| [Hupkes et al., 2020, Hermann, 2014, Lake et al., 2015a]. The research conducted in this thesis is an effort to design novel architectures to address some of the challenges of compositional generalization when the models are required to recognize the novel composition of objects and attributes in the visual modality and express it in natural language.

1.2 Compositional Learning

The compositionality is considered as one of the key elements in human intelligence and explained by [Partee et al., 1995] as: *the meaning of a whole is a function of the meanings of the parts and of the way they are syntactically combined.*

However, in terms of computational modeling, compositional learning can have multiple aspects, including primitiveness, systematicity, productivity, and substituivity, which are identified

in cognitive science literature. These aspects are explained in Table 1.1. Compositional abilities of computational models have been widely studied with different lenses using a variety of benchmarks [Chang et al., 2016, Gao et al., 2023, Mancini et al., 2021]. Figure 1.1 provides an overview of multiple aspects in compositional learning, recently proposed compositional benchmarks and related modalities. These benchmarks are proposed to evaluate the compositional ability of different neural networks within different modalities, including natural language processing (NLP), Computer Vision (CV) and vision-language fields. For example, scan is a pure textual compositional learning task that requires models generating action sequences from compositional navigation commands. In the vision-language field, many benchmarks are proposed to measure models' compositional ability via downstream tasks such as question-answering, image-text retrieval, action generation and compositional zero-shot learning (CZSL). In this thesis, we mainly focus on Compositional Zero-Shot Learning benckmarks and study the primitiveness and systematicity in compositional learning.

Aspect	**Description**
Primitiveness	Concept seen in isolation during train can be applied compositionaly at test time.
Systematicity	Generalize to unseen compositions of known elements.
Productivity	Generating longer sequences than those seen in the training data.
Substitutivity	Model robustness when replacing words with synonyms.

Table 1.1 Different aspects of compositional learning identified in compositional generalization literature. In this thesis, we focus on primitiveness and systematicity aspects in compositional learning.

An example of our focused compositional zero-shot learning problem is shown in Figure 1.2. As shown in Figure 1.2(a), suppose the training set has images with compositional concepts sliced-tomato, sliced-cake, ripe-apple, peeled-apple, etc. Given a new image, our goal is to assign a novel compositional concept sliced-apple to the image by composing the element concepts, sliced and apple, learned from the training data. Although sliced and apple have appeared with other objects or attributes, the combination of this attribute-object pair is not observed in the training set. Representative CZSL datasets include MIT-States [Isola et al., 2015a], UT-Zappos [Yu and Grauman, 2014] and C-GQA [Hudson and Manning, 2019]. Based on these datasets, we further

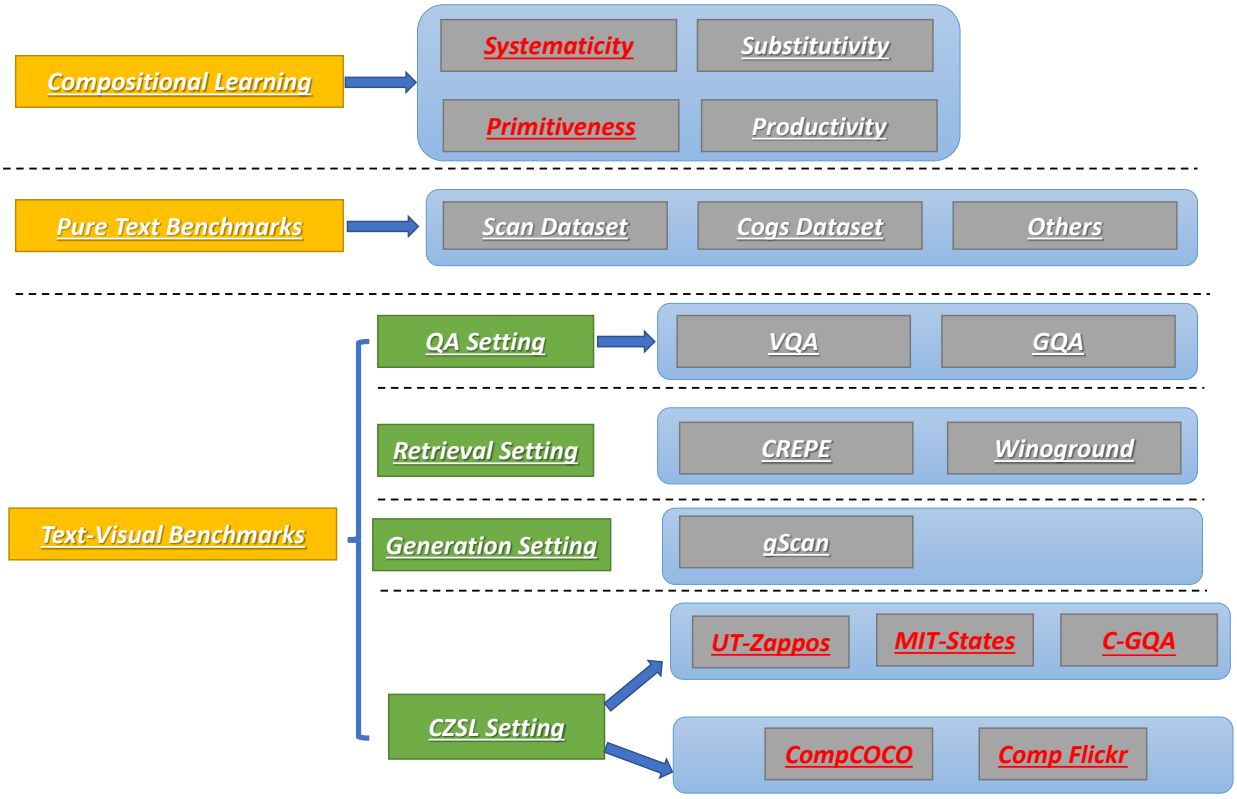


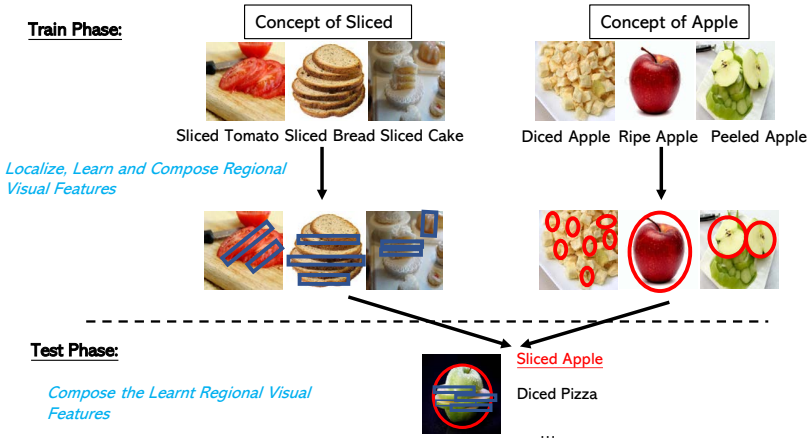
Figure 1.1 The aspects of compositional learning are shown on the top of the figure. Examples of Benchmarks and datasets of different modalities are shown in the rest of this Figure. The compositional aspects and datasets marked in red are the ones we focused on in this thesis.

propose two more CZSL benchmarks , including CompCOCO and CompFlickr as shown in Figure 1.2(b). Different from the previous CZSL benchmarks, these two datasets add more textual information to testify the current deep learning model's compositional learning ability, especially the large visual-language models (VLMs).

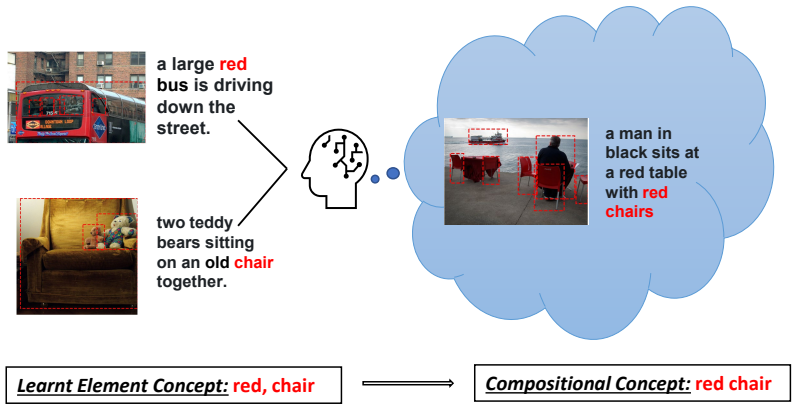
1.3 Challenges of Compositional Learning

Challenge 1: Zero-Shot Learning. Despite the success of deep learning (DL) models, traditional DL models require training on a massive amount of labeled data for each class. However, the distribution of compositional concept samples naturally follows a zero-shot setting: novel compositional concepts do not appear in the training phase. In this respect, collecting large-scale labeled samples to address compositional learning is a challenge. Because there are no training data available for the novel pairs, the learned models will bias to seen pairs.

Challenge 2: Grounded Concept Learning. The second challenge is grounding ability. Ground-



(a)



(b)

Figure 1.2 Compositional learning examples for MITStates and CompCOCO.

ing means the ability to connect words to the real-world entities, events, and ideas that they refer to and it is obviously necessary and fundamental for compositional concept learning. Mostly language models that are trained with huge amounts of data, are not fed with the explicit alignments of occurring words in natural language expression and their real-world manifestations. After the introduction of self-attention, more and more works use the attention weights as indicator for grounding. However, due to the complex self-attention mechanism and the large number of layer and heads in self attention implementation, it is difficult to enforce attention to represent the gronding during training time.

Challenge 3: Capturing the rules of composition. Capturing compositionality and learning principles of compositions in language has been a long-term challenge for neural networks. Most

of the prior work focus on designing new architectures with the guide of explicit compositional structures [Fodor and Pylyshyn, 1988a, Andreas, 2019, Huynh and Elhamifar, 2020]. However, such designs are customized to the task setting and have a limited generalization ability. Current models rely on large amounts of data to capture the encoded patterns of compositionality. Such a framework has difficulty in handing out-of-distribution compositions. In order to address the compositional problem, the models should learn both 1) primitive concepts and 2) the rules for composing them. Current studies show the limitations of data-driven models in generalizing over composing rules.

1.4 Contributions of the Thesis

Contribution 1: Transductive Episodic Training. To address challenges 1 and 3 in CZSL, we propose an episode-based training scheme. We perform model optimization over batches of tasks instead of batches of data. Within this framework, we treat each composition in the training set as a compositional learning task. Through training over multiple tasks, the model is expected to progressively accumulate knowledge on compositional generalization rules and, learn the unseen compositions based on the seen ones within each episode. In addition, we utilize the unlabeled data to augment the supervision for episodic learning and compositional generalization in a transductive learning framework. Experiments have shown the importance of of the transductive learning setting which increase the accuracy by 1.5% in pair accuracy.

Contribution 2: Meta-Learning. To further address challenges of grounding and composing rule learning, we develop a meta-learning framework to train the vision & language models VLMs, we call (MetaReVision), for compositional concept learning. Specifically, MetaReVision uses DG-MAML (Domain-Generalization Model-Agnostic Meta-Learning) proposed by [Li et al., 2018], a variant of Model-Agnostic Meta-Learning (MAML) [Finn et al., 2017], to learn the primitive concepts and the compositional strategy by training through episodes, in a more principled way. In MetaVL, each episode consists of a support set and a query set. The support set mimics the primitive concept learning scenario, while the query set encourages the compositional strategy learning by the DG-MAML's bi-level optimization objective.

Contribution 3: Prompting VLMs for Compositional Learning Given the huge influence of larger pre-trained VLMs in various vision and language tasks [Zhu et al., 2023], our third contribution is to effectively utilize them for compositional concept learning using the prompting methods. We propose a new prompting approach, called GIPCOL, to inject information about the composition of objects and attributes into the prompting design. Specially, we use CLIP as the large VLMs backbone in our experiments and change its hard prompting strategy by combining learnable prefix vectors and element concept vectors. In particular, we achieved SoTA AUC resutls on all three benchmarks.

Contribution 4: In-Context Learning (ICL) for CZSL. Although the generative large models represented by GPT-series [Brown et al., 2020, Achiam et al., 2023] have achieved huge success in many downstream tasks within the in-context learning framework, evaluation and application of such models in a multi-modal problem setting is not straightforward, especially in the zero-shot setting. Main challenges include 1) adapt the current evaluation benchmarks for a sound evaluation of generative large language models for zero-shot compositional learning and, 2) improve foundation models for better compositional generalization by introducing in-context example retriever and ranker modules. To address the above challenges, we propose GenCZSL which introduce the retriever and ranker modules. The retriever is to select informative examples and the ranker is to further sort the retrieved example to help Flamingo recognize the novel compositions.

1.5 Organization of Dissertation

The reminder of this dissertation is organized as follows: in Chapter 2, we introduce background and previous works which this dissertation builds on. In Chapter 3, we present the work on recognizing compositional attribute-object concepts within the zero-shot setting. We propose an episode-based cross-attention (EpiCA) network which combines merits of cross-attention mechanism and episode-based training strategy to recognize novel compositional concepts, which aim to address the grounding and compositional challenges. In Chapter 4, we present MetaReVision, a meta-learning framework to train vision-and-language models for compositional concept learning. The episodic training and the bi-level optimization within the meta-learning framework

encourages gradients learnt from support set to be beneficial for compositional concept learning in the query set, In Chapter 5, we will present PromptCompVL, which explores the compositional zero-shot learning(CZSL) ability of large pre-trained vision-language models(VLMs) within the prompt-based learning framework. PromptCompVL gives a general prompting-based framework for compositional learning and makes two design choices: first, it uses a soft-prompting instead of hard-prompting to inject learnable parameters for compositional learning. Second, it uses the soft-embedding layer to learn primitive concepts in different combinations. In Chapter 6, we explore the possibility of utilizing the in-context learning (ICL) paradigm compositional learning. ICL provides the foundation models, like GPT4 [Achiam et al., 2023] or LLaMa [Touvron et al., 2023], with a few labeled examples as input before asking them to make a prediction on a new example. Different from previous works, we try to address the in-context example selectio for ICL. In Chapter 7, we draw the conclusion of our current state of research and provide the research proposal for the next steps toward completing this PhD thesis. We provide the timelines of the proposed research accordingly.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Compositional Zero-Shot Learning

Compositional learning is the key component of human intelligence and has been widely studied in the deep learning field under the contexts of human-object interactions(HOI) [Kato et al., 2018, Hou et al., 2020], compositional zero-shot learning [Nagarajan and Grauman, 2018b, Misra et al., 2017a], natural language processing [Lake, 2019, Nye et al., 2020] and language acquisition [Jin et al., 2020, Surís et al., 2020]. In this thesis, we study the Compositional Zero-Shot Learning (CZSL) problem and this topic falls into the language acquisition category. As a specific zero-shot learning (ZSL) problem, compositional zero-shot learning (CZSL) tries to learn complex concepts by composing element concepts. Previous solutions can mainly be categorized as:

- **Classifier-based methods** train classifiers for element concepts and combine the element classifiers to recognize compositional concepts [Chen and Grauman, 2014, Misra et al., 2017a, Li et al., 2019a].
- **Metric-based methods** learn a shared space by minimizing the distance between the projected visual features and concept features [Nagarajan and Grauman, 2018b, Li et al., 2020b].
- **Generative-based methods** learn to generate samples from the semantic information and transfer CZSL into a traditional supervised classification problem [Wei et al., 2019].
- **Prompting-based methods** [Nayak et al., 2022] tries to explore the compositional knowledge from the large visual-language models (CLIP)[Radford et al., 2021] by constructing the textual prompting input.

In our work, we try to address the key challenges in CZSL, including 1) grounding, 2) compositional rule learning, 3) zero-shot setting, and based on different visual-language models (VLMs) proposing novel parameter-efficient methods to solve the CZSL problem. Moreover, we contribute two more realistic datasets CompFlickr and CompCOCO which give more textual information for CZSL and the realted prolem settings are as below:

- For UT-Zappos, MIT-Staes and C-GQA datasets, the textual input is acting as the class labels. In this CZSL setting, given an image, we need to retrieve or generate the most relevant pair label. As a zero-shot learning problem, the pair label has never been seen during training time and therefore we can formulate CZSL as a open-vocabulary problem. CLIP [Radford et al., 2021] and Flamingo [Alayrac et al., 2022] can be utilized for their zero-shot or few-shot learning ability.
- For CompFlickr and CompCOCO datasets, because we have more textual input as our contextual information, we formulate CZSL as masked token predictio problem. Then we can utilize VLMs, like VL-BERT [Su et al., 2020], as multi-modal encoder to help predict the masked compositional concepts. In this setting, we modify the VLMs in two ways: 1) add retrieval module to retrieve related element concepts to construct episodes. 2) meta-train such VLMs [Finn et al., 2017] to accumulate compositional knowledge from the constructed compositional tasks.

2.2 Large Foundation Models

Large-scale datasets, self-supervision training technique, and attention mechanism [Vaswani et al., 2017a] have led to the emergence of powerful uni-modal encoders for images [Dosovitskiy et al., 2020], videos [Arnab et al., 2021], language models [Devlin et al., 2019] and other modalities [Girdhar et al., 2022]. These uni-modal encoders form the basis for large vision-language models (VLMs). Popular VLMs such as CLIP [Radford et al., 2021] and ALIGN [Jia et al., 2021] are trained using the above uni-modal encoders and fusing multi-modal information from the massive web datasets in form of images and alt-text. In this section, we mainly introduce three VLMs related to our thesis, including VL-Bert [Su et al., 2020], CLIP [Radford et al., 2021] and Flamingo [Alayrac et al., 2022] separatelyas below:

2.2.1 Generic Vision-Language Encoder: VL-BERT

VL-BERT [Su et al., 2020] is designed to extract generic representation for visual-linguistic tasks through pre-trained tasks, including masked language modeling (MLM) and masked RoI classification. Such pre-trained models are expected to have a joint understanding of image features

and language phrases that correspond to them. Specially, after extracting visual tokens using Fast R-CNN [Girshick, 2015] from images and textual tokens from texts, VL-BERT adopts the Transformer model [Vaswani et al., 2017a] as the backbone to extract multi-modal representation from massive-scale Conceptual Captions dataset [Sharma et al., 2018], together with text-only corpus. VL-BERT follows the pre-training and fine-tuning framework. After obtaining the generic representation for vision-lanuge tasks, for each downstream tasks, like Visual Question Answering (VQA) [Antol et al., 2015], Visual Commonsense Reasoning (VCR) [Zellers et al., 2019] and Referring Expression task [Yu et al., 2016]. In our work, we aim to explore the compositional ability of VL-BERT's vision-language representation. In order to achieve this goal, we propose two new benchmarks for compositional learning testing and our experiments on these two benchmarks show that the extracted representations have difficulty for representing novel compositions. Furthermore, we propose a new framework which combines retrieval and meta-learning to enhance VL-BERT and similar models' compositional ability which is detailed in Chapter 5.

2.2.2 Contrastive Image-Text Pretraining: CLIP

The recently released contrastively trained vision-language model, CLIP [Radford et al., 2021], has enabled a diverse of downstream applications at the intersection of Computer Vision (CV) and Natural Language Processing (NLP) fields in the form of Language Guided Vision Processing [Huang et al., 2023a, Zhang et al., 2023, Huang et al., 2023b]. Pre-training using 400 million of image-text pairs, CLIP-based models have demonstrated remarkable zero-shot capabilities [Ma et al., 2023a]. Moreover, through the pre-trained visual encoder, textual encoder and the latent space which align images and texts, CLIP provide many downstream application scenarios. 1) In CV, its pre-trained visual and textual encoders have been used for semantic segmentation, object detection and image captioning [Rao et al., 2022a]. 2) In diffusion models, CLIP has been used as a loss and acted as an automated evaluation metric [Hessel et al., 2021]. 3) In feature extractor, CLIP has been incorporated into architectures for various tasks, such as video summarization [Xu et al., 2021]. In our work, we aim to study and improve CLIP's compositional ability within using prompting paradigm. We conduct our experiments on three compositional datasets, including

MIT-States [Isola et al., 2015a], UT-Zappos [Yu and Grauman, 2014] and CGQA [Hudson and Manning, 2019] and find that improved prompting design can help CLIP's compositional learning which is detailed in Chapter 4.

2.2.3 Few-Shot Vision-Language Model: Flamingo

In order to utilize the increasing ability of large VLMs, in-context learning (ICL) has become a new paradigm for multi-modal tasks [Brown et al., 2020]. However, most VLMs only accept one image and utilized this single input image for downstream tasks [Bugliarello et al., 2020, Bagad et al., 2023]. Such VLMs can not be directly used in ICL's compositional learning since ICL requires multiple images as demonstration input in compositional learning. Recently proposed Flamingo [Alayrac et al., 2022] can consume sequences of arbitrarily interleaved visual and textual data as input in few-shot setting. It introduce two components to address the arbitrarily interleaved challenge: 1) perceiver which uses query vectors to fuse and compress visual input and produce a small fixed number of visual tokens per image, 2) cross-attention mechanism to fuse the multi-modal information from the query vectors. However, different from Flamingo's few-shot application, we need to retrieve and construct episodes in compositional learning. We will discuss the episode construction and optimiazaion in Chapter 6.

2.3 Parameter-Efficient Paradigm For Applying Large Models

The full-model fine-tuning (FT) for large language models (LLM) is expensive and could affect the learnt knowledge acquired during the large scale pre-training phase [Sun et al., 2023]. Therefore, more parameter-efficient techniques are recently explored to increase the accessibility of large models. In this section, we give a detailed discussion about the parameter-efficient fine-tuning methods and talk about these methods' application in compositional learning.

2.3.1 Prompt-Based Learning

Prompt-based learning is an emerging technique originated from NLP field. Different from traditional supervised fine-tuning techniques, prompting-based methods freeze most parts of the large pre-trained NLP model, like T5 [Raffel et al., 2020] and GPT [Brown et al., 2020], and concatenate a small number of additional learnable parameters to the test input which learns to

solve downstream tasks [Liu et al., 2021] as Equation 4.2.

$$\text{Input}_{\text{PT}} = \text{concat}\left(\mathbf{P}; \mathbf{X}_{\text{test}}\right). \tag{2.1}$$

where P is the learnable embeddings. Because of these learnable embeddings, prompt-based learning requires access to a training set $\mathbf{X}_{\text{train}}$ for the target downstream task.

As the prevalence of large pre-trained visual-language(VL) models, prompting-based methods are introduced to explore the multi-modal knowledge encoded in such VLMs [Tsimpoukelli et al., 2021, Radford et al., 2021, Jin et al., 2021]. Recently, [Zhou et al., 2022a] and [Zhou et al., 2022b] prompt CLIP by prepending learnable parameters to text input for low-resource image classification and achieves satisfactory resutls. Meanwhile, [Nayak et al., 2022] conducted compositional learning by modifying CLIP's original vocabulary embeddings and shows the possibility of prompting VL models for compositional learning. Our work proposes a novel prompting strategy to further imporve CLIP's compositional learning ability.

2.3.2 In-Context Learning

In-context learning (ICL) is an important paradigm for adapting LLM and VLMs to new tasks which is first introduced by [Brown et al., 2020]. Different from prompt-based learning, ICL paradigm enables the adaptation of these large models to new tasks by prompting them with instructions (zero-shot) or demonstrations (few-shot) without any additional learnable parameters as shown in Equation 6.1.

$$\text{Input}_{\text{ICL}} = \text{concat}\left([\mathbf{X}_{icl}; \mathbf{Y}_{icl}]_1^k; \mathbf{X}_{\text{test}}\right) \tag{2.2}$$

where $[\mathbf{X}_{icl}; \mathbf{Y}_{icl}]_1^k$ are the k demonstraing examples.

Compared with traditional learning paradigms, ICL has several advantages. First, data efficiency, the ability to do few-shot learning directly reduces the need for human-labeled data. Second, computing efficiently, in contrast to other popular training paradigms, ICL enables inference without any gradient updates. Lastly, good performance, ICL also displays amazing versatility through different modes of prompting. However, ICL's performance is highly sensitive to prompting

input and three key components affect its performance, including example selection, example order and template design [Nguyen and Wong, 2023]. In this thesis, we explore ICL for compositional learning. Specially, we focus on example selection and ranking to improve VLMs' compositional ability.

2.4 Meta Learning

Humans learn in a compositional manner from their previous experience [Fodor, 1975]. This process could be formalized within meta-learning framework. Meta learning, also known as learning to learn, deal with the problem of efficient learning so that they can learn new concepts or skills fast with just a few seen examples (few-shot setting) or even no seen examples (zero-shot setting). It aims to solve a low-resource problem by leveraging the learnt experience from a set of related tasks. Through learning from the compositional tasks, meta-learning could be used to learn the compositional rules.There are mainly three categories of meta-learning methods: 1) Metric-based methods learn a metric or distance function over tasks [Sung et al., 2018a, Snell et al., 2017b]. 2) Model-based methods aim to design an architecture or a training process for rapid adaption across tasks [Ravi and Larochelle, 2016, Munkhdalai et al., 2018]. 3) Optimization-based methods directly adjust the optimization algorithm to enable quick adaptation with just a few examples [Nichol et al., 2018a, Finn et al., 2017]. Meta learning has also been widely deployed in NLP field [Gu et al., 2018, Dou et al., 2019, Holla et al., 2020] recently to address the low-resource language processing problems. In this thesis, we use optimization-based meta-learning methods to learn the generalizable initialization for CZSL by training on the constructed episodes. This process tries to mimic the human's compositional learning process and the compositional knowledge is encoded in the learnt parameter initialization.

CHAPTER 3

ZERO-SHOT COMPOSITONAL CONCEPT LEARNING

In this thesis, we study the problem of recognizing compositional attribute-object concepts within the zero-shot learning (ZSL) framework. We propose an episode-based cross-attention (EpiCA) network that combines the merits of the cross-attention mechanism and episode-based training strategy to recognize novel compositional concepts. Firstly, EpiCA bases on cross-attention to correlate concept-visual information and utilizes the gated pooling layer to build contextualized representations for both images and concepts. The updated representations are used for a more in-depth multi-modal relevance calculation for concept recognition. Secondly, a two-phase episode training strategy, especially the transductive phase, is adopted to utilize unlabeled test examples to alleviate the low-resource learning problem. Experiments on two widely-used zero-shot compositional learning (ZSCL) benchmarks have demonstrated the effectiveness of the model compared with recent approaches on both conventional and generalized ZSCL settings [1].

3.1 Introduction

Humans can recognize novel concepts through composing previously learned knowledge - known as compositional generalization ability [Lake et al., 2015b, Lake and Baroni, 2018]. As a key critical capacity to build modern AI systems, this thesis investigates the problem of zero-shot compositional learning (ZSCL) focusing on recognizing novel compositional attribute-object pairs appeared in the images. For example in Figure 5.1, suppose the training set has images with compositional concepts sliced-tomato, sliced-cake, ripe-apple, peeled-apple, etc. Given a new image, our goal is to assign a novel compositional concept sliced-apple to the image by composing the element concepts, sliced and apple, learned from the training data. Although sliced and apple have appeared with other objects or attributes, the combination of this attribute-object pair is not observed in the training set.

This is a challenging problem because objects with different attributes often have a significant

[1] Zero-Shot Compositional Concept Learning. Guangyue Xu, Parisa Kordjamshid, Joyce Chai. MetaNLP@ACL, 2021

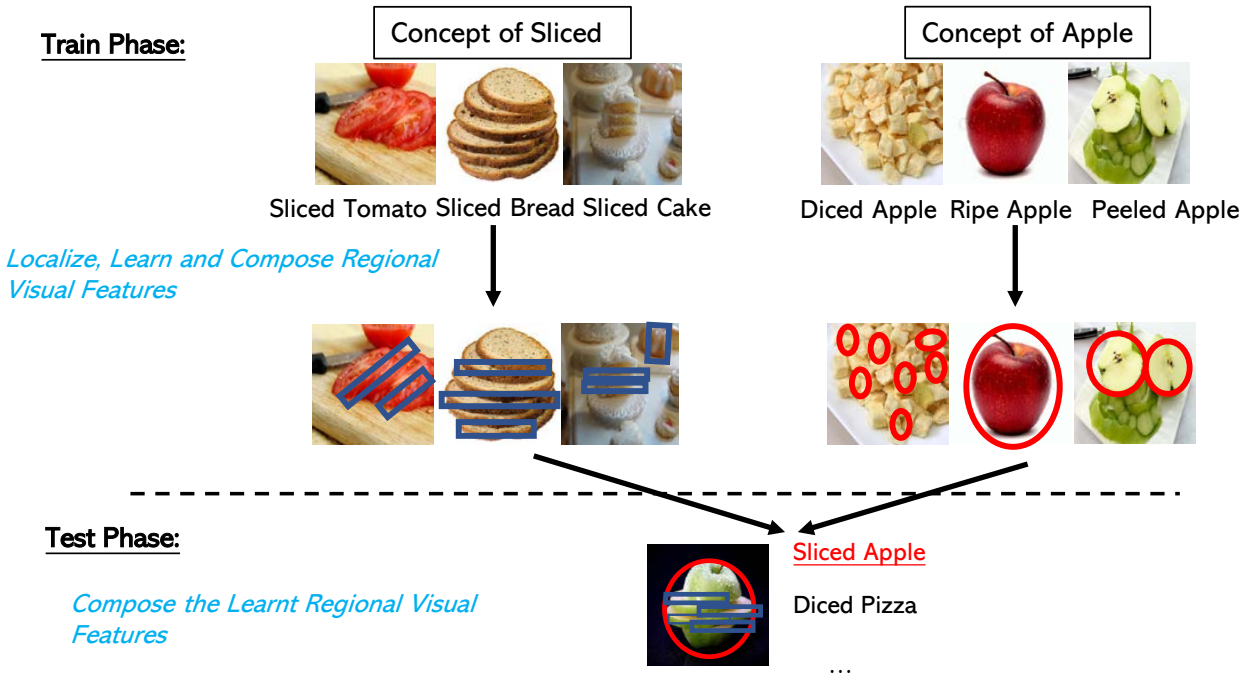


Figure 3.1 Given the concepts of sliced and apple in the training phase, our target is to recognize the novel compositional concept slice apple which doesn't appear in the training set by decomposing, grounding and composing concept-related visual features.

diversity in their visual features. While red apple has similar visual features as the apple prototype, sliced apple presents rather different visual features as shown in Fig 5.1. Similarly, the same attributes can have different visual effects depending on the modified objects. For example, old has a different visual effect in objects of old town compared to objects of old car. Despite recent progress [Misra et al., 2017b, Li et al., 2020c], previous works still suffer several limitations: (1) Most existing methods adopt metric learning framework by projecting concepts and images into shared latent space, and focus on regularizing the structure of the latent space by adding principled constraints without considering the relationship between concepts and visual features. Our work brings a new perspective, the relevance-based framework inspired by [Sung et al., 2018b], to conduct compositional concept learning. (2)Previous works represent concept and image by the same vector regardless of the context it occurs. However, cross concept-visual representation often provides more grounded information to help in recognizing objects and attributes which will consequently help in learning their compositions.

Motivated by the above discussions, we propose an Episode-based Cross Attention (EpiCA)

network to capture multi-modal interactions and exploit the visual clues to learn novel compositional concepts. Specifically, within each episode, we first adopt cross-attention encoder to fuse the concept-visual information and discover possible relationships between image regions and element concepts which corresponds to the localizing and learning phase in Fig.5.1. Second, a gated pooling layer is introduced to obtain the global representation by selectively aggregating the salient element features corresponding to Fig. 5.1's composing phase. Finally, relevance score is calculated based on the updated features to update EpiCA.

The contribution of this work can be summarized as follows: 1) Different from previous work, EpiCA has the ability to learn and ground the attributes and objects in the image by cross-attention mechanism. 2) Episode-based training strategy is introduced to train the model. Moreover, we are among the first works to employ transductive training to select confident unlabelled examples to gain knowledge about novel compositional concepts. 3) Empirical results show that our framework achieves competitive results on two benchmarks in conventional ZSCL setting. In the more realistic generalized ZSCL setting, our framework significantly outperforms SOTA and achieves over 2× improved performance on several metrics.

3.2 Related Work

Compositional Concept Learning. As a specific zero-shot learning (ZSL) problem, zero-shot compositional learning (ZSCL) tries to learn complex concepts by composing element concepts. Previous solutions can mainly be categorized as: (1) classifier-based methods train classifiers for element concepts and combine the element classifiers to recognize compositional concepts [Chen and Grauman, 2014, Misra et al., 2017b, Li et al., 2019a]. (2) metric-based methods learn a shared space by minimizing the distance between the projected visual features and concept features [Nagarajan and Grauman, 2018a, Li et al., 2020c]. (3) GAN-based methods learn to generate samples from the semantic information and transfer ZSCL into a traditional supervised classification problem [Wei et al., 2019].

Attention Mechanism. The attention mechanism selectively use the salient elements of the data to compose the data representation and is adopted in various visiolinguistic tasks. Cross

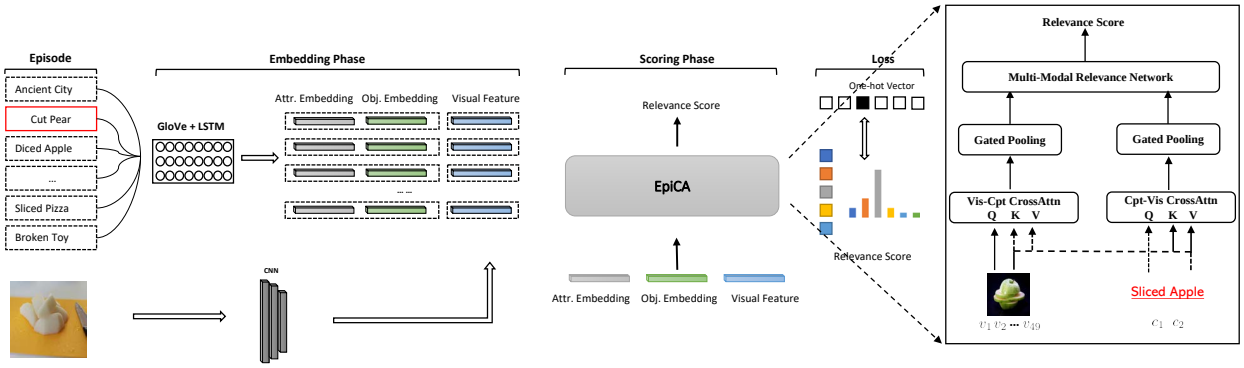


Figure 3.2 Illustration of the proposed EpiCA framework. It is a two-stage training framework, including inductive learning and transductive learning. Both phases are trained on episodes as illustrated in Alg. 1.

attention is employed to locate important image regions for text-image matching [Lee et al., 2018]. Self-attention and cross-attention are combined at different levels to search images with text feedback [Chen et al., 2020b]. More recent works refer to Transformer [Vaswani et al., 2017b] to design various visiolinguistic attention mechanism [Lu et al., 2019].

Episode-based Training. The data sparsity in low-resource learning problems, including few-shot learning and zero-shot learning, makes the typical fine-tuning strategy in deep learning not adaptable, due to not having enough labeled data and the overfitting problem. Most successful approaches in this field rely on an episode-based training scheme: performing model optimization over batches of tasks instead of batches of data. Through training multiple episodes, the model is expected to progressively accumulate knowledge on predicting the mimetic unseen classes within each episode. Representative work includes Matching network [Vinyals et al., 2016], Prototypical network [Snell et al., 2017a] and RelNet [Sung et al., 2018b].

The related works to EpiCA are RelNet [Sung et al., 2018b] and cvcZSL [Li et al., 2019a]. Compared with these methods, we have two improvements including an explicit way to construct episodes which is more consistent with the test scenario and a cross-attention module to fuse and ground more detailed information between the concept space and the visual space.

3.3 Approach

3.3.1 Task Definition

Different from the traditional supervised setting where training concepts and test concepts are from the same domain, our problem focuses on recognizing novel compositional concepts of attributes and objects which are not seen during the training phase. Although we have seen all the attributes and objects in the training set, their compositions are novel [2].

We model this problem within the ZSL framework where the dataset is divided into the seen domain $\mathcal{S} = \{(v_s, y_s)|v_s \in \mathcal{V}^s, y_s \in \mathcal{Y}^s\}$ for training and the unseen domain $\mathcal{U} = \{(v_u, y_u)|v_u \in \mathcal{V}^u, y_u \in \mathcal{Y}^u\}$ for test, where v is the visual feature of image $\mathcal{I}$ which can be extracted using deep convolution networks and y is the corresponding label which consists of an attribute label a and a object label o as $y = (a, o)$ satisfying $a_u \subseteq a_s$, $o_u \subseteq o_s$ and $\mathcal{Y}_s \cap \mathcal{Y}_u = \phi$. Moreover, we address the problem in both conventional ZSCL setting and generalized ZSCL setting. In conventional ZSCL, we only consider unseen pairs in the test phase and the target is to learn a mapping function $\mathcal{V} \mapsto \mathcal{Y}^u$. In generalized ZSCL, images with both seen and unseen concepts can appear in the test set, and the mapping function changes to $\mathcal{V} \mapsto \mathcal{Y}^s \cup \mathcal{Y}^u$ which is a more general and realistic setting.

3.3.2 Overall Framework

As summarized in Fig. 5.4, EpiCA consists of the cross-attention encoder, gated pooling layer and multi-modal relevance network to compute the relevance score between concepts and images. In order to accumulate the knowledge between images and concepts, EpiCA is trained by episodes including the following two phases:

- Inductive training phase constructs episodes from the seen concepts and trains EpiCA based on these constructed episodes.
- Transductive training phase employs the self-taught methodology to collect confident pseudo-labeled test items to further fine-tune EpiCA.

[2] We refer concept as compositional concept, element concept as the attribute and the object in the rest of the thesis.

3.3.3 Unimodal Representation

Concept Representation. Given a compositonal concept (a, o), we first transform attribute and object using 300-D GloVe [Pennington et al., 2014a] separately. Then we use one layer BiLSTM [Hochreiter and Schmidhuber, 1997] to obtain contextualized representation for concepts with d_k hidden units. Instead of using the final state, we maintain the output features for both attribute and object and output feature matrix $C \in \mathbb{R}^{2\times d_k}$ for each compositional concept.

Image Representation. We extract the visual features using pretrained ResNet [He et al., 2016] from a given image. In order to obtain more detailed visual features for concept recognition, we keep the output from the last convolutional layer of ResNet-18 to represent the image and therefore each image is split into $7 \times 7 = 49$ visual blocks with each block as a 512-dim vector denoted as $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{49})$. Each element represents a region in the image. We further convert v_i with a linear transformation $v_i = \mathbf{W}^\top v_i$, where $\mathbf{W} \in \mathbb{R}^{512\times d_k}$ is the weight matrix to transfer the image into the joint concept-image space.

3.3.4 Cross Attention Encoder

Motivation. Previous works usually utilize vector representation for both concepts and images and construct a metric space by pushing aligned images and concepts closer to each other. The potential limitation of such frameworks is that the same vector representations without context information will miss sufficient detailed information needed for grounding and recognizing objects and attributes appeared in the images. We observe that certain visual blocks in the image can be more related to certain element concept and certain element concept may highlight different visual blocks. Inspired by this observation, our model addresses the previous limitation by introducing cross-attention encoder and constructs more meaningful cross-modality representation for both images and element concepts for compositional concept recognition.

Cross Attention Layer. To fuse and ground information between visual space and concept space, we first design a correlation layer to calculate the correlation map between the two spaces, which is used to guide the generation of the cross attention map. Given an image and a candidate concept, after extracting unimodal representations, the correlation layer computes the semantic

relevance between visual blocks $\{v_i\}_{i=1}^{49}$ and element concepts $\{c_j\}_{j=1}^{2}$ [3] with cosine distance and output the final image-to-concept relevance matrix as $R \in \mathbb{R}^{49\times 2}$ with each element r_{ij} calculated using Eq. 3.1. We can easily have another concept-to-image relevance matrix by transposing R.

$$r_{ij} = \left(\frac{v_i}{\|v_i\|_2}\right)^T \left(\frac{c_j}{\|c_j\|_2}\right), i \in [1, 49], j \in [1, 2] \tag{3.1}$$

In order to obtain attention weights, we need to normalize the relevance score r_{ij} as Eq. 3.2 as [Chen et al., 2020a].

$$\bar{r}_{ij} = \frac{\text{relu}\left(r_{ij}\right)}{\sqrt{\sum_{j=1}^{n} \text{relu}\left(r_{ij}\right)^2}} \tag{3.2}$$

After obtaining the normalized attention score, we can calculate the cross-attention representation based on the selected query space Q and the context space V, where $V = K$ in our setting as shown in Fig. 5.4. Taking image-to-concept attention for example, given a visual block feature v_i as query, cross attention encoding is performed over the element concept space C using Eq. 3.3.

$$\widehat{v}_i = \sum_{j=1}^{n} \alpha_{ij} c_j, \quad \text{s.t.} \quad \alpha_{ij} = \frac{\exp\left(\lambda \bar{r}_{ij}\right)}{\sum_{j=1}^{n} \exp\left(\lambda \bar{r}_{ij}\right)} \tag{3.3}$$

where λ is the inverse temperature parameter of the softmax function [Chorowski et al., 2015] to control the smoothness of the attention distribution.

Visually-Attended Concept Representation. The goal of this module is to align and represent concepts with related visual blocks and help further determine the alignment between element concepts and image regions. We use concept embedding as query and collect visual clues using Eq. 3.3 and the final visually-attended features for compositional concept is $\widehat{c} \in R^{2\times d_k}$.

Concept-Attended Visual Representation. An image representation grounded with element concept would be beneficial for compositional concept learning. Following the similar procedure as visually-attended concept representation, we take visual block features as query and concept embedding as context. We can calculate the concept-attended visual representation using Eq. 3.3.

[3] Each compositional concept only has two elements, attribute and object.

The final result $\widehat{v} \in \mathbb{R}^{49\times d_k}$ represents the concept-attended block visual features with the latent space dimension d_k.

3.3.5 Gated Pooling Layer

After the cross-attention encoder, the output image features $V = [v_1, \ldots, v_{49}] \in \mathbb{R}^{49\times d_k}$ and concept features $C = [c_1, c_2] \in \mathbb{R}^{2\times d_k}$ are expected to contain rich cross-modal information. Our target of gated pooling layer is to combine elements to form the final representation for concepts and images separately. Pooling techniques can be directly deployed to obtain such representation. However, we argue that elements should have different effect on the final concept recognition. For example, background visual blocks shouldn't be paid much attention during concept recognition. To address the assumption, we propose gated pooling layer to learn the relative importance of each element and dynamically control the contribution of each element in the final representation. Specially, We apply one linear layers with parameter $W \in \mathbb{R}^{d_k\times 1}$ on the element feature x_i and normalize the output to calculate an attention weight α_i that indicates the relative importance of each element using Eq. 3.4.

$$x = \textstyle\sum_i \alpha_i x_i \quad \text{s.t.} \quad \alpha_i = \frac{\exp((Wx_i))}{\sum_{k=1}^{N}\exp((Wx_k))} \tag{3.4}$$

3.3.6 Multi-Modal Relevance Network

After obtaining the updated features for both images $\widehat{v}_i$ and concepts $(\widehat{a}, \widehat{o})_j$, we introduce the multimodal relevance network shared the spirit as [Sung et al., 2018b] to calculate the relevance score as shown in Eq. 3.5

$$s_{i,j} = g_\phi\left(\text{concat}[(\widehat{v}_i), (\widehat{a}, \widehat{o})_j]\right) \tag{3.5}$$

where g is the relevance function implemented by two layer feed-forward network with trainable parameters ϕ.

In order to train EpiCA, we add Softmax activation on the relevance score to measure the probability of image i belonging to concept j within the current episode as Eq. 3.5 and update

Algorithm 1: Training EpiCA for ZSCL:

Input: $\mathcal{D}_{train} = \{(v_m, (a_m, o_m)\}_{m=1}^{|Tr|}$, $\mathcal{D}_{test} = \{v_n\}_{i=n}^{|Ts|}$, task size S, sample interval t
Output: Multi-Modal Rel. Function f

```
// Inductive Learning Phase
for epoch ← 1 to E_ind_max do
    for each image and the corresponding pair in the training set do
        Construct an episode [v_p, (a_p, o_p), (a_n1, o_n1), ··· , (a_ns, o_ns)].
        Gated Cross-Attention Encoding using Eq. 3.1, 3.2, 3.3 and 3.4
        Calculating multi-modal relevance score using Eq 3.5.
        Updating EpiCA.
    end
end
// Transductive Learning Phase
for epoch ← 1 to E_trans_max do
    if epoch % t == 0 then
        Pick confident samples from unseen set by Eq. 3.7.
    end
    Updating EpiCA by Eq 3.9.
end
```

EpiCA using cross-entropy loss.

$$p_j(\widehat{v}_i) = \frac{\exp(s_{i,j})}{\sum_{k=1}^{C} \exp\left(s_{i,k}\right)} \tag{3.6}$$

3.3.7 Training and Prediction

Inductive Training. For each image and the corresponding pair label, we randomly sample negative pairs to form an episode which consists of an image v_p, a positive pair (a_p, o_p) and a predefined number n_t of negative pairs in the form of $[v_p, (a_p, o_p), (a_{n_1}, o_{n_1}), \cdots, (a_{n_t}, o_{n_t})]$. Then within each episode, we calculate the relevance score between image and all candidate pairs using Eq. 3.5. Finally, we calculate the cross entropy loss using Eq. 3.6 and update EpiCA as shown in Alg. 1.

Transductive Training. The disjointness of the seen/unseen concept space will result in domain shift problems and cause the predictions biasing towards seen concepts as pointed by [Pan and Yang, 2009]. Transductive training utilizes the unlabeled test set to alleviate the problem [Dhillon et al., 2019]. Specifically, transductive training has a sampling phase to select confident test samples and

utilize the generalized cross entropy loss as Eq. 3.8 to update EpiCA.

Following previous work [Li et al., 2019b], we use threshold-based method as Eq. 3.7 to pick up confident examples.

$$\frac{p_1(\widehat{v}_i)}{p_2(\widehat{v}_i)} > \gamma \tag{3.7}$$

where p is calculated by Eq. 3.6 and the threshold is the fraction of the highest label probability $p_1(\widehat{v}_i)$ and the second highest label probability $p_2(\widehat{v}_i)$ which measures the prediction peakiness in current episode. Only confident instances are employed to update EpiCA which is controlled by γ.

Moreover, the recently proposed generalized cross-entropy loss [Zhang and Sabuncu, 2018] is used to calculate the loss for pseudo-labeled test examples as Eq. 3.8.

$$\mathcal{L}_u = \sum_{(v_i,(a,o)_j)\in\mathcal{U}} \frac{1-(p_j(\widehat{v}_i))^q}{q} \tag{3.8}$$

where $p_j(\widehat{v}_i)$ is the probability of $\widehat{v}_i$ belonging to pair $(\widehat{a},\widehat{o})_j$ calculated using Eq. 3.6. $q \in (0,1]$ is the hyper-parameter related to the noise level of the pseudo labels, with higher noisy pseudo labels requiring larger q.

Finally, the transductive loss is calculated as Eq. 3.9, where $\mathcal{L}_u$ corresponds to the generalized cross entropy loss from pseudo-labeled test examples and $\mathcal{L}_s$ is the cross entropy loss for the training examples

$$\mathcal{L} = \mathcal{L}_u + \mathcal{L}_s. \tag{3.9}$$

Prediction. Given a new image with extracted feature v_i, we iterate over all the candidate pairs and select the pair with the highest relevance score as $(\hat{a},\hat{o}) = \text{argmax}_{\hat{a},\hat{o}}\, s_{i,j}(\hat{v}_i,(\hat{a},\hat{o})_j)$ as Eq. 3.5 using EpiCA.

3.4 Experiments

Dataset. We use similar dataset as in [Nagarajan and Grauman, 2018a, Purushwalkam et al., 2019a] for both conventional and generalized ZSCL settings with the split shown in Tab. 3.1. Notably, generalized ZSCL setting has additional validation set for both benchmarks which allows

cross-validation to set the hyperparameters. The generalized ZSCL evaluates the models on both seen/unseen sets.

- MIT-States [Isola et al., 2015b] has 245 objects and 115 attributes. In conventional ZSCL, the pairs are split into two disjoint sets with 1200 seen pairs and 700 unseen pairs. In generalized ZSCL, the validation set has 600 pairs with 300 pairs seen in the training set and 300 pairs unseen during training and the test set has 800 pairs with 400 pairs seen and remaining 400 pairs unseen in the training set.
- UT-Zappos [Yu and Grauman, 2017] contains images of 12 shoe types as object labels and 16 material types as attribute labels. In conventional ZSCL, the dataset is split into disjoint seen set with 83 pairs and unseen set with 33 pairs. In generalized ZSCL, the 36 pairs in the test set consists 18 seen and 18 unseen pairs. 15 seen pairs and 15 unseen pairs composes the validation set.

Implementation Details. We develop our model based on PyTorch. For all experiments, we adopt ResNet-18 pre-trained on ImageNet as the backbone to extract visual features. For attr-obj pairs, we encode attributes and objects with 300-dim GloVe and fix it during the training process. We randomly sample 50 negative pairs to construct episodes. We use Adam with 10^{-3} as the initial learning rate and multiply the learning rate by 0.5 every 5 epoch and train the network for total 25 epochs. We report the accuracy at the last epoch for conventional ZSCL. For generalized ZSCL, the accuracy is reported based on the validation set. Moreover, the batch size is set to 64, λ in Eq. 3.3 is set to 9, q in Eq. 3.8 is set to 0.5 and the threshold in Eq. 3.7 is set to 10. [4]

Baselines. We compare EpiCA with the following SOTA methods: 1) Analog [Chen and Grauman, 2014] trains a linear SVM classifier for the seen pairs and utilizes Bayesian Probabilistic Tensor Factorization to infer the unseen classifier weights. 2) Redwine [Misra et al., 2017b] leverages the compatibility between visual features v and concepts semantic representation to do the recognition. 3) AttOperator [Nagarajan and Grauman, 2018a] models composition by treating attributes as matrix operators to modify object state to score the compatibility. 4) GenModel [Nan et al., 2019]

[4] Our code is publicly available at: https://github.com/HLR/CrossAttnCptLearn

	Conventional ZSCL		Generalized ZSCL	
	MIT-States	Zappos	MIT-States	Zappos
# Attr.	115	16	115	16
# Obj.	245	12	245	12
# Train Pair	1262	83	1262	83
# Train Img.	34562	24898	30338	22998
# Test Pair	700	33	800	36
# Test Img.	19191	4228	12995	2914
# Val. Pair			600	30
# Val. Img.			10420	3214

Table 3.1 Data Statistics about Conventional and Generalized Data Split for MIT-States and UT-Zappos Datasets.

adds reconstruction loss to boost the metric-learning performance. 5) TAFE-Net [Wang et al., 2019] extracts visual features based on the pair semantic representation and utilizes a shared classifier to recognize novel concepts. 6) SymNet [Li et al., 2020c] builds a transformation framework and adds group theory constraints to its latent space to recognize novel concepts. We report the results according to the above baseline papers and the released official code [5] [6] of the aforementioned baselines.

Methods	MIT-States(%)	UT-Zappos(%)
Random	0.14	3.0
ANALOG	1.4	18.3
REDWINE	12.5	40.3
ATTOPERATOR	14.2	46.2
GenModel	17.8	48.3
TAFE-Net	16.4	33.2
SymNet	**19.9**	52.1
EpiCA(Inductive)	15.68	**52.56**
EpiCA(Transductive)	18.13	**55.48**

Table 3.2 Results of Conventional ZSCL setting.

3.4.1 Conventional ZSCL Setting

Quantitive Results. Top-1 accuracy metric is reported in this setting to compare different methods. The top-1 accuracy of the unseen attr-obj pairs for conventional ZSCL is presented in Tab. 4.3.

[5]https://github.com/Tushar-N/attributes-as-operators

[6]https://github.com/ucbdrive/tafe-net.git

	Mit-States						UT-Zappos					
	Val AUC			Test AUC			Val AUC			Test AUC		
Model Top k →	1	2	3	1	2	3	1	2	3	1	2	3
AttOperator	2.5	6.2	10.1	1.6	4.7	7.6	21.5	44.2	61.6	25.9	51.3	67.6
RedWine	2.9	7.3	11.8	2.4	5.7	9.3	30.4	52.2	63.5	27.1	54.6	68.8
LabelEmbed+	3.0	7.6	12.2	2.0	5.6	9.4	26.4	49.0	66.1	25.7	52.1	67.8
TMN	3.5	8.1	12.4	2.9	7.1	11.5	36.8	57.1	69.2	29.3	**55.3**	69.8
SymNet	4.3	9.8	14.8	3.0	7.6	12.3	‖	‖	‖	‖	‖	‖
Inductive EpiCA	7.73	12.19	22.93	6.55	13.07	20.01	25.13	50.19	61.97	25.59	50.06	63.08
Transductive EpiCA	**9.01**	**17.63**	**24.01**	**7.18**	**14.02**	**21.31**	**53.18**	**68.71**	**77.89**	**35.04**	54.83	**70.02**

Table 3.3 AUC in percentage (multiplied by 100) on MIT-States and UT-Zappos. Our EpiCA model outperforms the previous methods by a large margin on MIT-States based on most of the metrics on UT-Zappos.

EpiCA outperforms all baselines on Zappos benchmark and exceeds the state-of-the-art by 3.3%. It achieves comparable performance on MITStates benchmark. We will empirically analyze the model's behavior in later sections.

3.4.2 Generalized ZSCL Setting

In this setting, following the related work [Purushwalkam et al., 2019a], we measure the performance with AUC metric. AUC introduces the concept of calibration bias which is a scalar value added to the predicting scores of unseen pairs. By changing the values of the calibration bias, we can draw an accuracy curve for seen/unseen sets. The area below the curve is the AUC metric as a measurement for the generalized ZSCL system.

Quantitative results. Tab. 3.3 provides comparisons between our EpiCA model and the previous methods on both the validation and testing sets. As Tab. 3.3 shows, the EpiCA model outperforms the previous methods by a large margin. On the challenging MIT-States dataset which has about 2000 attribute-object pairs, all the baseline methods have a relatively low AUC score while our model is able to double the performance of the previous methods, indicating its effectiveness.

3.4.3 Ablation Study

We conduct ablation study on EpiCA and compare its performance in different settings.

Importance of Transductive Learning.

The experimental results in Tab. 4.3 and Tab. 3.3 show the importance of transductive learning.

There are about 2% and 3% performance gains for MIT-States and UT-Zappos in conventional ZSCL. A significant improvement is observed for both datasets in generalized ZSCL. This is within our expectation because 1) our inductive model has accumulated knowledge about the elements of the concept and has the ability to pick confident test examples. 2) after training the model with the confident pseudo-labeled test data, it acquires the knowledge about unseen concepts.

Importance of Cross-Attention (CA) Encoder. To analyze the effect of CA encoder, we remove CA (w/o CA) and use unimodal representations for both concepts and images. From Tab. 3.4, it can be seen that EpiCA does depend on multi-modal information to do concept recognition and the results also verifies the rationale to fuse multi-modal information by cross-attention mechanism.

Importance of Gated Pooling (GP) Layer. We replace GP layer by average pooling (w/o GP). Tab. 3.4 shows the effectiveness of GP in filtering out noisy information. Instead of treating each element equally, GP help selectively suppress and highlight salient elements within each modality.

Importance of Episode Training. We also conduct experiments by removing both CA and GP (w/o GP and CA). In this setting, we concatenate unimodal representation of images and concepts and use 2-layer MLP to calculate the relevance score. Although simple, it still achieves satisfactory results, showing episode training is vital for our EpiCA model.

EpiCA variants	MIT-States(%)	UT-Zappos(%)
Full EpiCA	15.79	52.56
- w/o cross attention (CA)	12.05	42.77
- w/o gated pooling (GP)	13.46	50.47
- w/o GP and CA	14.13	48.76

Table 3.4 Ablation study of EpiCA components. The episode training and cross-attention encoder are import to our model. Adding gated pooling layer further boosts the accuracy.

3.4.4 Qualitative Analysis.

Fig. 3.3 shows some examples and their predicted labels by EpiCA. Although it gives the correct predictions for the two examples in the first row, EpiCA still struggles in distinguishing the similar, even opposite attributes, like New and Old. For example, the second highest prediction for the image with true label new truck is old car. The predicted object is reasonable, but the predicted

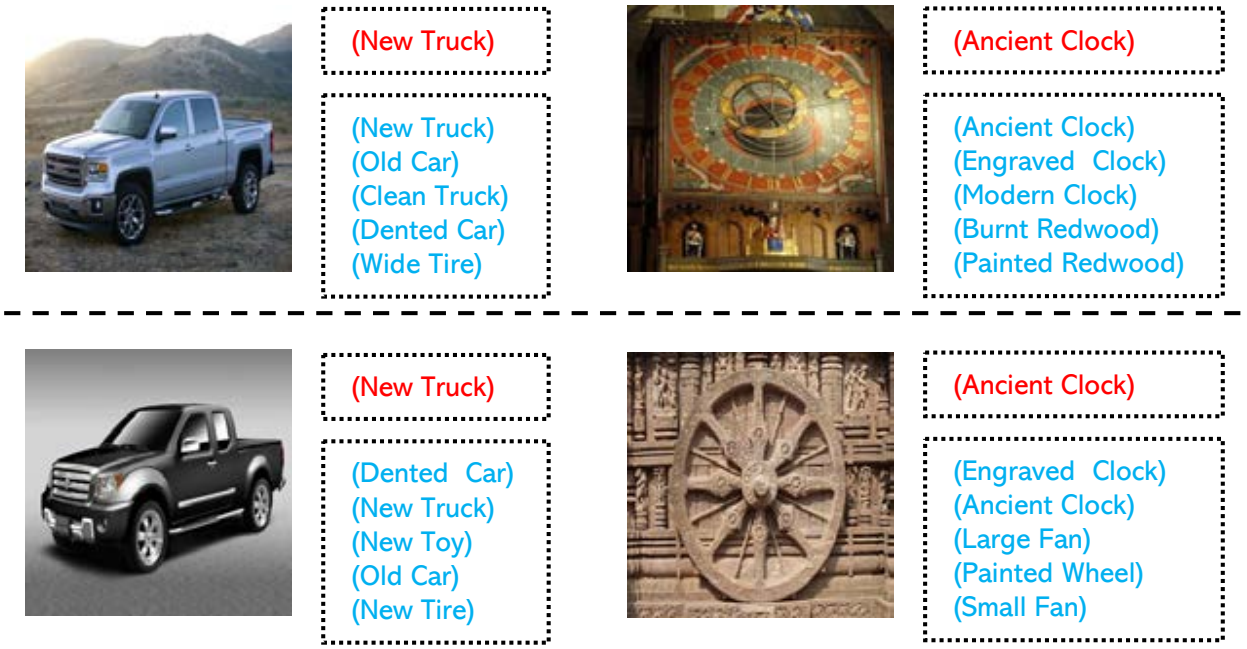


Figure 3.3 Predicting examples of EpiCA from MIT-States dataset. True label and predicted labels are in red and blue text respectively.

attribute is opposite. Meanwhile, for the incorrect predictions, the predicted labels are meaningful and remain relevant to the image. For example, Engraved Clock may be a better label than Ancient Clock for the bottom image. These examples show that EpiCA learns the relevance between images and concepts. But the evaluation of the models is hard and in some cases additional information and bias is needed to predict the exact labels occurring in the dataset.

3.5 Conclusion

In this thesis, we propose EpiCA which combines episode-based training and cross-attention mechanism to exploit the alignment between concepts and images to address ZSCL problems. It has led to competitive performance on two benchmark datasets. In generalized ZSCL setting, EpiCA achieves over 2× performance gain compared to the SOTA on several evaluation metrics. However, ZSCL remains a challenging problem. Future work that explores cognitively motivated learning models and incorporates information about relations between objects as well as attributes will be interesting directions to pursue.

CHAPTER 4

GIPCOL: GRAPH-INJECTED SOFT PROMPTING FOR COMPOSITIONAL ZERO-SHOT LEARNING

Pre-trained vision-language models (VLMs) have achieved promising success in many fields, especially with prompt learning paradigm. In this work, we propose GIPCOL (**G**raph-**I**njected Soft **P**rompting for **CO**mpositional **L**earning) to better explore the compositional zero-shot learning (CZSL) ability of VLMs within the prompt-based learning framework. The soft prompt in GIPCOL is structured and consists of the prefix learnable vectors, attribute label and object label. In addition, the attribute and object labels in the soft prompt are designated as nodes in a compositional graph. The compositional graph is constructed based on the compositional structure of the objects and attributes extracted from the training data and consequently feeds the updated concept representation into the soft prompt to capture this compositional structure for a better prompting for CZSL. With the new prompting strategy, GIPCOL achieves state-of-the-art AUC results on all three CZSL benchmarks, including MIT-States, UT-Zappos, and C-GQA datasets in both closed and open settings compared to previous non-CLIP as well as CLIP-based methods. We analyze when and why GIPCOL operates well given the CLIP backbone and its training data limitations, and our findings shed light on designing more effective prompts for CZSL[1].

4.1 Introduction

Compositional ability is a key component of human intelligence and should be an important building block for current autonomous AI agents. Fig. 5.1 demonstrates a compositional learning example where after learning the element concepts sliced and apple, the autonomous agent is expected to recognize the novel composition sliced apple, by composing the leared element concepts[2] which has not been observed during the training time. This example shows the compositional attribute-object learning problem and this type of compositional ability is essential for language grounding in the vision-language tasks, such as instruction following [Chai et al., 2018], navigation

[1]GIPCOL: Graph-Injected Soft Prompting for Compositional Zero-Shot Learning. Guangyue Xu, Joyce Chai, Parisa Kordjamshid. WACV, 2024

[2]Element concepts also known as primitive concepts including both attributes and objects in CZSL

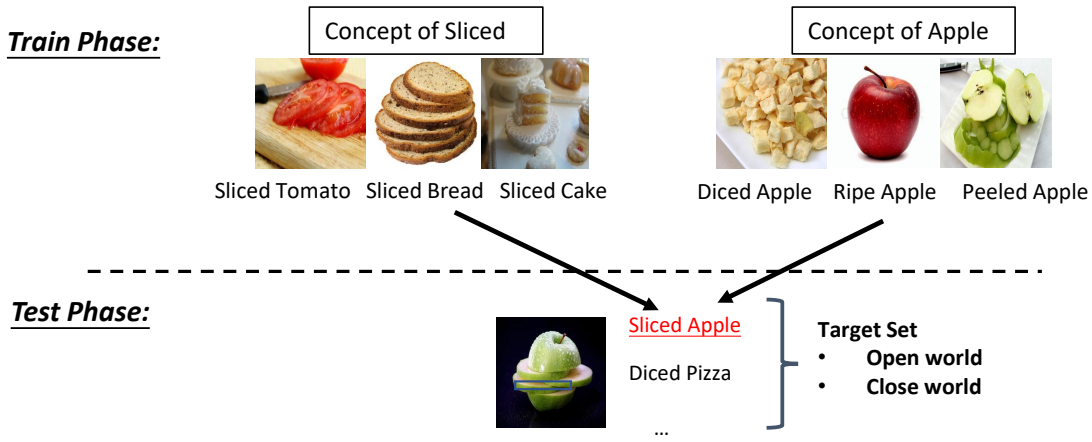


Figure 4.1 CZSL setting: given the element concepts of sliced and apple, our target is to recognize the compositional concept sliced apple.

[Anderson et al., 2018] , and image captioning [Vinyals et al., 2015].

In this chapter, we investigate the compositional zero-shot learning (CZSL) problem as shown in the example. It requires agents to recognize novel compositions of the attribute-object (attr-obj) pairs appearing in an image by composing previously learned element concepts (e.g., "sliced" and "apple" individually are considered as element concepts). The main challenges of CZSL are 1) zero-shot setting in which we do not have training data for the novel compositions. 2) the model should learn the compositional rules to compose the learned element concepts. 3) the distribution shift from the training data to the test data cased by zero-shot setting. Such shift causes the learned models overfitting the seen compositions and makes it difficult to generalize to novel compositions. Previous solutions usually construct a shared embedding space to calculate the matching scores between images and seen pairs and add different generalizing constraints to regularize the space expecting the learnt embeddings capable of encoding compositional properties [Nagarajan and Grauman, 2018b, Naeem et al., 2021, Mancini et al., 2021]. Given impressive performance of large VLMs on downstream tasks, in this work, we attempt to solve CZSL from the lens of prompting large VLMs specifically using CLIP [Radford et al., 2021] as in [Nayak et al., 2022].

Different from traditional zero-shot learning (ZSL) settings where each class is represented by a single text label [Zhou et al., 2022a, Zhou et al., 2022b], CZSL needs to consider the compositional

information among the concepts. Therefore, the prompt design which can efficiently encode the compositional information is the main challenge for our work. We expect the designed prompt can re-program CLIP for compositional learning [Tsai et al., 2020] and the compositional labels in the prompt should consider the compositonal information. Motivated by above expectations, we propose GIPCOL (Graph-Injected Soft Prompting for COmpositional Learning) to design a better prompt to apply VMLs in CZSL. The core idea of GIPCOL is to re-program CLIP for CZSL by setting the prefix vectors in the soft prompt as learnable parameters which is different from CSP [Nayak et al., 2022]. Moreover, GIPCOL captures the compositional structure between concepts by constructing a compositional graph from the seen pairs in the training dataset. The concepts, both element concept and compositional concept, are acting as nodes in the graph and the compositional graph models the feasible topological combinations between these concepts. GIPCOL uses a GNN module to update the element label's representations based on their neighbor information in the constructed compositional graph. And the updated element embedding is used as class labels int the soft prompt. Concretely, the learnable prefix vectors and GNN-updated element concepts consist of the soft prompt for GIPCOL and work together to explore CLIP's knowledge for CZSL. The contributions of this work can be summarized as follows,

- Novel prompting design. Our technique introduces a novel way of utilizing the compositional structure of concepts for constructing the soft prompts. Though we use GNN for capturing this structure, any other differentiable architectures can be used here to enrich the prompt's compositional representation.
- GIPCOL achieves SoTA AUC results on all three CZSL benchmarks, including MIT-States, UT-Zappos, and the more challenging C-GQA datasets. Moreover, it shows consistent improvements compared to other CLIP-based methods on all benchmarks.

4.2 Related Work

Compositional Zero-Shot Learning (CZSL) is a special field of Zero-Shot Learning (ZSL). The CZSL is a challenging problem as it requires generalization from seen compositions to novel compositions by learning the compositional rules between element concepts. There are mainly four

lines of research to address this problem. 1) Classifier-based methods train classifiers for attributes and objects separately and combine the element predictions for compositional predictions [Misra et al., 2017a]. 2) Embedding-based methods construct a shared embedding space for both textual pairs and images. Different methods add different constraints on the space to enhance compositionality [Nagarajan and Grauman, 2018b]. 3) Generation-based methods learn to generate visual features for the novel compositions and train classifiers from the generated images [Xian et al., 2018a]. 4) Newly proposed prompt-based methods utilize CLIP and introduce learnable element concept embedding or soft prefix vectors in the soft prompt to solve CZSL problems [Nayak et al., 2022, Xu et al., 2022].

Prompt-based Learning. Parallel to 'fine-tuning', prompt learning provides an efficient mechanism to adapt large pretrained language models(PLMs) or vision-language models (VLMs) to downstream tasks by treating the input prompt as learnable parameters while freezing the rest of the foundation model. Prompt learning is a parameter-efficient framework originated from the NLP field aiming at utilizing knowledge encoded in PLMs for downstream tasks [Liu et al., 2021, Brown et al., 2020, **?**]. Recently, as the prevalence of large vision-language models (VLMs), prompt learning is introduced into multimodal settings to solve VL-related problems [Tsimpoukelli et al., 2021, Yang et al., 2022, Jin et al., 2021], including the CZSL problems [Nayak et al., 2022, Xu et al., 2022]. In both linguistic and multi-modal settings, prompt engineering plays an important role. How to design a suitable prompt template for downstream tasks is a challenge and GIPCOL proposes a novel approach to address this challenge.

Vision-Language Models. Large VMLs are pre-trained to learn the semantic alignment between vision and language modalities in different levels [Jia et al., 2021, Radford et al., 2021]. Attention-based encoder, large mini-batch contrastive loss, and web-scaled training data are the main factors to boost the performance of such vision-language models. Recent advances in these pre-trained VLMs have presented a promising direction to promote open-world visual understanding with the help of language. Besides the open-world image classification, VLMs are used in other visual fields, like dense prediction [Rao et al., 2022b] and caption generation [Mokady et al., 2021].

Among existing methods, the most relevant to ours are CSP [Nayak et al., 2022] and CGE[Naeem et al., 2021]. CSP treat the element concept labels as learnable parameters to prompt CLIP for CZSL and can be considered as a baseline for GIPCOL. CGE encodes compositional concepts using GNN and constructs a shared embedding space to align images and compositional concepts. It is a task-specific architecture and needs to fine-tune the visual encoder to achieve satisfactory performance. Compared with such task-specific models, GIPCOL is a general prompting method and uses GNN to capture interactions among the concepts for its soft prompting design. GIPCOL fixes CLIP's pre-trained visual and textual encoders and achieves better performance in a more general and parameter-efficient manner. It is worth noting that GNN used in CGE and GIPCOL have different nature, CGE for compositional encoding and GIPCOL for soft prompt construction.

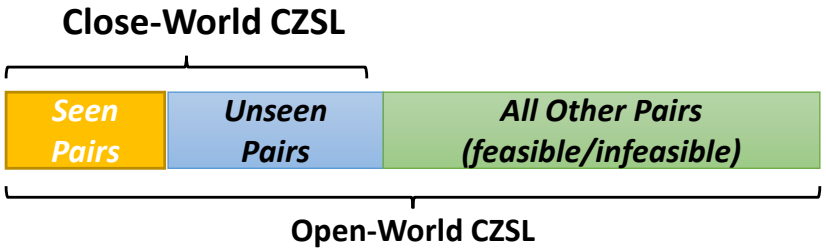


Figure 4.2 Illustration of different CZSL settings based on the target compositional set. GIPCOL is evaluated under closed-world and open-world settings.

4.3 Problem Formulation

In this section, we formally define the CZSL task. Let $\mathbb{A} = \{a_0, a_1, \ldots, a_n\}$ be the attribute set and $\mathbb{O} = \{o_0, o_1, \ldots, o_m\}$ be the object set. All possible compositional label space $\mathbb{C}$ is the Cartesian product of these two element concept sets, $\mathbb{C} = \mathbb{A} \times \mathbb{O}$ with size $n \times m$. At training time, we are given a set of seen[3] examples $\mathbb{C}_{\text{seen}} = \{(x_1, c_1), \ldots, (x_k, c_k)\}$, where x_i is an image and $c_i = (a_i, o_i)$[4] is its compositional label from the seen set $\mathbb{C}_{seen} \subset \mathbb{C}$. The goal of CZSL is to learn a function f to assign a compositional label from the target set $\mathbb{C}_{target} \subseteq \mathbb{C}$ to a given image . Based on different target set settings as shown in Fig. 4.2, CZSL can be categorized into 1)

[3]seen examples also mean training examples, we use them interchangeably in this work.

[4]We use the pair index to denote the object and attribute indexes for the sake of simple notation. The object and attribute indexes do not refer to their original sets in this case.

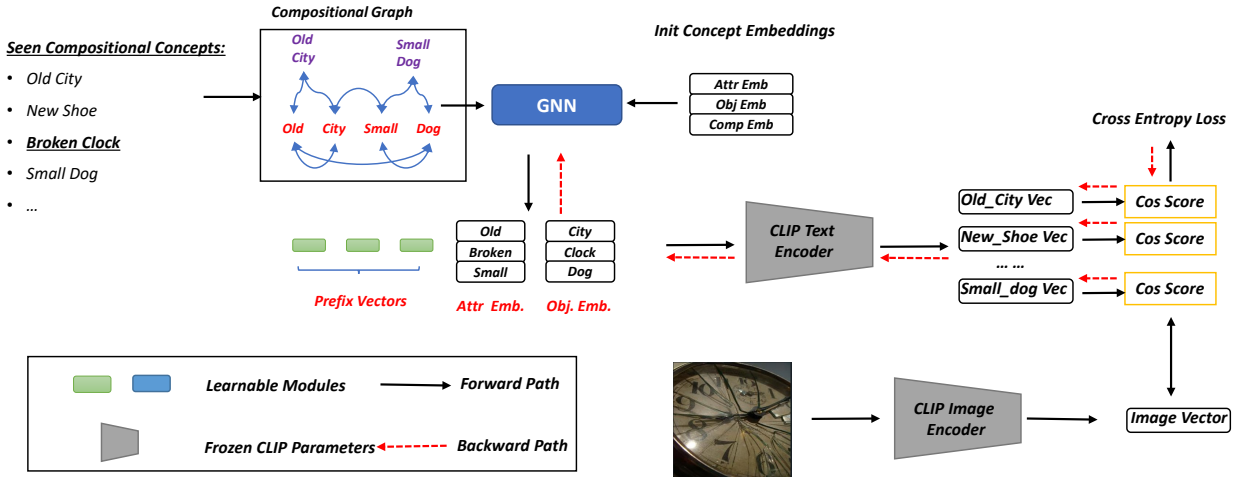


Figure 4.3 GIPCOL Architecture. Besides CLIP's frozen text and visual encoders, GIPCOL consists of two learnable components: a soft-prompting module and a GNN. GIPCOL calculates the cosine similarity between the given image and all candidate pairs and the cross-entropy loss is back-propagated through the frozen LM in order to update soft-prompt and GNN.

Closed-world CZSL, where $\mathbb{C}_{target} = \mathbb{C}_{seen} \cup \mathbb{C}_{unseen}$, the target set consists of both seen and unseen pairs as introduced in [Purushwalkam et al., 2019b]. In this setting, both seen and unseen pairs are feasible. This setting is called a closed-world setting because the test pairs are given in advance. 2) Open-world CZSL, where $\mathbb{C}_{target} = \mathbb{C}$. The target set contains all attr-obj combinations including both feasible and infeasible pairs. This is the most challenging case introduced in [Mancini et al., 2021]. We evaluate our models under both closed-world and open-world settings.

4.4 GIPCOL

By pre-training on 400 million image-text association pairs, CLIP has already learned the general knowledge for images recognition. In order to fully utilize CLIP's capability in compositional learning, GIPCOL freezes CLIP's textual and visual encoders and focuses on structuring its textual prompt to address compositional concept learning. The GIPCOL's architecture is shown in Fig. 5.4. In particular, GIPCOL adds two learnable components to construct the soft prompt for CZSL: the learnable prefix vectors and the GNN module. The prefix vectors are used to add more learnable parameters to represent the compositional concepts and reprogram CLIP for compositional learning. Notably, in the whole architecture in Fig 5.4, soft-prompt and soft-embedding are the only modules that need to be learnt. The GNN module is to capture the compositional structure of the objects

and attributes for a better compositional concept representation in the constructed soft prompt. We describe the details of GIPCOL, including the learnable prefix vectors, GNN, and CLIP's visual/textual encoder in the following section.

4.4.1 GIPCOL Architecture

Learnable Prefix Vectors. We designate k learnable prefix vectors $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$ where $\theta_i \in \mathbb{R}^d$ in soft prompt for compositional concept encoding. d is set to 768 to be consistent with CLIP embedding size. Here, larger k means more learnable parameters and learning ability for compositional concept representation. These vectors are used to prepend to the attr-obj embeddings and act as part of the compositional representation. These prefix vectors are fine-tuned by gradients flowing back through CLIP during the training time.

GNN as Concept Encoder. Different from traditional zero-shot learning (ZSL) problems where output labels are treated independently, CZSL requires modeling the interactions between element concepts. For example, given the compositional concept red apple, we need to learn both the concept apple and how red changes apple's state instead of treating red and apple as two independent concepts. Graph Neural Networks (GNN) have been proved to be able to capture such dependencies [Naeem et al., 2021, Mancini et al., 2022]. We introduce GNN in GIPCOL to enrich the concept's representations by fusing information from their compositional neighbors as follows,

$$(\hat{a}_i, \hat{o}_i) = GNN_\Phi(a_i, o_i) \tag{4.1}$$

where Φ is GNN's parameter, (a_i, o_i) and $(\hat{a}_i, \hat{o}_i)$ are the original and updated compositional concept's representation. The updated node representations from GNN will serve as class labels in soft prompt. The whole soft prompt represents the compositional concept and will be put into CLIP's textual encoder for compositional learning.

Frozen CLIP's Text Encoder. After obtaining the updated compositional representations $(\hat{a}_i, \hat{o}_i)$, GIPCOL adds the learnable prefix vectors $\Theta = [\theta_1, \theta_2, ..., \theta_m]$ prepending in front of $(\hat{a}_i, \hat{o}_i)$ to represent compositional concept as follows,

$$\underbrace{[SOS,\ \overbrace{\theta_1, \theta_2, ..., \theta_m}^{\textbf{prefix Vectors}},\ \overbrace{\hat{a}_i, \hat{o}_i}^{\textbf{GNN-Updated Concept}},\ EOS]}_{\textbf{Soft Prompt as Compositional Concept Representation}}. \tag{4.2}$$

Then we use CLIP's frozen text encoder, a Bert encoder [Devlin et al., 2019], to extract the normalized EOS vector as the compositional concept's representation for further multi-modal alignment as follows,

$$\boldsymbol{c}_i = \frac{TxtEnc\left(\Theta, (\hat{a}_i, \hat{o}_i)\right)}{\|TxtEnc\left(\Theta, (\hat{a}_i, \hat{o}_i)\right)\|} \tag{4.3}$$

where $(\hat{a}_i, \hat{o}_i)$ is the GNN-updated attribute and object vectors and c_i is the i-th compositional concept vector encoded by CLIP.

Frozen visual encoder. Following CLIP's pre-processing routine, we first rescale the image's size to 224×224. Then we use ViT-L/14 as the visual encoder ViT to encode the image and extract the [CLASS] token as the image's representation. The extracted image vector x_i needs to be normalized as follows for further similarity calculation.

$$\boldsymbol{x}_i = \frac{VisEnc(v_i)}{\|VisEnc(v_i)\|} \tag{4.4}$$

where v_i is the given image and x_i is its vector representation.

Aligning Image and Compositional Concept. After obtaining the vectors for the compositional concept c_i and the image x, GIPCOL calcualtes the probability of x belonging to class c_i as follows,

$$p(c_i \mid x) = \frac{\exp\left((x \cdot c_i)/\tau\right)}{\sum_{k=1}^{K} \exp\left((x \cdot c_k)/\tau\right).} \tag{4.5}$$

where τ is a temperature parameter from CLIP, $\cdot$ denotes the inner product of the concept vector and the image vector and K is the number of attr-obj pairs in the training set.

4.4.2 GNN in Soft Prompting

As disussed previously, a key idea to address the CZSL problem is to learn concept representations that are able to internalize the compositional information. Graph could be the tool to model such compositional dependencies. And this idea has been used in previous work [Naeem et al.,

2021, Mancini et al., 2022] by applying Graph Neural Networks(GNN) as encoders to represent the compositional concepts. Although we adopt similar graph-based methods for compositional encoding, our novelty is to use the graph's compositional structure to facilitate the automated prompt engineering in compositional learning as shown in Fig. 4.4. We model the element concepts and their composition explicitly in GNN for the soft prompting construction. In principle, the GNN module can be replaced by other differentiable architectures that are able to capture the concept's compositional information. We describe the detailed GNN application in GIPCOL next.

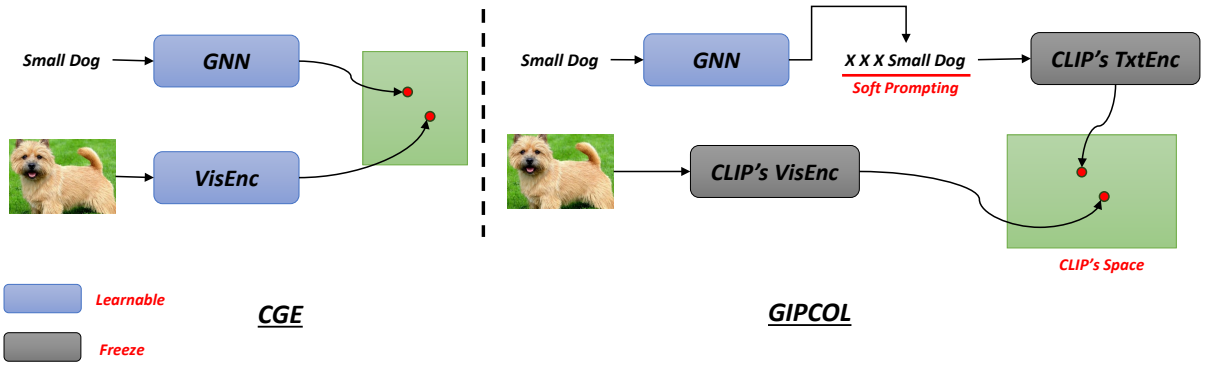


Figure 4.4 Comparison between CGE and GIPCOL. The main difference is that GIPCOL uses GNN to help prompt construction instead of compositional concept encoder.

Node Embedding $\mathcal{V}$. There are two types of nodes in GIPCOL's compositional graph: element concept node and compositional concept node. The node embedding's size is $R^{(|a|+|o|+|c|)*d}$, where $|a|$ is the attribute number, $|o|$ is the object number, $|c|$ is the training pair number and d is the feature dimension. For the element nodes, we initialize them using CLIP's embedding vectors. For the compositional nodes, we initialize them using the average embedding of the element nodes, that is, $\frac{att\_vec+obj\_vec}{2}$. GIPCOL relies on GNN to fuse information from the constructed compositional graph and update the concept's representation.

Compositional Graph Constructions $\mathcal{E}$. We use a graph to capture the compositional dependencies and learn richer concept representations. The connection design among concepts is the key challenge for such graph. In order to utilize the feasible compositional information, GIPCOL considers the training pairs and construct one single compositional graph for both closed-world

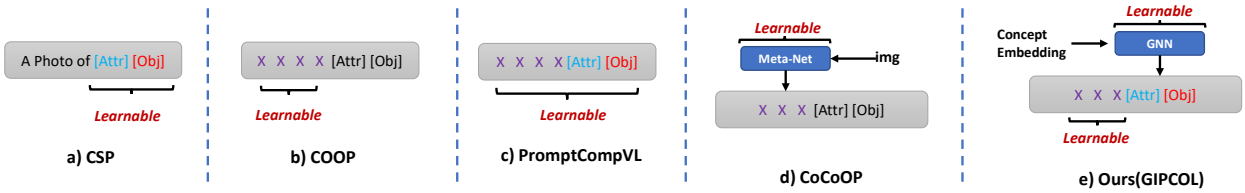


Figure 4.5 Different prompting strategies. GIPCOL combines both soft prefix vector and GNN for prompt construction.

CZSL and open-world CZSL to conserve the computing and storage resources. Specifically, given a pair $c = (a, o)$, besides the self-connected edge, GIPCOL adds three undirected edges $(c \leftrightarrow a)$, $(c \leftrightarrow o)$ and $(a \leftrightarrow o)$ in the graph where the adjacency matrix $A \in \mathbb{R}^{K \times K}$ is symmetric with $K = |a| + |o| + |c|$. The compositional concept plays the bridging role to help connect element concepts and only the element concepts are used to construct the compositional prompting due to the zero-shot setting.

GNN Module: Once we have the compositional graph and the initialized concept features, we can update the concept's embedding by fusing the compositional information from its neighbors. Any GNN models could be applied here and in GIPCOL, we use Graph Convolution Network (GCN) [Kipf and Welling, 2016] in Eq. 4.6 for compositional encoding.

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}\Phi^{(l)}\right) \tag{4.6}$$

where H^l denotes the node's representations in the l^{th} layer, σ is the non-linearity ReLU function, $\tilde{A}$ is the adjacency matrix with added self-connections, $\tilde{D}$ is a diagonal node degree matrix and Φ^l is the learnable weight matrix in layer l. Notably, other graph constructing methods, like using external knowledge [Karthik et al., 2022], and other GNN models, like GAT [Velickovic et al., 2017], could be further explored to improve CZSL performance based on GIPCOL's architecture. However, these are not target of this work. Here, GIPCOL shows the effectiveness of utilizing compositional knowledge in prompting construction in CZSL.

4.4.3 Training

After obtaining the concept and image representations, we calculate the class probability using Eq. 4.5. And the regularized Cross-Entropy loss is used to update GIPCOL's prefix vectors Θ and

GNN parameters Φ as follows,

$$-\frac{1}{|\mathbb{N}|}\sum_{i\in\mathbb{N}}\log p_{\theta}(c_i \mid x) + \lambda_1\|\mathbf{\Theta}\|^2 + \lambda_2\|\mathbf{\Phi}\|^2 \tag{4.7}$$

where λ_1 and λ_2 are the hyper-parameters to control the weight decay for prefix vector and GCN separately. GIPCOL keeps CLIP's pre-trained textural and visual encoders fixed during the training time. And more details about the training process can be found in Alg. 2.

Algorithm 2: GIPCOL:

1: Initialize GIPCOL using CLIP's pre-trained textual and visual encoders.
2: Update element concept's representation using GNN as Equation 4.1 and Equation 4.6.
3: Construct textual prompt for compositional labels using the updated element concepts and learnable prefix vectors as Equation 4.2.
4: Extract and normalize image/text vectors using CLIP's image/text encoder based on Equation 4.3 and Equation 4.4 separately.
5: Calculate the class probability as Equation 4.5 using the cosine similarity and update GIPCOL's soft-prompting layer Θ and GNN layer Φ using Cross-Entropy loss.

4.4.4 Inference

During inference, given an image, we first construct the soft prompts for all target concepts using the fine-tuned prefix vectors and GNN. Then, we use CLIP's frozen textual and visual encoders to obtain the image vector x and the target concept vector set $\mathbb{C}_{tareget}$. Then we use cosine measurement to select the most similar attr-obj pair from $\mathbb{C}_{target}$ as the compositional label as follows,

$$\hat{c} = \underset{c_i\in\mathbb{C}_{\text{target}}}{\arg\max}\ cos\left(c_i, x\right). \tag{4.8}$$

where c_i is the i-th compositional vector from the target set.

4.4.5 CLIP-Prompting Method Comparison

In this section, we clarify the difference between all CLIP-prompting methods used in CZSL as shown in Fig. 4.5. Generally, all current CLIP-prompting methods keeps the image representation fixed and learn constructing the CLIP's textual prompt to represent the compositional concept as shown in Eq. 4.2. The main difference is that CSP[Nayak et al., 2022] learns the element

embedding, COOP[Zhou et al., 2022a] learns the prefix vectors and PromptCompVL learns both the element embedding and the prefix vectors. All these three methods do not explicit consider the compositional structures between concepts. In order to inject more semantic information into soft prompt, CoCoOP[Zhou et al., 2022b] introduces a Meta-Net and tries to modify the prefix vectors based on each image input. It uses the instance-level information not the global compositional information for CZSL. Such instance-level prompting also causes training inefficient and consumes a significant amount of computing resources as discussed in that work. Different from all previous methods, GIPCOL proposed a novel prompting strategy by combining the learnable prefix vectors and the GNN module and the detailed comparition is in Append. **??**. Although both CGE[Naeem et al., 2021] and GIPCOL use GNN to encode compositional concepts, the GNN module functions in a fundamentally different manner in these two models. GNN in GIPCOL helps construct the soft prompting for CZSL. However, GNN in CGE plays the text encoder role which projects the concept into the embedding space. GIPCOL freeze CLIP's textual and visual encoders to utilize CLIP's multi-modal aligning ability for CZSL which is more efficient. In contrast, CGE needs to train both the GNN and visual encoder to obtian competitive performance as compared in Fig. 4.4.

4.5 Experiments

4.5.1 Experimental Setting

Datasets. We conduct experiments on three datasets, MIT-States [Isola et al., 2015a], UT-Zappos [Yu and Grauman, 2014] and C-GQA [Naeem et al., 2021]. MIT-States and C-GQA consist of images with objects and their attributes in the general domain. In contrast, UT-Zappos contains images of shoes paired with their material attributes which is a more domain-specific dataset. Our experiments follow the previous works [Purushwalkam et al., 2019b, Naeem et al., 2021] on the data split for training and testing. More details about the data splits and statistics can be found in Tab. 4.1.

Implementation details. We extend on the codebase of [Nayak et al., 2022][5] and [Naeem et al., 2021][6] for GIPCOL's implementation. Moreover, for a fair comparison, the length of the prefix

[5]https://github.com/BatsResearch/csp
[6]https://github.com/ExplainableML/czsl

	MIT-States	UT-Zappos	C-GQA
$\#\,Attr.$	115	16	413
$\#\,Obj.$	245	12	674
$\#\,Attr. \times Obj.$	28175	192	278362
# Train Pair	1262	83	5592
# Train Img.	30338	22998	26920
# Val. Seen Pair	300	15	1252
# Val. Unseen Pair	300	15	1040
# Val. Img.	10420	3214	7280
# Test Seen Pair	400	18	888
# Test Unseen Pair	400	18	923
# Test Img.	19191	2914	5098

Table 4.1 Dataset Statistics for MIT-States, UT-Zappos and C-GQA.

vector, k, is set to 3 which is the same length of CLIP hard-prompting 'a photo of'. The dimension of soft-prompting d is set to 768 which is consistent with CLIP's model setting. Moreover, we use two-layer GCN to encode concepts and the corresponding GNN's learnable parameters are $\Phi = \{\Phi^1, \Phi^2\}$ Our code will be made publicly available on GitHub[7].

Evaluation Metrics. Zero-shot models are biased to the seen classes as shown in previous woks [Chao et al., 2016, Mancini et al., 2021]. As a standard method in zero-shot learning, we introduce a scalar value adding to the unseen classes to adjust the bias towards the seen classes as used in [Purushwalkam et al., 2019b, Nayak et al., 2022]. By varying the added bias from $-\infty$ to $+\infty$, we report GIPCOL's performance using the following four metrics in both the closed-world and the open-world settings as discussed in Sec. 4.3: 1) Best seen accuracy (S), testing only on seen compositions when bias is $-\infty$; 2) Best unseen accuracy (U), testing only on unseen compositions when bias is $+\infty$; 3) Best harmonic mean (HM) which balances the performance between seen and unseen accuracies; 4) Area Under the Curve (AUC), the area below the seen-unseen accuracy curve by varying the scalar added to the unseen compositional concepts.

Baselines. We compare GIPCOL with two types of baselines: 1) non-CLIP methods (top seven models in the closed setting and top six in the open setting) namely Attributes as Operators (AoP)[Nagarajan and Grauman, 2018b], Label Embed+ (LE+)[Misra et al., 2017a], Task Mod-

[7]https://github.com/HLR/GIPCOL

ular Networks (TMN)[Purushwalkam et al., 2019b], SymNet[Li et al., 2020b], Compositional Graph Embeddings (CGE)[Naeem et al., 2021], Compositional Cosine Logits (CompCos)[Mancini et al., 2021] and Siamese Contrastive Embedding Network(SCEN)[Li et al., 2022]. 2) CLIP-based methods (the bottom three models), namely CLIP[Radford et al., 2021], Context Optimization(COOP)[Zhou et al., 2022a] and compositional soft prompting (CSP)[Nayak et al., 2022].

Feasibility Calibration in Open-World Setting. Open-world CZSL is more challenging compared with the closed-world setting as the class space contains all possible combinations of attributes and objects including both feasible compositions and infeasible compositions. In order to filter out the infeasible compositions, we apply the feasibility calibration as used in [Mancini et al., 2021, Nayak et al., 2022]. For each unseen pair (a, o), we first collect two sets from the training data. One is the applicable attribute set $A = \{a_1, a_2, \ldots, a_M\}$ for the target object o and the other is the applicable object set $O = \{o_1, o_2, \ldots, o_N\}$ for the target attribute a where (a_i, o) and (a, o_j) has been observed in training time. Then we calculate the similarity between a and each element in A and use the maximum similarity score as this pair's attribute feasibility score as follows,

$$f_a(a, o) = \max_{(a_i, o) \in \mathbb{C}_{\text{seen}}} \frac{e(a) \cdot e(a_i)}{\|e(a)\| \|e(a_i)\|}, \tag{4.9}$$

where e is the GloVe embedding [Pennington et al., 2014b]. On the other hand, this pair's object feasibility score is calculated in a similar way based on the applicable object set. Finally, the unseen pair feasibility score is calculated as the average of the two scores, $\frac{f_a + f_o}{2}$. After obtaining the feasibility score for all unseen pairs, we can filter out infeasible compositions by setting a threshold T whcih can be tuned based on the validation set. The final prediction for image x in the open-world setting is computed as follows,

$$\hat{c} = \underset{c_i \in \mathbb{C}_{\text{target}},\, c_i \geq T}{\arg\max} \; cos\left(c_i, x\right). \tag{4.10}$$

Different from the closed-world setting, we require the feasibility score of the predicted label c to be larger than a threshold. The threshold uses in our experiments is shown in Tab. 4.2. In open-world CZSL (OW-CZSL), we use the validation set to choose a feasible threshold to remove less feasible compositions from the output space and the adopted threshold in GIPCOL is shown in Tab. 4.2.

Dataset	Feasibility Score
MIT-States	0.40691
UT-Zappos	0.51878
C-GQA	0.49941

Table 4.2 GIPCOL's feasibility threshold score.

4.5.2 Results

Results on MIT-States. As shown in Tab. 4.3 and 4.4, GIPCOL achieves the new SoTA results on MIT-States on both closed-world and open-world settings compared with CLIP and non-CLIP baselines (except for the best-unseen metric (U)). The CLIP-based models have consistently better performance compared to the non-CLIP methods[8]. CLIP-prompting methods, including COOP, CSP and ours, further boost the performance compared to the vanilla CLIP model.

Results on UT-Zappos. On UT-Zappos, previous CLIP-based approaches under-perform the SoTA performance achieved by CGE which is a non-CLIP model. However, GIPCOL successfully surpasses the CGE model. Note that UT-Zappos is a domain-specific dataset that consists of shoe types and the materials. There may exist two reasons for to explain the accurary drop: 1) CLIP doesn't see many images from this domain during training time; 2) As a fashion data, there is a appearance shift between CLIP's training data set and UT-Zappo's test data set We suspect that CLIP may not have seen sufficient similar samples from this specific domain and therefore purely tuning the prompting is not helpful to solve the problem. In contrast, GIPCOL adds additional compositional information to learn the element concept embedding which appears to boost the compositional learning ability within this specific domain.

Results on C-GQA. On the more challenging C-GQA dataset, GIPCOL also achieves new SoTA results on both closed and open world settings with an exception for the seen accuracy in the open world. However, the key metric is AUC which is consistently higher for GIPCOL in all settings.

Comparing GIPCOL with other CLIP-based method. Besides the absolute SOTA improvement on MIT-States, another interesting observation is the GIPCOL achieves a consistent improvement

[8]In principle CLIP-based and non-CLIP-based methods cannot be directly compared as we have no information about the training data used for CLIP training. Here we follow previous work and include these baselines for the sake of comparison and consistency with the previous work.

Method	MIT-States				UT_Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP [Nagarajan and Grauman, 2018b]	14.3	17.4	9.9	1.6	59.8	54.2	40.8	25.9	17.0	5.6	5.9	0.7
LE+ [Misra et al., 2017a]	15.0	20.1	10.7	2.0	53.0	61.9	41.0	25.7	18.1	5.6	6.1	0.8
TMN [Purushwalkam et al., 2019b]	20.2	20.1	13.0	2.9	58.7	60.0	45.0	29.3	23.1	6.5	7.5	1.1
SymNet [Li et al., 2020b]	24.2	25.2	16.1	3.0	49.8	57.4	40.4	23.4	26.8	10.3	11.0	2.1
CompCos [Mancini et al., 2021]	25.3	24.6	16.4	4.5	59.8	62.5	43.1	28.7	28.1	11.2	12.4	2.6
CGE [Naeem et al., 2021]	32.8	28.0	21.4	6.5	64.5	71.5	60.5	33.5	33.5	15.5	16.0	4.2
SCEN [Li et al., 2022]	29.9	25.2	18.4	5.3	63.5	63.1	47.8	32.0	28.9	25.4	17.5	5.5
CLIP [Radford et al., 2021]	30.2	40.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
COOP [Zhou et al., 2022a]	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 [Nayak et al., 2022]	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
GIPCOL (Ours)	48.5	49.6	36.6	19.9	65.0	68.5	48.8	36.2	31.92	28.4	22.5	7.14

Table 4.3 Closed-World CZSL results on UT-Zappos, Mit-States and C-GQA datasets.

on MIT-States in both settings and on UT-Zappos in the close setting compared with other CLIP-based methods. This empirically shows the effectiveness of introducing both soft-embedding and soft-prompting in CZSL. Comparing with CSP [Nayak et al., 2022], we only introduce additional 3 learnable prompt vectors and obtain satisfactory improvements on MIT-States. This shows the importance of soft-prompting. It reprograms CLIP for CZSL. For soft-embedding, we learn the element concept embedding instead of using fixed CLIP's embedding which is better for compositional learning compared with COOP [Zhou et al., 2022a].

We give some qualitative analysis in next section. CZSL dataset usually is a multiple-label dataset which means we can describe an object from different dimensions. For example, giraffe in C-GQA, we can describe its color or it size. Both of the compostions should be right. Therefore, we need to develop more reasonable metrics to evaluate compositional learning performance. Moreover, there exists wrong labeled items, such as black point in the last row, meaning we also needs a more clean benchmark for CZSL.

4.5.3 Qualitative Analysis

Predicted Examples. We looked into a number of randomly selected predictions from GIPCOL shown in Fig. 4.6. The red colored texts are the ground-truth labels, the blue colored texts are GIPCOL's correctly predicted labels and the black colored texts are GIPCOL's wrongly predicted labels. The first two columns present examples with correctly predicted compositional labels and the last two columns show the wrongly predicted labels, either wrong in attributes or wrong in objects.

Method	MIT-States				UT_Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP [Nagarajan and Grauman, 2018b]	16.6	5.7	4.7	0.7	50.9	34.2	29.4	13.7	-	-	-	-
LE+ [Misra et al., 2017a]	14.2	2.5	2.7	0.3	60.4	36.5	30.5	16.3	19.2	0.7	1.0	0.08
TMN [Purushwalkam et al., 2019b]	12.6	0.9	1.2	0.1	55.9	18.1	21.7	8.4	-	-	-	-
SymNet [Li et al., 2020b]	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CompCos [Mancini et al., 2021]	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	-	-	-	-
CGE [Naeem et al., 2021]	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.1	1.8	2.9	0.47
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., 2022a]	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., 2022]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
GIPCOL (Ours)	48.5	16.0	17.9	6.3	65.0	45.0	40.1	23.5	31.6	5.5	7.3	1.30

Table 4.4 Open-World CZSL results on UT-Zappos, Mit-States and C-GQA datasets.

From this figure, we can see that GIPCOL can recognize objects in most of the compositions in MIT-States and C-GQA datasets. However, it has difficulty to precisely predict the attributes for these two datasets. For example, it predicts modern clock instead of ancient clock which is the antonym of the actual attribute. But for UT-Zappos, the more domain-specific dataset, GIPCOL even has difficulty in recognizing the objects.

Figure 4.6 We show the top-3 predictions of our proposed model for some images. Red colors are ground-truth labels, blue colors are correctly predicted labels and black colors are wrongly predicted labels.

Differences in Domains: In this section, we try to explain why GIPCOL works in CZSL by checking the CLIP's training data. From Tables 4.3 and 4.4, we observe that CLIP without any prompt-tuning can achieve better performance compared to non-CLIP models on the MIT-States dataset, but not on the UT-Zappos dataset. We hypothesize that this issue can be related to the

distribution difference between the pre-training data used by CLIP and the data domain of the downstream task. To validate this hypothesis, we further look into some concrete examples from MIT-stats and UT-Zappos. We take burnt boat from MIT-Stats and Faux Fur-Shoes Clogs and Mules from UT-Zappos for comparison as shown in Fig. 4.7. From this figure, we can see that MIT-States have similar visual appearance with CLIP's pre-trained data. However, for UT-Zappos, because of the fashion style change overtime, shoes have significant visual appearance between the pre-training dataset and the target dataset. Results in Tab. 4.3 and Tab. 4.4 have shown the domain similarity plays an important role in prompting-based method. Prompting CLIP without any training can achieve better performance on MIT-State then UT-Zappos. GIPCOL helps address this challenge partially by prompting design based on the results. The CLIP's training data is not publicly available. However, LAION-400M [Schuhmann et al., 2021] used the released CLIP model and obtained the closest 400M image-text pairs[9] from their crawled dataset from Web by reverse engineering. We based our analysis on this constructed LAION-400M subset. By querying LAION-400M with burn boat, we could retrieve about 600 relevant images. By querying with Faux Fur_Shoes Clogs and Mules we can retrieve about 200 relevant images. The first interesting difference is in the quantity of the retrieved relevant images which is significantly lower for the shoe dataset. The second difference is the data quality differences. As can be seen from Fig. 4.7, the retrieved shoes are less similar to the UT-Zappos' shoes when compared to the similarity of the retrieved boats to MIT-Stats boats. We note that UT-Zappos is about shoe fashion and was constructed in 2014 while CLIP is pretrained using recent 2020's images. The change in fashion trends has made the images look different for the same compositional concept. Based on these observations, it is evident that the quantity and quality of CLIP's pre-training data play an important role in its performance.

Covering the Performance Gap. Despite the above-mentioned issues, GIPCOL improves the UT-Zappos dataset. While we found that CLIP's pre-training data is important in its performance in the Zero-shot setting, introducing the additional compositional knowledge in GIPCOL positively

[9]https://rom1504.github.io/clip-retrieval

Figure 4.7 Comparison between retrieved images from Laion400M and UT-Zappos/MIT-States.

impacts CLIP's ability in recognizing the novel compositional concepts. GIPCOL uses GNN to inject compositional information into concept representations which turned out to be helpful. The improvement is important, especially for UT-Zappos which is a special domain with not many shared similar examples with CLIP's training.

t-SNE Comparison between CLIP and GIPCOL The compositiponal concepts learnt by GIPCOL and CLIP are visualized separately in Fig. 4.8. Each figure randomly sample 5 compositional concepts with related images and draw their representation using t-SNE [Van der Maaten and Hinton, 2008]. All prompting-based methods share the same image representation because they freeze CLIP's visual encoder during training. Then compositional encoding is the difference between all these prompting models. From the figures, we can see that GIPCOL's compositional vectors (+) are closer to the related image cluster compared with CLIP's vectors (-) which empirically shows that GIPCOL has better compositional encoding ability.

4.5.4 Ablation Study

To better understand the influence of each component in GIPCOL, Tab. 4.5 shows the performances of its variations on UT-Zappos' closed-world setting. From Table 4.5, both GNN and soft-prompting are important for GIPCOL.

Effects of GNN. We remove the GNN module and directly set attribute and object embeddings as learnable parameters as in [Xu et al., 2022]. The performance decreases. Especially the AUC drops from 36.2% to 32.2%.

Effect of Learnable Prefix Vectors. Another variant of GIPCOL is to fix the prefix vectors and

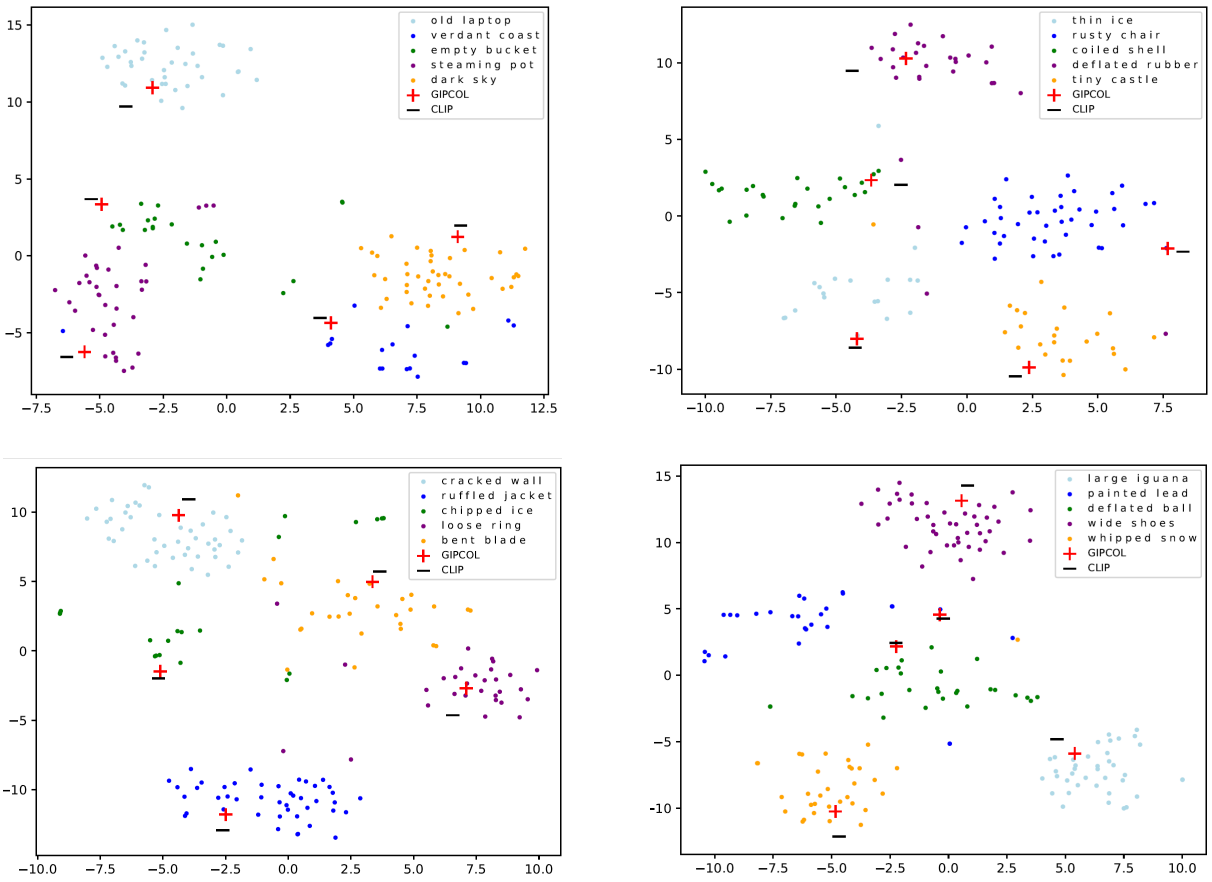


Figure 4.8 t-SNE comparison between CLIP and GIPCOL.

Model	S	U	H	AUC
GIPCOL	65.0	68.5	48.8	36.2
- without GNN	64.4	64.0	46.12	32.2
- without prefix	64.7	62.3	45.9	31.0
- without both (CLIP)	15.8	49.1	15.6	5.0

Table 4.5 Performance of GIPCOL's variations.

only tune the GNN module to update the class embeddings. From Tab. 4.5, we can see that learnable prefix vectors play a more important role than GNN. In fact, adding the prefix vectors changes CLIP's textual input and makes it biased towards compositional learning, which is a key component in GIPCOL.

Comparison to vallina CLIP. Although CLIP has seen many of the compositional concepts during training, applying CLIP directly achieves no satisfactory results in CZSL. This result shows the

importance of prompting learning in CZSL.

4.5.5 Higher-Order Compositional Learning

Previous work (CSP) [Nayak et al., 2022] introduced another challenging dataset: AAO-MIT-States, a subset derived from MIT-States to evaluate the higher-order compositional learning ability in the form of attribute-attribute-object (AAO) compositions. After learning the prefix vectors and GNN-encoded element concepts, GIPCOL can be easily adapted to solve AAO by modifying the compositional prompt to $(\theta_1, \theta_2, ..., \theta_m, \hat{a_i}, \hat{a_j}, \hat{o_k})$ to represent the higer-order compositions. We report the AAO results in Tab. 4.6. We can see that GIPCOL has a better higher-order compositional leaning ability, with a 3% absolute improvement compared with CSP.

Model	Accuracy
CLIP	62.7
CSP	72.6
GIPCOL (Ours)	75.9

Table 4.6 AAO Performance of different CLIP-based models.

4.6 Conclusion

In this chapter, we propose GIPCOL, a new CLIP-based prompting framework, to address the compositional zero-shot learning (CZSL) problem. The goal is to recognize compositional concepts of objects with their states and attributes as described by images. The objects and attributes have been observed during training in some compositions, however, the test-time compositions could be novel and unseen. We introduce a novel prompting strategy for soft prompt construction by treating element concepts as part of a global GNN network that encodes feasible compositional information including objects, attributes and their compositions. In this way, the soft-prompt representation is influenced not only by the pre-trained VLMs but also by all the compositional representations in its neighborhood captured by the compositional graph. Our results have shown that GIPCOL performs better and achieves SoTA AUC results on all three benchmarks including MIT-States, UT-Zappos, and C-GQA . These results demonstrate the advantages and limitations of prompting large vision and language models (such as CLIP) for compositional concept learning.

CHAPTER 5

METAREVISION: META-LEARNING WITH RETRIEVAL FOR VISUALLY GROUNDED COMPOSITIONAL CONCEPT ACQUISITION

Humans have the ability to learn novel compositional concepts by recalling and generalizing primitive concepts acquired from past experiences. Inspired by this observation, in this thesis, we propose MetaReVision, a retrieval-enhanced meta-learning model to address the visually grounded compositional concept learning problem. The proposed MetaReVision consists of a retrieval module and a meta-learning module which are designed to incorporate retrieved primitive concepts as a supporting set to meta-train vision-language models for grounded compositional concept recognition. Through meta-learning from episodes constructed by the retriever, MetaReVision learns a generic compositional representation that can be fast updated to recognize novel compositional concepts. We create CompCOCO and CompFlickr to benchmark the grounded compositional concept learning. Our experimental results show that MetaReVision outperforms other competitive baselines and the retrieval module plays an important role in this compositional learning process [1].

5.1 Introduction

Learning to compose from previous experience is an important integral part of human intelligence [Fodor and Pylyshyn, 1988b, Biederman and Vessel, 2006]. Generally, compositional learning refers to the ability to learn a set of basic primitives and generalize these primitives in a novel scenario different from training time [Kemp and Tenenbaum, 2009, Ontanón et al., 2021]. It includes various learning aspects, such as systematic generalization, productivity and substitutivity [Hupkes et al., 2020]. In this work, we focus on systematic generalization within the multi-modal setting and propose a multi-modal compositional problem: Grounded Compositional Concept Learning (GCCL). As shown in Figure 5.1, in the GCCL setting, the models are trained with primitive concepts, such as red and chair, from the training data. The trained models are then applied to predict novel compositional concepts e.g., red chair in the testing phase although these concepts were never seen during training.

[1] MetaReVision: Meta-Learning with Retrieval for Visually Grounded Compositional Concept Acquisition. Guangyue Xu, Parisa Kordjamshid, Joyce Chai. EMNLP-Finding, 2023

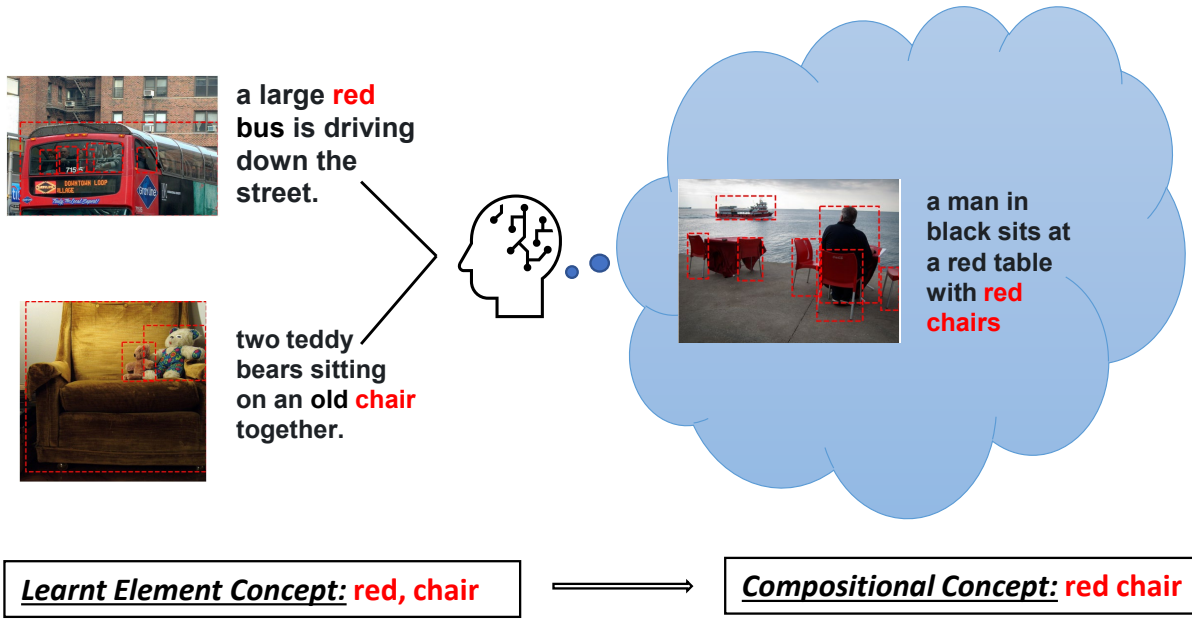


Figure 5.1 An illustration of Grounded Compositional Concept Learning(GCCL). For example, given concepts (red, bus) and (old, chair) in the training data, the goal is to learn to predict novel compositional concepts (red, chair) as masked token prediction at test time.

The ideal vision-language system should have the compositional ability to solve the GCCL problem. Recently, significant efforts have been made to the development of pre-training vision-language models (VLMs) [Tan and Bansal, 2019, Su et al., 2020, Radford et al., 2021]. These VLMs have demonstrated impressive performance in various downstream tasks, including Visual Question Answering (VQA) [Li et al., 2020a], Vision-Language Navigation (VLN) [Hao et al., 2020] and image captioning [Zhou et al., 2020]. Despite their success in related fields, it remains unclear whether these models can truly perceive the world in a compositional manner or generate language compositionally to cooperate with humans in a shared physical world. Such composition-related questions are important from both the theory and the application perspectives. From the theory perspective, compositional learning allows the model to process and understand objects by breaking them down into smaller, interpretable units. Therefore, compositional learning helps improve large models' efficiency and generalization [Andreas et al., 2016]. From the application perspective, it is not realistic to give the model all possible compositions in training data. For example, in Vision Language Navigation (VLN), it is not feasible to observe a sofa with all possible

colors e.g. red sofa and blue sofa. The vision-language models applied in VLN are expected to recognize these compositions after learning the element concepts [2]. Compositional learning can be viewed as a special case of zero-shot learning problems. Moreover, the domain-shift problem is commonplace in zero-shot learning because the statistical distribution of the data in the training set (seen compositions) and the testing set (novel compositions) could be significantly different. While compositionality can be reliably interpreted by humans, State-of-the-art VLMs, which are trained on vast amounts of image-text pairs and employ diverse loss functions, still encounter challenges in compositional learning [Ma et al., 2023a, Thrush et al., 2022].

To address these limitations, this thesis takes a closer look at the compositionality in VLM with an attempt to improve its ability. More specifically, we create two grounded compositional concept learning datasets, CompFlickr and CompCOCO curated from MSCOCO [Chen et al., 2015] and Flickr30K [Plummer et al., 2015], for VLMs' token-level compositional analysis. Moreover, we present MetaReVision, Meta-Learning with Retrieval for Visually Grounded Compositional Concept Acquisition, a retrieval-enhanced meta-learning framework for compositional concept acquisition, which introduces retriever into GCCL. The retrieval mechanism plays a crucial role in human learning. It facilitates long-term retention, understanding enhancement, and knowledge transfer during the learning process, which have been discussed by a large body of studies in cognitive science [Karpicke and Blunt, 2011, Karpicke, 2012]. To mimic such human's retrieving behavior[Roediger and Butler, 2011, Karpicke and Roediger III, 2008], MetaReVision retrieves relevant primitive concepts from a pre-constructed concept database and provides them as support evidence to do meta-learning for compositional concept learning. MetaReVision follows a *Learn-Retrieval-Compose* framework. It shares the compositional learning burden between VLMs and the retriever. Through meta-learning from the episodes constructed by the retriever, MetaReVision learns a generalized compositional representation that can be fast updated for novel compositional recognition. We evaluate MetaReVision on the proposed CompFlickr and CompCOCO datasets. The empirical results show that coupling retrieval and meta-learning performs

[2]Element concepts are also called primitive concepts in our setting. We use them interchangeably in this work.

better in GCCL compared with previous baselines.

Contributions of this work can be summarized as follows:

- This work explores a novel angle of retrieval-enhanced compositional concept learning. The model relies on retrieval to construct episodes for meta-learning. It addresses the domain-shift problem in compositional learning by learning from the retrieved instances.
- Two datasets are created to serve as benchmarks for grounded compositional concept learning. These datasets enrich existing zero-shot vision-language tasks, from the end-task level to the token-level.
- Our experiments show that MetaReVision demonstrates stronger performance in GCCL, especially in the novel setting. This empirically shows the effectiveness of combining retrieval and meta-learning techniques in the context of grounded compositional learning.

5.2 Related Works

Meta-Learning also known as learning to learn, aims to solve a low-resource problem by leveraging the learned experience from a set of related tasks. Meta-learning algorithms deal with the problem of efficient learning so that they can learn new concepts or skills fast with just a few seen examples (few-shot setting) or even without seen examples (zero-shot setting). Different from the typical meta-learning scenario where the training and test episodes are given in advance in few-shot learning [Sung et al., 2018a, Snell et al., 2017b, Nichol et al., 2018a, Finn et al., 2017], in GCCL, we need to construct episodes to employ meta-learning methods for compositional concept learning. In MetaReVision, we introduce a retriever to actively construct episodes to help compositional concept learning. During the test time, with additional retrieved support items, MetaReVision can further fast-update VLMs for current compositional concept recognition in the query set. This test-time fine-tuning is different from previous works which apply meta-learning in the zero-shot setting[Conklin et al., 2021].

Retrieval-Enhanced Learning. Retrieving related instances from a database, either the training set or external knowledge base, has been widely applied in tasks such as language modeling [Khandelwal et al., 2019], reinforcement learning [Goyal et al., 2022] and language tasks such as

NER [Wang et al., 2021]. Instead of distilling all training information into the model's parameters through gradient updates, retrieval-enhanced learning introduces a retriever to find related instances and based on these instances conduct further learning. For example, kNN-LM [Khandelwal et al., 2019] extends the pre-trained language model by linearly interpolating its next word distribution with a retrieval module. This design shows effective domain adaptation ability. [Wang et al., 2021] finds external contexts for the target instance by retrieving a set of semantically relevant texts to fine-tune the CRF module to address the NER problem. These studies highlight the significance of actively recalling information from a database to enhance learning outcomes. The general scheme of such methods is to combine a parametric model with a non-parametric retrieval system [Long et al., 2022]. Different from these settings, in GCCL, we train our own concept retriever and show retrieval's importance in compositional learning.

Compositional Learning. Recent research suggests that compositionality remains a challenge for state-of-the-art (SoTA) neural models such as Transformers and Graph Neural Networks [Nikolaus et al., 2019, Hupkes et al., 2020, SHAO et al., 2023]. To tackle this challenge, inspired by symbolic AI, some works try to add structural constraints into neural models [Bergen et al., 2021]. There are also some attempts to generate new data for the compositions [Naeem et al., 2023, Xian et al., 2018b]. Also, there have been noteworthy advancements in vision-language benchmarks that focus on probing and enhancing VLM's compositional abilities recently [Eisenschlos et al., 2023, Thrush et al., 2022, Ruis et al., 2020, Ma et al., 2023a]. Nevertheless, these works build end tasks in a compositional manner. They emphasize the performance of these compositional end tasks without giving consideration to the token-level compositional ability. However, GCCL targets VLM's token-level compositional ability. Moreover, different from symbolic and data-augment solutions, MetaReVision explores the retrieval method to solve the compositional problem.

5.3 Grounded Compositional Concept Learning (GCCL)

We start by introducing the settings of *Grounded Compositional Concept Learning (GCCL)* and further introduce the benchmarks we curated for this problem in this section.

5.3.1 Problem Definition

Existing VLMs try to learn a generic representation for multi-modal tokens in different contexts. These VLMs are expected to obtain generic token representations that have strong transfer ability for downstream tasks. We consider a setting that directly examines whether VLMs have the ability to acquire compositional meanings of tokens through the lens of language modeling. Different from the task-level compositional studies, GCCL approaches the compositional problem from the token-level and investigates whether VLMs possess the capability to acquire the compositional meanings of tokens.

Figure 5.2 GCCL task definition. Red highlights seen compositional concepts and blue highlights novel compositional concepts.

Figure 5.2 shows an example of the GCCL task. Given a set of image-caption pairs with the compositional concepts masked out from the caption, the model is tasked to learn the concept representations and predict the masked compositional concept conditioned on the contextual information. The learned model is then applied in the testing phase on both novel compositions as well as seen compositions. The model is evaluated based on its ability to learn novel compositions while maintaining (i.e., not forgetting) seen compositions.

Formally, given a set of text-image pairs $\{(x_{cap}, x_{img})\}_{i=1}^{n}$ where $x_{img} \in \mathcal{I}$ is the image with annotated bounding boxes, $x_{cap} \in \mathcal{T}$ is the caption with the compositional concepts replaced by *MASK*. The objective of GCCL is to predict the masked tokens based on the contextual information[Ma et al., 2023b, Jin et al., 2020]. Therefore, for BBoxes, only the locations are considered

as input, not their label information. A model capable of solving GCCL can be described as a functional $f : \mathcal{I} \times \mathcal{T} \rightarrow \mathcal{V}_{attr} \times \mathcal{V}_{obj}$, where $\mathcal{V}_{attr} \times \mathcal{V}_{obj}$ is the target compositional concepts which could be either *adjective* + *noun* pairs or *noun* + *verb* pairs. Based on whether $\mathcal{V}_{attr} \times \mathcal{V}_{obj}$ have been seen during training, GCCL can be categorized into seen compositional testing and novel compositional testing. The desired compositional VLMs should achieve improved novel performance without sacrificing the seen performance.

5.3.2 GCCL Dataset Creation

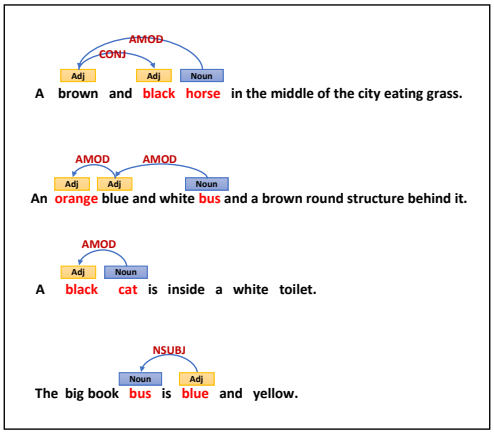
We build GCCL's benchmarks, CompFlickr and CompCOCO, from MSCOCO [Chen et al., 2015] and Flickr30K [Plummer et al., 2015]. We use the same data split introduced by [Nikolaus et al., 2019]. Their work studies the composition ability of image captioning systems by selecting 24 pairs as novel compositions by removing all images related to these 24 pairs from the training dataset. This ensures that novel compositions have never been seen during training. Other works adapt the same data split for compositional learning studies. For example, [Jin et al., 2020] utilized this split to check current VL models' compositional ability on phrases under the continual learning setting. However, in [Jin et al., 2020]'s work, most of the extracted phrases are in the form of article + noun, like the car and a man. They are single objects instead of compositional concepts. Such phase evaluation is not a good setting for compositional learning.

In order to evaluate the token-level compositional ability, we develop two benchmarks Compt-COCO and ComptFlickr to address the above limitation. Concretely, after paring the captions using Stanta [Qi et al., 2020], we use a number of rules to collect and mask the compositional concepts, the details are in the Figure 5.3. After parsing by Stanza, we can extract compositional pairs using the following rules. Compared with [Jin et al., 2020]' phase extracting rule, MetaReVision extracts more reasonable compositional pairs. Finally, the dataset is divided into 4 parts: training set without novel compositions, validation set with both seen and novel compositions for hyper-parameter tuning and model selection, seen test set, and novel test set. The detailed statistics of novel compositions for these two datasets are shown in Table 5.1. This table shows the statistics of the extracted novel compositional concepts. From the table, we can see that CompCOCO has more novel pairs

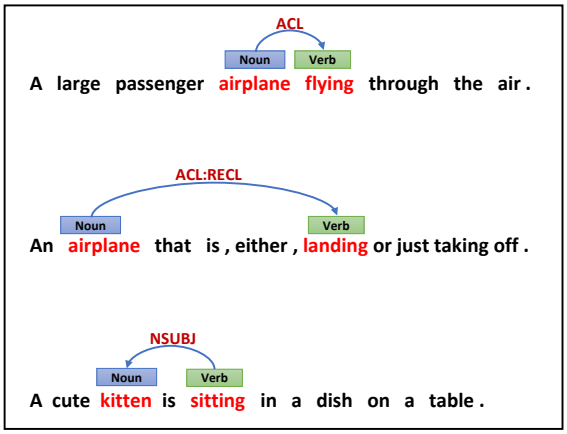
	MSCOCO				Flickr30K					
	Train Img.	Train Caps.	Test Img.	Test Caps.	Train Img.	Train Caps.	Val Img.	Val Caps.	Test Img.	Test Caps.
black bird	205	323	122	190	17	24	0	0	2	3
small dog	681	1067	316	481	360	612	11	12	17	33
white boat	373	261	196	134	69	85	0	0	3	8
big truck	417	601	191	288	28	38	0	0	1	1
eat horse	212	378	106	187	2	2	0	0	0	0
stand child	1288	1556	577	741	1048	1475	38	57	26	36
white horse	264	500	151	300	51	100	3	4	4	8
big cat	184	216	103	108	0	0	0	0	1	1
blue bus	276	506	143	243	11	16	0	0	0	0
small table	261	296	134	154	48	54	1	1	1	1
hold child	1328	1860	664	992	835	1289	27	37	35	60
stand bird	532	831	260	406	13	24	0	0	0	0
brown dog	613	878	291	430	934	1838	31	61	29	58
small cat	252	325	149	183	2	3	0	0	0	0
white truck	262	420	121	175	35	42	2	2	2	2
big plane	967	1345	357	494	5	5	0	0	0	0
ride woman	595	674	300	330	266	537	8	17	9	23
fly bird	245	526	132	283	29	53	0	0	0	0
black cat	840	1760	448	940	15	27	0	0	1	1
big bird	215	291	123	169	24	34	0	0	0	0
red bus	566	1212	232	474	11	20	0	0	1	1
small plane	481	833	158	279	13	20	0	0	0	0
eat man	555	698	250	314	153	272	4	5	5	10
lie woman	301	388	144	194	145	278	1	2	4	8

Table 5.1 Novel Pair Statistics for both CompCOCO and CompFlickr. We use the same 24 pairs to verify the compositional generalization.

than CompFlickr. And CompCOCO is a more reliable evaluation for novel compositional learning than And CompFlickr



(a) Rules to extract adj-noun pairs.



(b) Rules to extract verb-noun pairs.

Figure 5.3 Extracting rules to Construct CompFlickr and CompCOCO.

5.4 Meta-Learning with Retrieval for GCCL (MetaReVision)

Traditional word acquisition models typically learns a one-size-fits-all model from the entire training dataset and makes predictions for each test example in the inference phase. However,

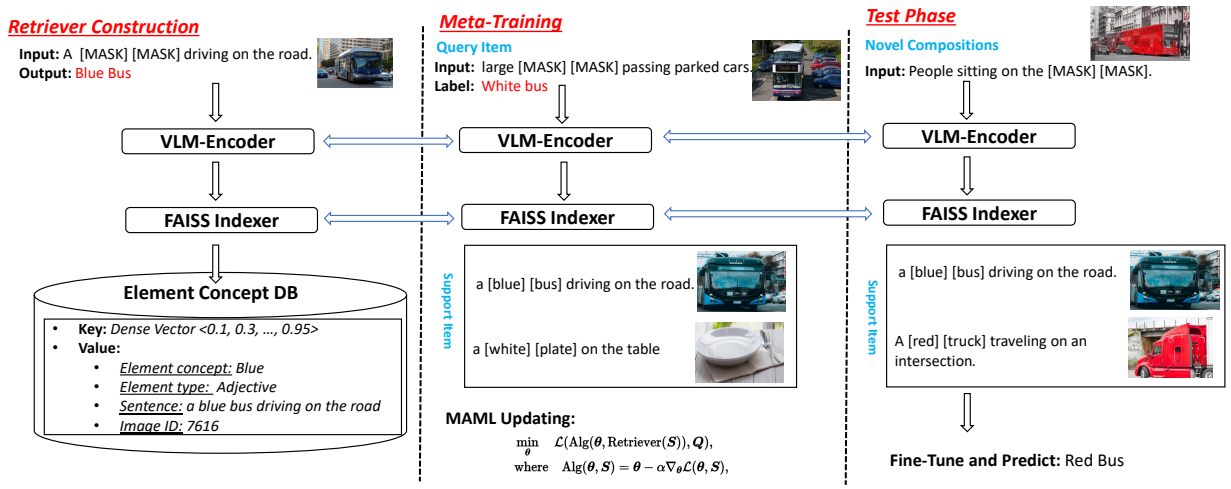


Figure 5.4 MetaReVision Architecture. The whole system includes two modules: retrieve and meta-trained VLM. During testing, MetaReVision retrieves related instances to fast-update VLM for novel compositional learning.

GCCL is a domain-shift problem and it is desirable to learn a customized model for each novel composition. In this work, we study to combine retrieval and meta-learning to address such customization and propose Meta-Learning with Retrieval for GCCLMetaReVision.

MetaReVision mainly consists of two modules: the retrieval model and the meta-learner as shown in Figure 5.4. The retrieval module learns to find similar element concepts from the training data. The meta-learner organizes the retrieved items as a pseudo task to meta-tune VLMs for compositional learning. In this part, we will discuss the base VLMs, retrieval module, and meta-learning module in detail and answer two key questions in MetaReVision's design: 1) How to retrieve related items, 2) How to utilize the retrieved items in the context of meta-learning.

5.4.1 Vision-Language Models (VLMs)

VLBERT [Su et al., 2020] and LXMERT [Tan and Bansal, 2019] are two representative VLMs that are suitable in our GCCL setting. They represent one-stream and two-stream VLMs separately. The difference is that two-stream VLMs have additional self-attention layers before cross-attention layers. We conduct experiments using these two types of VLMs to show the general effectiveness of the proposed framework. Moreover, all VLMs are trained from scratch to make sure that they do not see novel compositions during their training time.

5.4.2 Retriever and Element Concept Database

Given the compositional concepts, the ideal retriever is expected to retrieve the training examples that are the most beneficial for the target compositional concept learning. It is usually assumed that the examples that are the nearest neighbors of query examples are more likely to be beneficial ones for generalizing [Long et al., 2022]. GCCL retriever needs an encoder to encode the element concept, construct a database to organize these element concepts' information, and retrieve relevant concepts.

Element Concept Encoder. Given the linguistic and visual clues for the compositional concepts, the encoder is acting as a function $f(x_{cap}, x_{img})$ that maps a *MASK* concept to a fixed-length vector $\mathbb{R}^d$. Then for each primitive concept in the target compositions, $f(\cdot)$ can help retrieve related primitive concepts. MetaReVision relies on these retrieved concepts to conduct further compositional learning. In this way, MetaReVision enhances its own compositional capability by augmenting the input through the retrieval procedure. The encoding function $f(\cdot)$ is the key component for the retriever. In traditional vision-language tasks, like VQA and Visual Entailment[Song et al., 2022], CLIP [Radford et al., 2021] is usually used as the encoder to encode the whole visual or textural input and help build the retriever. However, in GCCL's token-level compositional setting, we focus on the token's representation and therefore use the VLMs as an encoder to extract *MASK* concept's representation for further compositional learning. These vectors are used as keys to construct the Element Concept Database and perform an approximate nearest neighbor search to augment compositional learning. We add a two-layer MLP and adopt Masked Language Modeling (MLM) to train vision-language retriever. For the encoder's training, since we focus on concept acquisition, words in compositional concepts are masked with a probability of 1.0, and others are not masked during training.

Element Concept Database. The element concept datastore $\mathcal{DB} = \{(k_i, v_i)\}$, which is constructed offline using the above-trained vision-language encoder, consists of dense representations of masked element concepts $k = Enc\left(x_{cap}, x_{img}\right) \in \mathbb{R}^d$ is as keys and the corresponding (x_{cap}, x_{img}) as values. To efficiently access this database, we implement the dense retriever for GCCL by an off-the-shelf-

retriever engine FAISS [Johnson et al., 2019] with a flat index (IndexFlatIP) without any training. Then given a masked concept, we can retrieve the top-K DB items by calculating the cosine similarity scores between the *[MASK]* concept with all DB items in nearly real-time as follows:

$$\text{Ret}(k) = \{(k_1, \text{Val}_1), \ldots, (k_M, \text{Val}_M)\} \tag{5.1}$$

where k is the mask concept's embedding vector, k_i is the DB item's key, $\text{Val}_i = (x_{cap_i}, x_{img_i})$ is the retrieved DB item's value, and Ret is the retrieved DB item set.

After adding the retrieval module into GCCL, the problem can be re-formulized as:

$$p(v \mid x) = \underbrace{p(v \mid x, Ret(x))}_{Learner} \underbrace{p(Ret(x) \mid x)}_{Retrieval} \tag{5.2}$$

where v is the *MASK* compositional concept's prediction, $x \in \mathbb{R}^d$ is the maksed concept's encoded vector and $Ret(x)$ is the retrieved DB items based on its vector x as Equation 5.1. The compositional learning happens in two levels: 1) retrieve related items from DB based on the encoding vector, 2) learn conditioned on contextual information and the retrieved items.

5.4.3 Meta-Learning for GCCL

Given the retrieved items, there are several ways to exploit these examples to facilitate compositional learning. The most direct method is to fine-tuning (FT). However, because the retrieved items are noisy and FT often faces over-fitting issues when they learn from a few labeled examples, FT does not help GCCL. Another choice in in-context learning [Wei et al., 2022]. However, as GCCL is a multi-modal problem. We have multiple image-caption pairs in the contextual input, current large multi-modals, like LLaVA [Liu et al., 2023] and GPT-4 [Achiam et al., 2023], can not be applied directly here. In MetaReVision, we choose meta-learning framework to utilize the retrieved items for GCCL. Meta-learning here is to train the base VLM with the ability to accumulate knowledge across episodes[3] and build internal generic representations for tokens that are suitable for compositional learning. Moreover, we introduce the verbalizer module to enforce the predicted concept for the query set coming from the retrieved support items. The verbalizer helps mitigate the

[3]episodes also called tasks in meta-learning.

memorization problem in meta-learning [Yin et al., 2019]. In the following part, we will discuss episode construction, the details about MAML, and *verbalizer* module used in MetaReVision.

Episode Constructions. We construct GCCL tasks τ_i for meta-learning as follows:

$$\tau_i = \left(\mathcal{D}_{\tau_i}^{\text{support}}, \mathcal{D}_{\tau_i}^{\text{query}} \right), \tag{5.3}$$

where $\mathcal{D}_{\tau_i}^{\text{support}}$ indicates the support set and $\mathcal{D}_{\tau_i}^{\text{query}}$ indicates the query set. Specifically, for one task, we randomly select one compositional concept as the query set. Then we retrieve a small number of examples that are similar to the query concepts. These retrieved items make up the support set. Meta-learning's objective in GCCL is to predict the compositional concepts in the query set after learning the element concepts in the support set. Here, episodes help VLMs to accumulate compositional knowledge and learn a generic compositional representation for masked concepts from the task-level instead of instance-level.

Meta-Learner. We use MAML [Finn et al., 2017] as our meta-learning algorithm. As an optimization-based method, MAML has two optimizing steps within each episode: the meta-train step and the meta-test step. In the meta-train step, MAML learns a task-specific learner θ' based on the current parameter θ and retrieved support items S. In the meta-test step, MAML updates the parameter θ based on the fast-updated parameter θ' and the compositional query items Q as shown in Figure 5.5. Moreover, MAML can be solved by formulating it as a bi-level optimization problem. Equation 5.2 can be extended to Equation 5.4.

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathcal{L}\left(\text{Alg}\left(\boldsymbol{\theta}, \text{Retriever}\left(\boldsymbol{S}\right)\right), \boldsymbol{Q}\right), \\ \text{where} \quad & \text{Alg}(\boldsymbol{\theta}, \boldsymbol{S}) = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{S}), \end{aligned} \tag{5.4}$$

where $\boldsymbol{\theta}$ is the learnt parameters, Retriever($\boldsymbol{S}$) stands for the retrieved DB items, $\boldsymbol{Q}$ is target compositional concept and Alg represents the optimization algorithm adapting to the support instances. There are different versions regarding Alg [Nichol et al., 2018b, Finn et al., 2017]. We use MAML which unrolls the optimizing process and tries to find a good initial parameter configuration for all compositions.

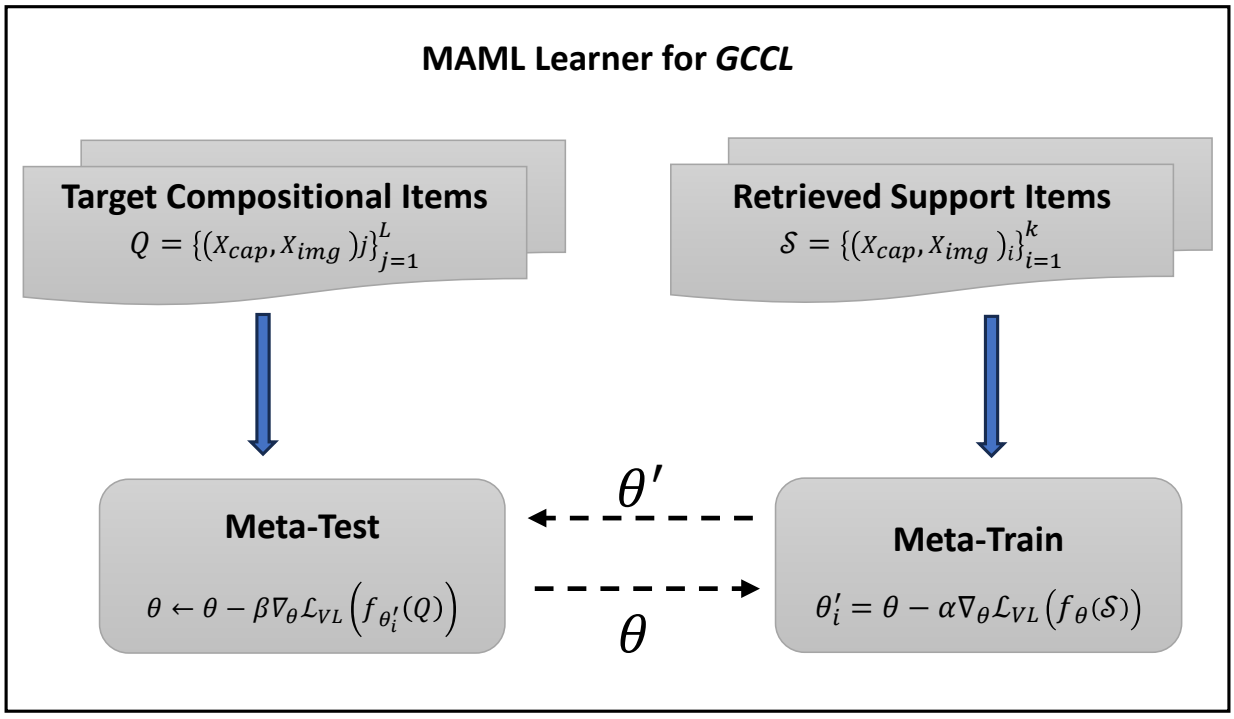


Figure 5.5 MAML's computing procedure.

Verbalizer. MAML's classical application is in few-shot learning, where class-to-label assignment needs to be conducted within each episode, that is, the same class has different labels among different episodes. Without such re-assignment, the models can memorize the class information and conduct prediction directly without considering the items in the support set. This is known as *memorization* problem in MAML discussed in [Yin et al., 2019]. To help MetaReVision learn from the retrieved instances, we introduce the *verbalizer* module into MetaReVision. It enforces prediction for the query set by selecting concepts from the support set as shown in Figure 5.6. In this way, MetaReVision will rely on the retrieved element concepts rather than memorizing the labels to do compositional learning. This helps alleviate the MAML's memorization problem.

5.4.4 Inference

During inference time, we consider each test compositional concept as a query item and retrieve relevant instances from concept DB as support instances. Therefore, we construct a specific task for the current compositional concept. Instead of applying the general model θ directly, MetaReVision retrieves support instances to fast-update the model to adapt to current compositions and make

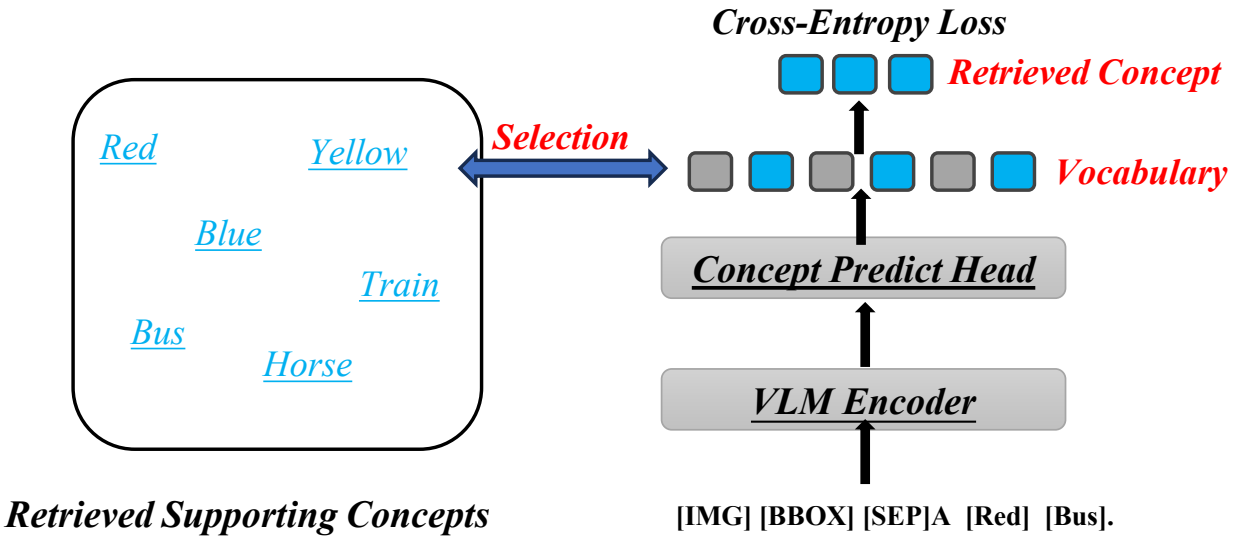


Figure 5.6 Verbalizer helps VLM consider retrieved instances when learning.

	VL-Model	VLBERT			LXMERT		
	Metric	Pair Accu.↑	Attr. Accu.↑	Obj. Accu.↑	Pair Accu.↑	Attr. Accu.↑	Obj. Accu.↑
COCO	Train-Scratch	7.73%	25.88%	50.74%	8.14%	26.36%	55.06%
	MAML w/o Ret.	9.03%	27.08%	50.04%	9.04%	27.01%	56.19%
	Ours(Top 4)	11.15%	29.84%	50.17%	12.01%	29.36%	58.81%
	Ours(Div 4)	13.50%	31.85%	50.92%	13.79%	33.76%	59.87%
Flickr	Train-Scratch	6.04%	17.53%	65.21%	5.12%	18.10%	61.68%
	MAML w/o Ret.	8.60%	22.06%	64.38%	7.52%	18.45%	64.55%
	Ours(Top 4)	10.7%	24.58%	65.54%	9.38%	20.45%	65.10%
	Ours(Div 4)	11.50%	25.49%	66.58%	10.58%	22.45%	65.15%

Table 5.2 MetaReVision's Results on Novel Compositional Concept.

predictions as $v_i = \text{argmax}_{v \in Sup} P(v)$, where the prediction comes from the retrieved concepts. In MAML's testing, it is observed that a larger number of updates can give a considerable performance boost. Thus, we choose the inner loop updates to 20 before testing.

5.5 Experiments

In this section, we introduce the GCCL's datasets, demonstrate the implementing details of MetaReVision, and compare its results with other baselines. Ultimately, we empirically analyze the retriever importance in MetaReVision.

	VL-Model	VLBERT			LXMERT		
	Metric	Pair Accu.↑	Attr. Accu.↑	Obj. Accu.↑	Pair Accu.↑	Attr. Accu.↑	Obj. Accu.↑
COCO	Train-Scratch	32.45%	49.06%	60.03%	34.12%	50.33%	61.96%
	MAML w/o Ret.	32.23%	49.05%	59.20%	34.09%	49.97%	61.93%
	Ours(Top 4)	32.27%	49.15%	59.98%	34.02%	49.90%	61.90%
	Ours(Div 4)	32.46%	50.01%	60.05%	34.15%	50.32%	62.00%
Flickr	Train-Scratch	24.34%	42.72%	52.53%	22.68%	40.86%	50.11%
	MAML w/o Ret.	23.73%	41.92%	49.01%	22.15%	41.21%	49.97%
	Ours(Top 4)	23.75%	41.95%	49.04%	22.75%	41.19%	50.01%
	Ours(Div 4)	26.52%	46.11%	53.23%	23.41%	42.02%	51.61%

Table 5.3 MetaReVision's Results on Seen Compositional Concept.

5.5.1 Dataset

CompCOCO is constructed from MSCOCO [Chen et al., 2015] using its 2014's split. In this split, COCO-captions has 103175 training images and 15112 validation images [Chen et al., 2015]. Because MSCOCO does not provide test data, we use the validation data as the testing data in CompCOCO. Moreover, in order to extract more compositional concepts, we modify [Lu et al., 2018]'s category and change the drier synonym list as: hair drier, hairdryer, hair dryer, blow dryer, blow drier, which helps to extract more clean concepts.

CompFlickr is constructed from Flickr30k Entities [Plummer et al., 2015]. Flickr30k contains $276k$ manually annotated bounding boxes for $31,783$ images and a total of $158,915$ English captions (five per image). We use the given train/val/test split to construct CompFlickr.

5.5.2 Evaluation Metrics.

We use accuracy as our primary metric to measure the GCCL performance and report object, attribute, and compositional accuracy separately. [Jin et al., 2020] uses perplexity as the forgetting metric in continual learning which is not appropriate in our work due to MetaReVision's offline setting.

5.5.3 Implementation Details

The implementation of MetaReVision uses the HuggingFace Transformers library [Wolf et al., 2020]. For MAML, we use Adam optimizer [Kingma and Ba, 2014] as both inner and outer optimizers. We set the inner learning rate to $5e-5$, the outer learning rate to $1e-5$, and based on

Target Context	Target Concepts	Retrieved Context	Retrieved Concepts
A white truck parked in front of a house that is being built.	White Truck	Several bikes parked next to a white van.	White Van
		A man in a suit poses by an colored truck.	Colored Truck
		A woman smiling in front of a big bus.	Big Bus
		People waiting on the side of the road for the yellow bus.	Yellow Bus.
A couple of birds flying through a cloudy sky.	bird fly	Two geese are flying in the air near trees.	Geese Fly
		Two hawks flying near a snow covered mountain.	Hawk Fly
		Two birds sit in the grass next to each other.	Bird Sit
		Two black birds are sitting on top of a mountain.	Black Bird.
a small boy is eating from a green plate	boy eat	A young boy is enjoying his pizza at the dinner table.	Boy Enjoy
		The little girl is eating lunch and having milk.	Girl Eat
		The woman is eating her meal at the table by herself.	Woman Eat
		An elderly couple is having a small snack in their kitchen.	Couple Have.
A brown dog is on the deck of a boat on water.	Brown Dog	A white and black dog laying on top of a yellow boat.	Black Dog
		a brown and black horse some green grass and some houses	Brown Horse.
		The black and white puppy is playing with a small toy.	Dog Play
		A white and black animal lays on a bench that is on grass outdoors.	White Animal
a blue bus with a large sign on the side of it.	Blue Bus	A red bus driving down a street in front of a red double decker bus.	Red Bus
		a red car driving down a city road on a cloudy day	Red Car.
		A red bus driving next to an orange and green bus.	Green Bus
		a red double decker bus a regular bus and a tow truck outdoors.	Regular Bus
blue bus parked in front of an azure building. A	Blue Bus	Two men in suits stand in front of a blue and white semi truck.	Blue Truck
		a white and black bus with a rainbow colored flag on the front	Black Bus.
		Four friends stand in front of an orange van.	Orange Van
		A large blue RV parked outside a large brick building.	Blue RV

Table 5.4 Episode examples constructed by MetaReVison's retrieval modules.

HIGHER [4] to calculate the higher gradients. The code for this chapter will be released at [5].

5.5.4 Episode Examples

Table 5.4 shows episode examples constructed in MetaReVision. From the table, we can see that MetaReVision can retrieve true element concepts for target compositional concepts, such as white truck, bird fly, boy eat. But there also exist cases we can not find true element concepts in the retrieved support set, such as blue bus. In this example, MetaReVision can retrieve many similar objects, but has a challenge to retrieve the true color blue. Also, from these randomly sampled episodes, we can see that in GCCL, objects are easier to be retrieved compared to objects.

5.5.5 Baselines

We use two types of baselines in this evaluation. The first is the *train-from-scratch baseline* which trains VLMs from random initialized parameters. Another baseline is *MAML without retriever*. In this setting, VLMs are meta-trained using the same retrieved tasks, but VLMs can not access the support set. It predicts directly during test time. This baseline is used to show the importance of the retriever during test time for GCCL. Moreover, we also compare two variants of MetaReVision, including *Top 4* and *Div 4*. *Top 4* retrieves top 4 similar concepts, which may contain duplicated concepts. The same concept could have different vector representation which

[4] https://github.com/facebookresearch/higher

[5] https://github.com/HLR/MetaReVision

is affected by different visual and textual contexts. For example, car could have different vector values when modified by *red* or *blue*. *Div 4* retrieves the top 4 distinct similar concepts expecting that the true primitive concept will be in the retrieved set.

5.5.6 Main Results

We report the performance under both novel and seen settings as shown in Table 5.2 and Table 5.3. From the two tables, we can see that MetaReVision does help compositional learning, especially in the novel setting.

Novel Compositions. As shown in Table 5.2, MetaReVision improves the performance on the novel setting compared to the pre-trained model and MAML models. This suggests that MetaReVision captures a generic representation which is beneficial for compositional learning through meta-learning on the retrieved tasks. However, compared with seen compositions (i.e., Table 5.3), the performance on novel pairs drops significantly across the board. MetaReVision's accuracy drops by about 20% on CompCOCO dataset in novel setting compared with the seen setting. This indicates that such compositional generalization is still a very difficult and open task for current VL models.

Seen Compositions. Table 5.3 shows the performance in the seen setting. From the table, we can see that all models have similar accuracy in the seen setting. One possible reason is that all the models have been fully trained using the seen compositional concepts. MAML-based methods do not hurt the in-domain performance during this meta-learning phase.

5.5.7 Empirical Analysis of Retriever

Retrieval Accuracy. Figure 5.7 shows the retriever's top-4 accuracy for attributes, objects, and pairs under both seen and novel settings. Attribute recognition is the key challenge compared with object recognition in GCCL, even in the retrieval phase. In GCCL, the learned VLMs are biased to the seen attributes that need to be adjusted for effective compositional learning.

Importance of diverse sampling. Retrieving true concepts into the support set is important for GCCL. In this part, we assume an oracle situation where we can always select the true element concepts into the support set during test time. We study potential advantages that can be derived under this configuration. From Figure 5.8, we can see that the true concept in the support set

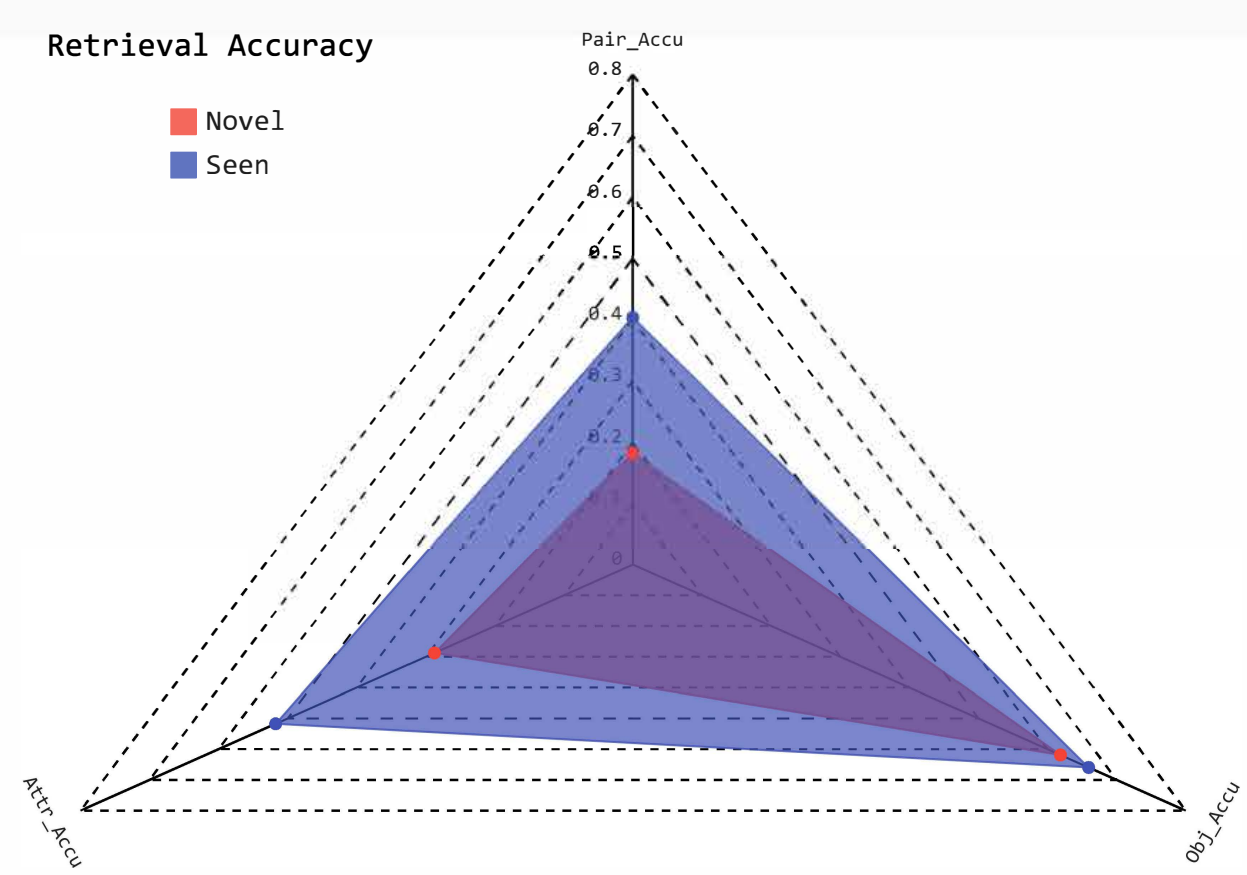


Figure 5.7 Comparison of the retriever accuracy between seen pairs and novel pairs in CompCOCO dataset.

does help the compositional learning. It also explains the importance of diverse sampling which increases the probability of selecting the correct elemental concepts.

5.6 Conclusions and Future Work

In this work, we propose MetaReVision, which combines retrieving method and meta-learning to train VLMs for grounded compositional concept learning. Our work highlights the significance of retrieval in compositional learning. Our empirical results on two proposed datasets, CompCOCO and CompFlickr, have shown that MetaReVision consistently outperforms conventional VLMs and meta-learning methods without retriever, especially in novel settings. However, GCCL is still a challenging open problem and many problems remain. Our future work will explore more cognitively plausible models and explicitly address the grounding ability in compositional concept learning.

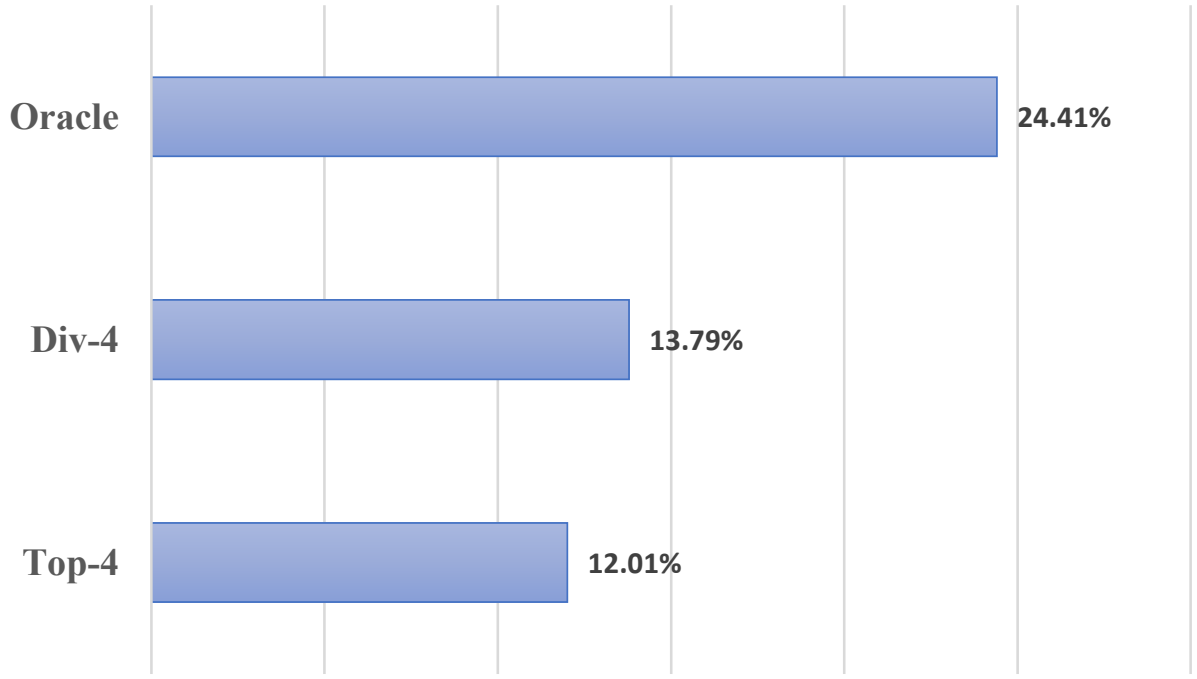


Figure 5.8 MetaReVision's accuracy on CompCOCO using different retrievers.

5.7 Limitations

The limitations of the proposed MetaReVision include 1) Grounding limitation. Currently, we rely on VLM's attention mechanism to do grounding. We do not have an explicit grounding design to align the textual concepts and visual regions. This could be an interesting direction for future GCCL works. 2) SoTA generative model comparisons. Currently, we can not directly apply SoTA generative models, such as BLIP-2 and MiniGPT, on GCCL due to the following reasons. One reason is the *GCCL problem setting*. In GCCL, it is not easy to transform the supporting items, including multiple images and captions, into contextual input for these generative models. Another reason is *controlled evaluation* which means that these huge generative models may have already seen the novel compositions during training and it is not a fair comparison with other models. 3) Updating retriever. We construct our element concept DB in advance and not updating this DB during the meta-learning time. Training both the learner and the retriever in an end-to-end manner could improve the performance for GCCL and other retrieval-enhanced models.

CHAPTER 6

GENCZSL: GENERATIVE COMPOSITIONAL ZERO-SHOT CONCEPT RECOGNITION

6.1 Introduction

The large generative language models, represented by GPT-series [Brown et al., 2020, Achiam et al., 2023], have achieved huge success in many natural language processing tasks. Moreover, with the scaling of model size and corpus size, these large language models demonstrate an in-context learning (ICL) ability. In this chapter, we aim to solve the compositional zero-shot learning problem through the application of the in-context learning paradigm. We propose leveraging foundation vision-language models to generate compositional concepts, thereby deviating from the conventional discriminative approaches which arim at aligning the compositional concepts with images in the constructed latent space.

While large language models have demonstrated remarkable in-context learning capabilities across various natural language processing tasks, deploying such paradigms in vision-language settings presents a considerable challenge. Applying in-context learning in vision-language models comes with the following set of challenges. However, directly applying Flamingo to generate compositional concepts for CZSL is challenging due to the following reasons: **1) Informative In-Context Example Selection**: Different from the few-shot setting, CZSL is essentially a zero-shot learning problem and the in-context examples are not available in CZSL problem. The ICL models need to find related examples to conduct compositional learning. Moreover, since both the selected examples and the order of the examples are important to the final performance in ICL [Zhang et al., 2022, Nguyen and Wong, 2023, Chang and Jia, 2023]. selecting and ranking informative in-context examples becomes a crucial element when applying Flamingo in the context of CZSL. **2) Mapping Between Predicted Tokens and Compositional Concept Labels**: Generative models can generate any tokens from its large vocabulary set based on the current contextual input. Mapping the generative tokens to the compositional labels is also challenging when applying ICL in CZSL. **3) Foundation Model Selection**: Handling sequences of arbitrarily interleaved visual and textual data

is a requirement for the foundation model. Effectively processing such mixed sequences demands the ability to seamlessly transition between visual and linguistic information. The recently proposed Flamingo aims to tackle the aforementioned challenges and gives a resolution for vision-language tasks in a few-shot setting where the input comprises interleaved textual and visual information.

To enable using generative models for CZSL, we propose a new approach called GenCZSL which is based on Flamingo to generate the compositional concepts. In our proposed technique, we use a retriever to select informative examples and ranker to further sort the retrieved example to help Flamingo in recognizing novel compositions. Concretely, given an image that corresponds to a novel compositional concept, GenCZSL applies CLIP's visual encoder to select related (img, concept) pairs to construct the candidate example pool for in-context learning.

Then GenCZSL introduces a ranker to sort the selected examples for in-context learning. For label mapping, instead of taking the **argmax** from the whole vocabulary, we restrict the model's output to a set of special tokens that correspond to the set of compositional labels, e.g., with the token "red car" corresponding to the compositional concepts. Overall, the contribution of this work can be summarized as follows:

- To the best of our knowledge, we are the first to apply the generative method to solve the compositional zero-shot learning problem. In contrast to previous discriminative models that train an alignment between images and compositional concepts, our work directly generates the corresponding compositional concept given an input image.
- We propose to use retrieval and ranking techniques for more effective in-context learning. Our experimental results show improved performance compared to basic in-context learning.

6.2 Preliminaries

6.2.1 In-Context Learning (ICL)

In this section, we present the background of in-context learning. We focus on in-context learning for CZSL using the vision-language model, Flamingo [Alayrac et al., 2022]. Given the vision-language model Flamingo, n relevant in-context examples for a specific task in hand, denoted as $\{x_i, y_i\}_{i=1}^{n}$ where x_i is an image and y_i is the related compositional concept (a_i, o_i), and a test

image input x_{test}, the compositional concept prediction for x_{test} is generated as follows:

$$y^* = \arg\max_{y \in \mathcal{Y}} p_G \left(y \mid x_1 \oplus y_1 \cdots x_m \oplus y_m \oplus x_{\text{test}} \right), \tag{6.1}$$

where $\mathcal{Y}$ is the compositional concept label space, $\oplus$ is the concatenation operation and m is the in-context example number. To deal with CZSL tasks, the original label is often mapped to word or words in Flamingo's vocabulary. As Equation 6.1 shows, Flamingo receives CZSL's supervision only from the concatenated $\{x_i, y_i\}_{i=1}^{m}$ and directly outputs the compositional concept prediction for the test image x_{test}. Typically, the number of in-context examples n is limited by the max input length of Flamingo. In previous works, the in-context examples are randomly sampled from the whole training dataset $\mathcal{D}$ [Brown et al., 2020]. However, recent researches have shown that ICL is sensitive to the provided examples and random in-context examples show significant instability and can cause inferior performance (Lu et al., 2022; Chen et al., 2022). In our work, we focus on selecting a small number of supporting in-context examples that are informative for the CZSL task and effective for in-context learning, from the entire dataset $\mathcal{D}$.

6.2.2 Foundation Model: Flamingo

Flamingo is a visual-language model that sets a new state-of-the-art in for few-shot learning on a wide range of open-ended multi-modal tasks. Flamingo can tackle a diverse spectrum of open-ended multimodal tasks with just a handful of task-specific examples in a few-shot setting, without any additional training required. Following the ICL paradigm, Flamingo takes input consisting of interleaved images and text and then outputs associated language as shown in Figure 6.1. Given a few example pairs of visual inputs and expected text responses composed in Flamingo's prompt, the model can be asked a question with a new image, and then generate an answer.

To address the challenge of fusing the information of interleaved images and texts, Flamingo introduces the following two key components in addition to standard auto-regression architecture:

- **Perceiver**. Flamingo uses perceiver to transforms image features from the vision encoder to a fixed number of visual outputs by the attention-based fusing mechanism. In particular,

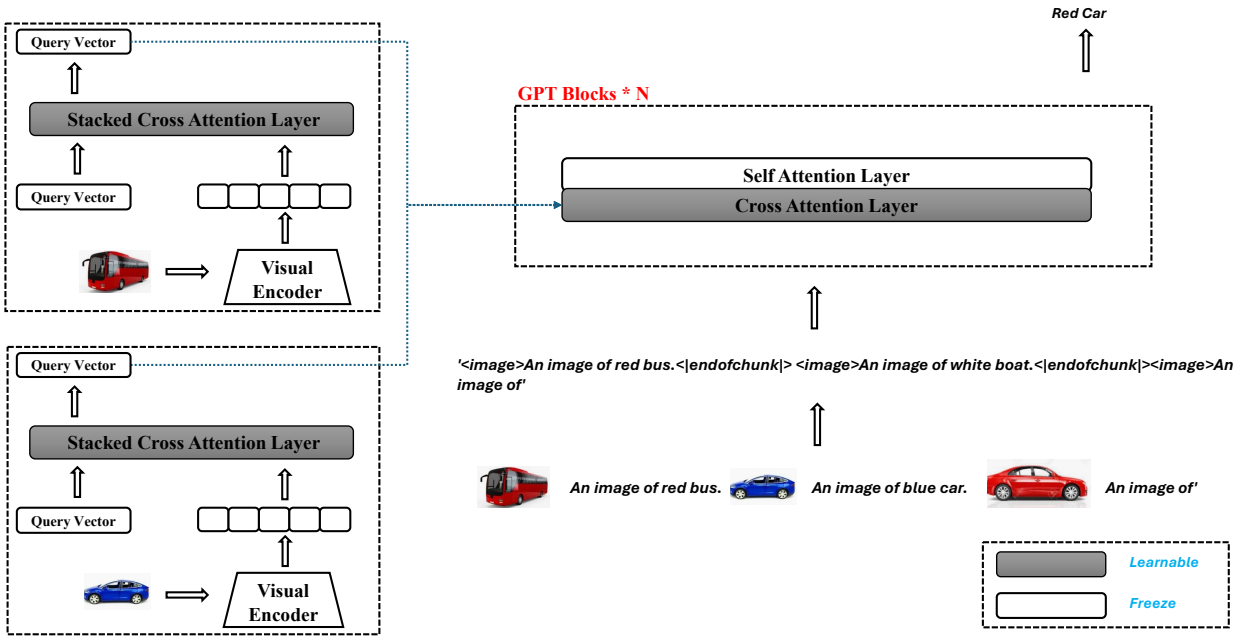


Figure 6.1 Flamingo architecture overview. Flamingo is a visual-language model that takes visual data interleaved with text as input and produces free-form text as output. It is originally proposed to address few-shot learning problem. Our work explores its in-context learning ability in CZSL.

Flamingo learns a predefined number of latent input queries which are fed to a Transformer and attend to the extracted visual features using attention machenism.

- **Multi-modal Fuser**. Flamingo freezes the pre-trained language model (LM) blocks, and inserts dense blocks of cross attention layers between the original LM layers for fuse information from visual input to textual input. And these inserted cross-attention layers are trained from scratch.

6.3 GenCZSL: Generative In-Context Learning for CZSL

In this section, we highlight the challenges of in-context learning for CZSL and explain our proposed solution based on Flamingo.

6.3.1 Challenges applying ICL in CZSL

Retrieving informative in-context examples is the critical challenge CZSL [Li et al., 2023b]. Different from the standard few-shot learning in ICL, CZSL requires selecting the examples for ICL first, And the difference is illustrated in Figure 6.2. As it is demonstrated in the figure, a few image-text examples are provided in advance in few-shot learning, and the pre-trained Flamingo

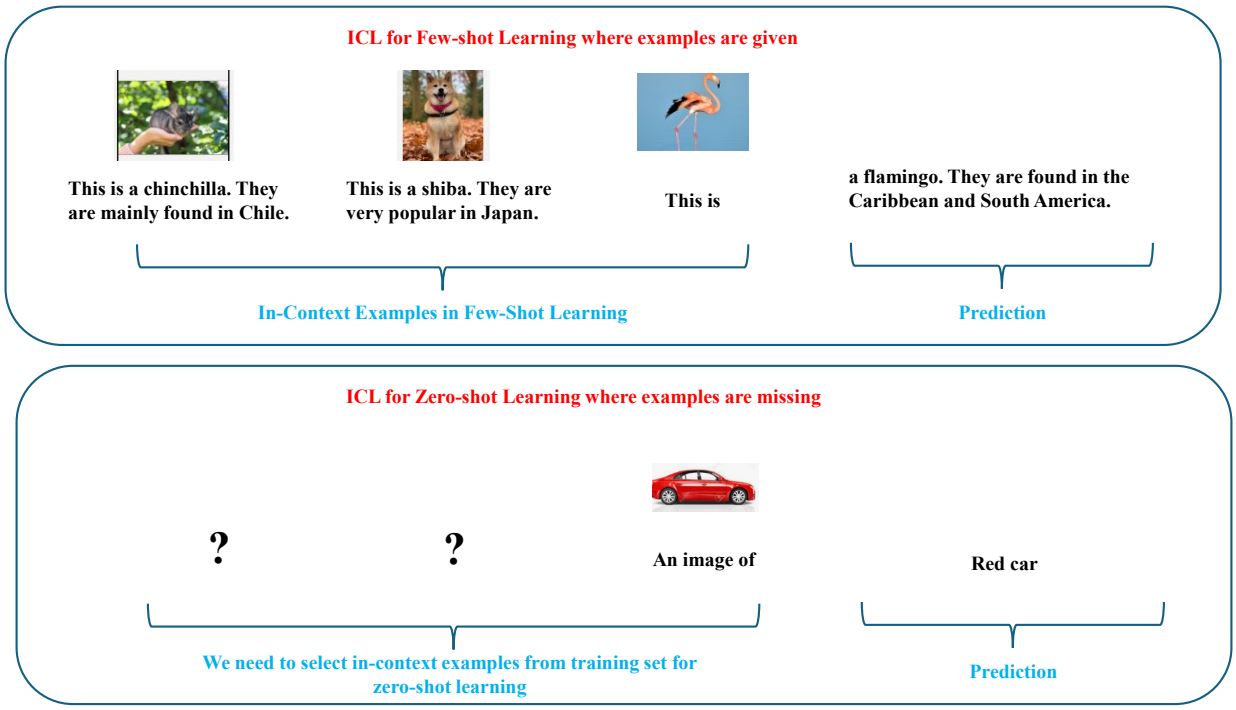


Figure 6.2 ICL difference between few-shot and zero-shot leanings.

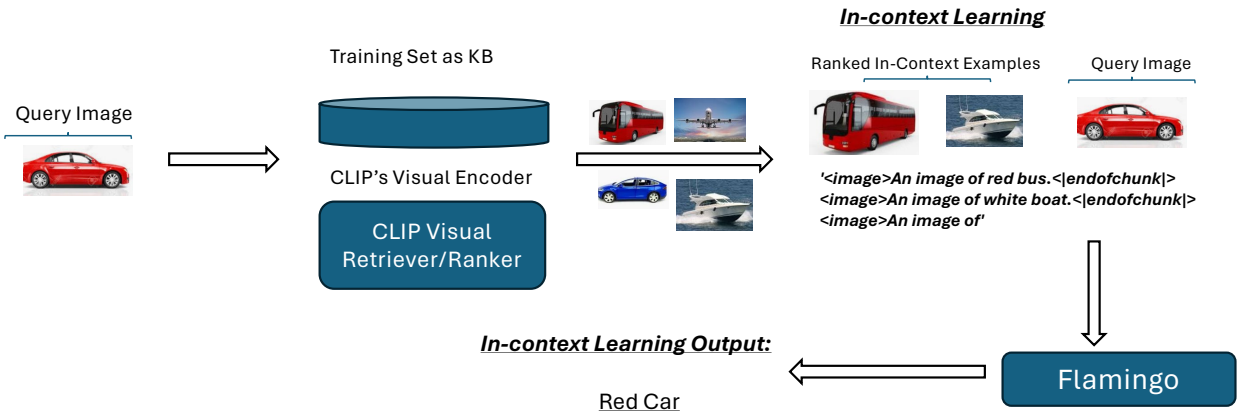


Figure 6.3 GenCZSL Architecture. GenCZSL uses the freezed CLIP's visual encoder to retrieve examples and uses ranker to sort the retrieved items. Flamingo is freezed in GenCZSL.

conducts prediction based on the concatenation of these demonstrated text-image pairs and the query image in a generative manner. However, since CZSL is a zero-shot setting problem, the in-context examples will not be provided. For a more accurate prediction, GenCZSL should retrieve related examples from the training set based on the query image and conduct prediction as Equation 6.1.

6.3.2 GenCZSL Architecture

In this section, we will the architecture of GenCZSL, especially focusing on how GenCZSL retrieve and rank the in-context examples to help the compositional learning. Overall, GenCZSL freezes the foundation vision-language model Flamingo and introduce two components, including retriever and ranker separately, to select and rank few-shot examples for Flamingo to conduct compositional learning as shown in Figure 6.3.

First Stage: Retriever. Selecting support examples within a context is a challenging task. The difficulty arises from the impracticality of considering all possible combinations and evaluating them, given the overwhelming complexity caused by the phenomenon of combinatorial explosion. Therefore, in the first stage, we aim to find those informative individual examples from the training dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the image, y_i is the related compositional label and N is the training set size. In this phase, we assume similar images to the query image will be more informative in ICL for CZSL. Based on this assumption, we introduce CLIP's pre-trained visual encoder [Radford et al., 2021] to select similar images from the training set based on the current query image. In such way, we first retrieve a set of relevant examples of size $n(n << N)$.

Second Stage: Ranker. Previous works on example selection for ICL [Chang and Jia, 2023, Ye et al., 2023, Lu et al., 2021] show that order of examples can impact the accuracy of the ICL's generation. Given these results, we introduce a ranking module for GenCZSL to reorder the retrived examples as shown in Figure 6.4. We approximate the ranking function using an MLP layer. Previous works mostly apply reinforcement-learning methods for example selection and ranking [Zhang et al., 2022] when using black box language models as the backbone generative model. However, the model architecture and parameter of Flamingo are all open-sourced and available which provides the possibility to back-propagate the gradients to ranker. Therefore, we adopt an easier and more efficient method to update the ranker to obtain a better ordering for the retrieved in-context examples.

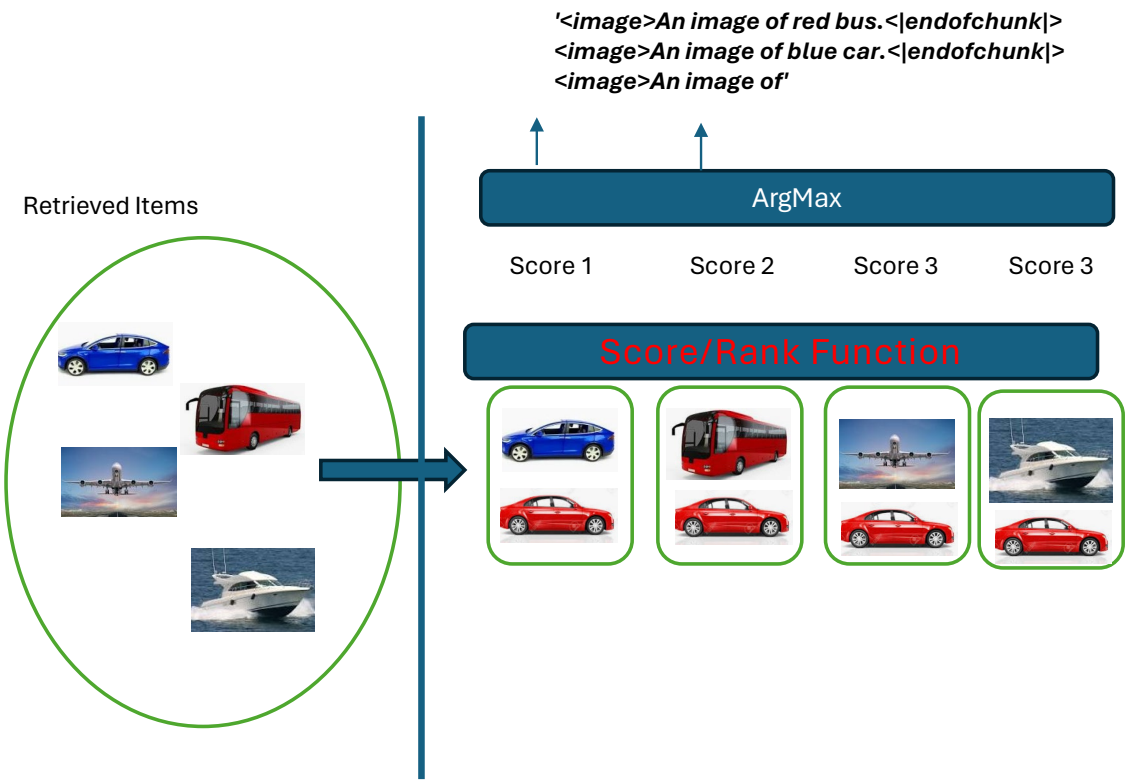


Figure 6.4 Ranker Architecture.

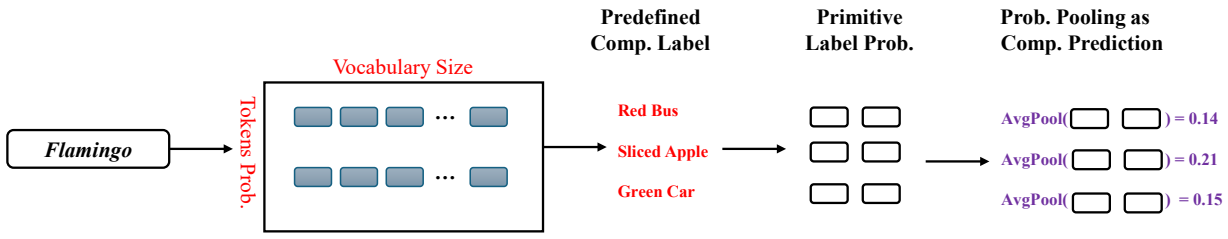


Figure 6.5 GenCZSL's Scoring Function. In CZSL, compositional labels include the element attribute and object labels. GenCZSL calculates the average of these element concept probability as the compositional prediction.

6.3.3 Scoring Functions for Composition Labels

As with other ICL methods, GenCZSL utilizes the scoring function to decide how the predictions of the generative model are mapped into an estimation of the likelihood of a specific label. GenCZSL uses the direct estimation method which uses the probability of candidate answers conditioned on the in-context inputs. The compositional labels are selected from the generated probability distribution for the tokens in Flamingo vocabulary and the most probable composition is selected afterwards.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP [Nagarajan and Grauman, 2018b]	14.3	17.4	9.9	1.6	59.8	54.2	40.8	25.9	17.0	5.6	5.9	0.7
LE+ [Misra et al., 2017a]	15.0	20.1	10.7	2.0	53.0	61.9	41.0	25.7	18.1	5.6	6.1	0.8
TMN [Purushwalkam et al., 2019b]	20.2	20.1	13.0	2.9	58.7	60.0	45.0	29.3	23.1	6.5	7.5	1.1
SymNet [Li et al., 2020b]	24.2	25.2	16.1	3.0	49.8	57.4	40.4	23.4	26.8	10.3	11.0	2.1
CompCos [Mancini et al., 2021]	25.3	24.6	16.4	4.5	59.8	62.5	43.1	28.7	28.1	11.2	12.4	2.6
CGE [Naeem et al., 2021]	32.8	28.0	21.4	6.5	64.5	71.5	60.5	33.5	33.5	15.5	16.0	4.2
SCEN [Li et al., 2022]	29.9	25.2	18.4	5.3	63.5	63.1	47.8	32.0	28.9	25.4	17.5	5.5
CLIP [Radford et al., 2021]	30.2	40.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
COOP [Zhou et al., 2022a]	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 [Nayak et al., 2022]	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
GIPCOL [Xu et al., 2024]	48.5	49.6	36.6	19.9	65.0	68.5	48.8	36.2	31.92	28.4	22.5	7.14
GenCZSL (random)	30.2	37.3	25.3	9.1	39.9	24.8	28.3	10.5				
GenCZSL (with Retriever)	34.6	42.4	30.6	12.5	48.7	30.2	32.5	16.3				
GenCZSL (with Ranker)	37.2	43.9	32.0	13.2	50.5	32.1	35.7	18.1				

Table 6.1 GenCZSL on Closed-World CZSL results on UT-Zappos, Mit-States and C-GQA datasets.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP [Nagarajan and Grauman, 2018b]	16.6	5.7	4.7	0.7	50.9	34.2	29.4	13.7	-	-	-	-
LE+ [Misra et al., 2017a]	14.2	2.5	2.7	0.3	60.4	36.5	30.5	16.3	19.2	0.7	1.0	0.08
TMN [Purushwalkam et al., 2019b]	12.6	0.9	1.2	0.1	55.9	18.1	21.7	8.4	-	-	-	-
SymNet [Li et al., 2020b]	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CompCos [Mancini et al., 2021]	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	-	-	-	-
CGE [Naeem et al., 2021]	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.1	1.8	2.9	0.47
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., 2022a]	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., 2022]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
GIPCOL [Xu et al., 2024]	48.5	16.0	17.9	6.3	65.0	45.0	40.1	23.5	31.6	5.5	7.3	1.30
GenCZSL (random)	37.6	9.7	10.3	2.6	46.8	28.5	18.4	9.1				
GenCZSL (with Retriever)	40.1	10.5	11.6	3.2	58.2	35.7	24.6	10.8				
GenCZSL (with Ranker)	41.2	10.4	12.0	3.6	60.1	38.5	23.7	11.2				

Table 6.2 Open-World CZSL results on UT-Zappos, MIT-States and C-GQA datasets.

This practice is similar to the way GPT is adapted to classification tasks [Brown et al., 2020]. In CZSL, we have a pre-defined set of compositional labels. We use AvgPool operation to average the element concept's probability and calculate each compositional label's probability accordingly, as shown in Figure 6.5.

6.4 Experiments

6.4.1 Dataset

We conduct experiments on three compositional zero-shot learning benchmarks, MIT-States [Isola et al., 2015a], UT-Zappos [Yu and Grauman, 2014] and C-GQA [Naeem et al., 2021]. MIT-States and C-GQA include images with the object and their attribute labels. The domain of these datasets

is very general. In contrast, UT-Zappos contains images of shoes paired with their material attributes which is a more domain-specific dataset. Our experiments follow the previous works [Purushwalkam et al., 2019b, Naeem et al., 2021] for the selection of train and test splits of the datasets. More details about the data splits and statistics can be found in Chapter 4.

6.4.2 Results

We compare our results with two types of baselines: 1) Task-specific architectures designed for CZSL and 2) CLIP-based methods in closed and open-world settings. The difference of these two settings is explained in Chapter 4. Both Table 6.1 and Table 6.2 show that although GenCZSL can not achieve SoTA results compared with CLIP-based methods, it obtains competitive results compared to task-specific architectures.

We conduct experiments for GenCZSL in three settings regarding different in-context example selecting methods: random example selection, retrieval-based example selection, and using an additional ranker to sort the retrieved examples. From the results, we can observe that compared with random sampled in-context examples, using CLIP's visual encoder help retrieve more informative examples for GenCZSL to solve the CZSL problem. Moreover, the introduction of the ranker can further improve the CZSL's performance using in-context learning methods.

6.5 Conclusion and Future Work

In this chapter, we provide an evaluation of in-context learning in solving the compositional zero-shot learning problem using the foundation vision-language model Flamingo. We propose an approach called GenCZSL to improve in-context learning for this multi-modal setting. To improve the efficacy of compositional zero-shot learning in GenCZSL, we focused on the selection and ranking of informative in-context examples. Especially, GenCZSL introduces a retriever to select more informative examples, and a ranker to reorder the selected examples, to help Flamingo conduct compositional learning. For future work, more analysis should be conducted to show what examples help Flamingo to do compositional learning, and whether these examples are as effective for human prediction.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this chapter, we summarize our work presented in this thesis, highlight the contributions and point to potential future directions.

7.1 Summary of Contributions

Compositional learning is a fundamental characteristic of human intelligence. This ability requires the computational models to understand that "the meaning of the whole is a function of the meanings of its parts". Although deep learning models have achieved huge success in many fields, they owe their success to training on large-scale datasets and have difficulties in adapting to new compositions. In this thesis, we focus on grounded compositional zero-shot learning (CZSL) and conduct experiments based on a variety of models. We demonstrate that large models struggle in compositional learning. Consequently, we provide various novel techniques and develop parameter-efficient methods to improve these models' compositional ability. Compared to previous CZSL methods, our proposed methods have achieved better performance on multiple benchmarks which demonstrates our significant contribution in advancing compositional learning. Our contribution which is explained in the chapters of this thesis can be summarized as follows.

- In Chapter 3, we study the problem of recognizing compositional attribute-object concepts within the zero-shot learning(ZSL) framework. We propose an episode-based cross-attention (EpiCA) network that combines the merits of the cross-attention mechanism and episode-based training strategy to recognize novel compositional concepts. Firstly, EpiCA is based on cross-attention to associate linguistic concepts with visual information and utilizes the gated pooling layer to build contextualized representations for both images and concepts. The updated representations are used for a more in-depth multi-modal relevance calculation for concept recognition. Secondly, a two-phase episode training strategy, especially the transductive phase, is adopted to utilize unlabeled test examples to alleviate the low-resource learning problem.
- In Chapter 4, we propose MetaReVision, a novel meta-learning framework to train vision-

language models for compositional learning. The episodic training and the bi-level optimization of meta-learning encourage gradients learned from the support set to be beneficial for compositional concept learning in the query set. Moreover, we created two datasets based on MSCOCO and Flicker30K to specifically target the evaluation of novel compositional concept learning with rich textual input.

- In Chapter 5, we propose GIPCOL, a new CLIP-based prompting framework, to solve the CZSL problem. GIPCOL models the interactions between element concepts via a graph neural network and learns rich compositional representation that are used to provide effective soft prompt to the CLIP model. Our experiments show that GIPCOL achieves better results compared with other prompting-based methods. Moreover, we analyze the importance of training data for compositional learning. Specially, our initial results have shown that GIPCOL performs better for a wider domain such as MIT-States and C-GQA, but less effective for a more specific domain such as UT-Zappos. These results demonstrate potential advantages and limitations in applying CLIP-based prompting approaches to compositional concept learning in the future.
- In Chapter 6, firstly we evaluate the large generative vision-language models in solving grounded CZSL problem and highlight their shortcomings. Moreover, we propose an effective in-context learning method to be used by such models. Our proposed approach is to select the most informative examples for in-context learning using a retriever module and use a ranker to reorder the selected examples and find their most effective order. This approach helps the VLM (Flamingo here) to better generalize over novel compositions. Our experiments show the effectiveness of the two retriever and ranker modules in the context of CZSL.

7.2 Future Directions

This dissertation explores different approaches to compositional learning, especially in the grounded compositional zero-shot learning field. To extend these approaches to a variety of real-world applications, possible future works are suggested as follows.

7.2.1 Explicit Grounded Compositional Learning

Visual-language models based on transformer architectures [Vaswani et al., 2017a] have achieved great success in many downstream tasks [Su et al., 2020, Tan and Bansal, 2019, Zhuge et al., 2021]. However, current pre-training strategies for these vision-language models usually depend on the attention mechanism to conduct implicit alignment between modalities that mainly focus on learning the coarse alignment between textual and visual input. Such methods lack the fine-grained alignment information between the visual regions and the textual tokens which could be a key component in compositional zero-shot learning [Thornberg et al., 2014]. Therefore, One possible research direction for Compositional Learning is explicit grounding vision-language models. Explicit grounding vision-language models are expected to consist of the following key features:

7.2.2 Exploring Diffusion Models for Compositional Learning

Recent research shows generative modeling is a crucial strategy for training artificial neural networks for discriminative tasks like image recognition [Hinton, 2007]. The recent large-scale text-to-image diffusion models have dramatically increased the text-based image generation abilities [Ho et al., 2020]. These generative models are trained to maximizes the evidence lower bound [Blei et al., 2017] of the given data's log-likelihood and learning to model the data distribution via an iterative noising and denoising procedure [Sohl-Dickstein et al., 2015]. These models can generate realistic images from textual prompts and exhibit impressive compositional generalization abilities. However, these diffusion models could be converted into classifiers which are useful for tasks beyond image generation, especially in the zero-shot setting. Diffusion classifier [Li et al., 2023a] is among the first works to apply diffusion models to discriminative tasks. However, directly utilizing such models in CZSl is still challenging. For example, how to filter out unfeasible compositional labels in the open CZSL setting is one challenge.

7.2.3 Retrieval-Based Compositional Learning

Humans recognize novel compositional concepts by recalling previously acquired primitive concepts and generalizing them to the novel compositional concepts even if they have never seen

the novel compositions before. However, deep learning models have difficulty in such compositional learning. In our thesis, we explored the importance of example selection for applying in-context learning in CZSL. In our current work, we freeze the foundation model Flamingo and train a ranker to sort the retrieved in-context examples. Such a design aims to adapt the ranker for the foundation model to conduct compositional learning. One possible direction is to design a better ranker to help explore the foundation model's compositional learning ability, such as retrieving more diverse in-context examples. Another direction is to learn the ranker and fine-tune the foundation model simultaneously. In the current design, the compositional learning burden mostly rests upon the ranker. Using bi-level optimization methods to train ranker and foundation models could be an interesting framework for solving compositional learning.

As more multi-modal applications start to enter our daily lives, it will be more important to equip intelligent agents with compositional ability and enable them to perceive complex environments they have never seen before. Despite the efforts we have made in this dissertation, a lot of important and interesting problems remain open. We believe that future research on this topic is of great value to make fundamental advances in AI.

BIBLIOGRAPHY

[Achiam et al., 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[Alayrac et al., 2022] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

[Anderson et al., 2018] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

[Andreas, 2019] Andreas, J. (2019). Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*.

[Andreas et al., 2016] Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. (2016). Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

[Antol et al., 2015] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

[Arnab et al., 2021] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

[Bagad et al., 2023] Bagad, P., Tapaswi, M., and Snoek, C. G. (2023). Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516.

[Baroni and Zamparelli, 2010] Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1183–1193.

[Bergen et al., 2021] Bergen, L., O'Donnell, T., and Bahdanau, D. (2021). Systematic generalization with edge transformers. *Advances in Neural Information Processing Systems*, 34:1390–1402.

[Biederman and Vessel, 2006] Biederman, I. and Vessel, E. A. (2006). Perceptual pleasure and the brain: A novel theory explains why the brain craves information and seeks it through the senses. *American scientist*, 94(3):247–253.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

[Brown et al., 2020] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Bugliarello et al., 2020] Bugliarello, E., Cotterell, R., Okazaki, N., and Elliott, D. (2020). Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint arXiv:2011.15124*.

[Carnap, 1988] Carnap, R. (1988). *Meaning and necessity: a study in semantics and modal logic*, volume 30. University of Chicago Press.

[Chai et al., 2018] Chai, J. Y., Gao, Q., She, L., Yang, S., Saba-Sadiya, S., and Xu, G. (2018). Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9.

[Chang et al., 2016] Chang, M. B., Ullman, T., Torralba, A., and Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*.

[Chang and Jia, 2023] Chang, T.-Y. and Jia, R. (2023). Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144.

[Chao et al., 2016] Chao, W.-L., Changpinyo, S., Gong, B., and Sha, F. (2016). An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 52–68. Springer.

[Chen and Grauman, 2014] Chen, C.-Y. and Grauman, K. (2014). Inferring analogous attributes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 200–207.

[Chen et al., 2020a] Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., and Han, J. (2020a). Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663.

[Chen et al., 2015] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[Chen et al., 2020b] Chen, Y., Gong, S., and Bazzani, L. (2020b). Image search with text feedback by visiolinguistic attention learning. pages 3001–3011.

[Chorowski et al., 2015] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. *Advances in neural information processing systems*, 28:577–585.

[Conklin et al., 2021] Conklin, H., Wang, B., Smith, K., and Titov, I. (2021). Meta-learning to compositionally generalize. *arXiv preprint arXiv:2106.04252*.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

[Dhillon et al., 2019] Dhillon, G. S., Chaudhari, P., Ravichandran, A., and Soatto, S. (2019). A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Dou et al., 2019] Dou, Z.-Y., Yu, K., and Anastasopoulos, A. (2019). Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

[Eisenschlos et al., 2023] Eisenschlos, J. M., Cole, J. R., Liu, F., and Cohen, W. W. (2023). WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

[Finn et al., 2017] Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:1126–1135.

[Fodor, 1975] Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard university press.

[Fodor and Pylyshyn, 1988a] Fodor, J. A. and Pylyshyn, Z. W. (1988a). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

[Fodor and Pylyshyn, 1988b] Fodor, J. A. and Pylyshyn, Z. W. (1988b). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

[Gao et al., 2023] Gao, K., Chen, L., Zhang, H., Xiao, J., and Sun, Q. (2023). Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*.

[Girdhar et al., 2022] Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., and Misra, I. (2022). Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

[Goyal et al., 2022] Goyal, A., Friesen, A., Banino, A., Weber, T., Ke, N. R., Badia, A. P., Guez, A., Mirza, M., Humphreys, P. C., Konyushova, K., et al. (2022). Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pages 7740–7765. PMLR.

[Gu et al., 2018] Gu, J., Wang, Y., Chen, Y., Cho, K., and Li, V. O. (2018). Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

[Hao et al., 2020] Hao, W., Li, C., Li, X., Carin, L., and Gao, J. (2020). Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13137–13146.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Hermann, 2014] Hermann, K. M. (2014). *Distributed Representations for Compositional Semantics*. PhD thesis.

[Hessel et al., 2021] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

[Hinton, 2007] Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Progress in brain research*, 165:535–547.

[Ho et al., 2020] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Holla et al., 2020] Holla, N., Mishra, P., Yannakoudakis, H., and Shutova, E. (2020). Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation. *arXiv preprint arXiv:2004.14355*.

[Hou et al., 2020] Hou, Z., Peng, X., Qiao, Y., and Tao, D. (2020). Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer.

[Huang et al., 2023a] Huang, X., Huang, Y.-J., Zhang, Y., Tian, W., Feng, R., Zhang, Y., Xie, Y., Li, Y., and Zhang, L. (2023a). Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310.

[Huang et al., 2023b] Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., and Zhang, L. (2023b). Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*.

[Hudson and Manning, 2019] Hudson, D. A. and Manning, C. D. (2019). Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

[Hupkes et al., 2020] Hupkes, D., Dankers, V., Mul, M., and Bruni, E. (2020). Compositionality decomposed: how do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795.

[Huynh and Elhamifar, 2020] Huynh, D. and Elhamifar, E. (2020). Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems*, 33:19849–19860.

[Isola et al., 2015a] Isola, P., Lim, J. J., and Adelson, E. H. (2015a). Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391.

[Isola et al., 2015b] Isola, P., Lim, J. J., and Adelson, E. H. (2015b). Discovering states and transformations in image collections.

[Jia et al., 2021] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

[Jin et al., 2021] Jin, W., Cheng, Y., Shen, Y., Chen, W., and Ren, X. (2021). A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

[Jin et al., 2020] Jin, X., Du, J., Sadhu, A., Nevatia, R., and Ren, X. (2020). Visually grounded continual learning of compositional phrases. In *EMNLP*.

[Johnson et al., 2019] Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

[Karpicke, 2012] Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, 21(3):157–163.

[Karpicke and Blunt, 2011] Karpicke, J. D. and Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018):772–775.

[Karpicke and Roediger III, 2008] Karpicke, J. D. and Roediger III, H. L. (2008). The critical importance of retrieval for learning. *science*, 319(5865):966–968.

[Karthik et al., 2022] Karthik, S., Mancini, M., and Akata, Z. (2022). Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning.

[Kato et al., 2018] Kato, K., Li, Y., and Gupta, A. (2018). Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251.

[Kemp and Tenenbaum, 2009] Kemp, C. and Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, 116(1):20.

[Khandelwal et al., 2019] Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. (2019). Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

[Kim and Linzen, 2020] Kim, N. and Linzen, T. (2020). Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Kipf and Welling, 2016] Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

[Lake and Baroni, 2018] Lake, B. and Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning*, pages 2873–2882.

[Lake, 2019] Lake, B. M. (2019). Compositional generalization through meta sequence-to-sequence learning. *arXiv preprint arXiv:1906.05381*.

[Lake et al., 2015a] Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015a). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

[Lake et al., 2015b] Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015b). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

[Lake et al., 2017] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

[Lee et al., 2018] Lee, K.-H., Chen, X., Hua, G., Hu, H., and He, X. (2018). Stacked cross attention for image-text matching. pages 201–216.

[Li et al., 2023a] Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., and Pathak, D. (2023a). Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*.

[Li et al., 2018] Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. (2018). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[Li et al., 2019a] Li, K., Min, M. R., and Fu, Y. (2019a). Rethinking zero-shot learning: A conditional visual classification perspective. *The IEEE International Conference on Computer Vision (ICCV)*.

[Li et al., 2023b] Li, X., Lv, K., Yan, H., Lin, T., Zhu, W., Ni, Y., Xie, G., Wang, X., and Qiu, X. (2023b). Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

[Li et al., 2019b] Li, X., Sun, Q., Liu, Y., Zhou, Q., Zheng, S., Chua, T.-S., and Schiele, B. (2019b). Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*, pages 10276–10286.

[Li et al., 2022] Li, X., Yang, X., Wei, K., Deng, C., and Yang, M. (2022). 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.

[Li et al., 2020a] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020a). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

[Li et al., 2020b] Li, Y.-L., Xu, Y., Mao, X., and Lu, C. (2020b). Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325.

[Li et al., 2020c] Li, Y.-L., Xu, Y., Mao, X., and Lu, C. (2020c). Symmetry and group in attribute-object compositions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325.

[Liu et al., 2023] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

[Liu et al., 2021] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

[Long et al., 2022] Long, A., Yin, W., Ajanthan, T., Nguyen, V., Purkait, P., Garg, R., Blair, A., Shen, C., and van den Hengel, A. (2022). Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6959–6969.

[Lu et al., 2019] Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. pages 13–23.

[Lu et al., 2018] Lu, J., Yang, J., Batra, D., and Parikh, D. (2018). Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

[Lu et al., 2021] Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2021). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

[Ma et al., 2023a] Ma, Z., Hong, J., Gul, M. O., Gandhi, M., Gao, I., and Krishna, R. (2023a). Crepe: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10910–10921.

[Ma et al., 2023b] Ma, Z., Pan, J., and Chai, J. (2023b). World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. *arXiv preprint arXiv:2306.08685*.

[Mancini et al., 2021] Mancini, M., Naeem, M. F., Xian, Y., and Akata, Z. (2021). Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5222–5230.

[Mancini et al., 2022] Mancini, M., Naeem, M. F., Xian, Y., and Akata, Z. (2022). Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on pattern analysis and machine intelligence*.

[Miller and Charles, 1991] Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

[Misra et al., 2017a] Misra, I., Gupta, A., and Hebert, M. (2017a). From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801.

[Misra et al., 2017b] Misra, I., Gupta, A., and Hebert, M. (2017b). From red wine to red tomato: Composition with context. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801.

[Mokady et al., 2021] Mokady, R., Hertz, A., and Bermano, A. H. (2021). Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

[Munkhdalai et al., 2018] Munkhdalai, T., Yuan, X., Mehri, S., and Trischler, A. (2018). Rapid adaptation with conditionally shifted neurons. In *International Conference on Machine Learning*, pages 3664–3673. PMLR.

[Naeem et al., 2023] Naeem, M. F., Khan, M. G. Z. A., Xian, Y., Afzal, M. Z., Stricker, D., Van Gool, L., and Tombari, F. (2023). I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15169–15179.

[Naeem et al., 2021] Naeem, M. F., Xian, Y., Tombari, F., and Akata, Z. (2021). Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962.

[Nagarajan and Grauman, 2018a] Nagarajan, T. and Grauman, K. (2018a). Attributes as operators. *ECCV*.

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

[Nan et al., 2019] Nan, Z., Liu, Y., Zheng, N., and Zhu, S.-C. (2019). Recognizing unseen attribute-object pair with generative model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8811–8818.

[Nayak et al., 2022] Nayak, N. V., Yu, P., and Bach, S. H. (2022). Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*.

[Nguyen and Wong, 2023] Nguyen, T. and Wong, E. (2023). In-context example selection with influences. *arXiv preprint arXiv:2302.11042*.

[Nichol et al., 2018a] Nichol, A., Achiam, J., and Schulman, J. (2018a). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

[Nichol et al., 2018b] Nichol, A., Achiam, J., and Schulman, J. (2018b). On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

[Nikolaus et al., 2019] Nikolaus, M., Abdou, M., Lamm, M., Aralikatte, R., and Elliott, D. (2019). Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

[Nye et al., 2020] Nye, M. I., Solar-Lezama, A., Tenenbaum, J. B., and Lake, B. M. (2020). Learning compositional rules via neural program synthesis. *arXiv preprint arXiv:2003.05562*.

[Ontanón et al., 2021] Ontanón, S., Ainslie, J., Cvicek, V., and Fisher, Z. (2021). Making transformers solve compositional tasks. *arXiv preprint arXiv:2108.04378*.

[Pan and Yang, 2009] Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

[Partee et al., 1995] Partee, B. et al. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

[Pennington et al., 2014a] Pennington, J., Socher, R., and Manning, C. (2014a). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Pennington et al., 2014b] Pennington, J., Socher, R., and Manning, C. D. (2014b). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Plummer et al., 2015] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

[Purushwalkam et al., 2019a] Purushwalkam, S., Nickel, M., Gupta, A., and Ranzato, M. (2019a). Task-driven modular networks for zero-shot compositional learning. *arXiv preprint arXiv:1905.05908*.

[Purushwalkam et al., 2019b] Purushwalkam, S., Nickel, M., Gupta, A., and Ranzato, M. (2019b). Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602.

[Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

[Radford et al., 2021] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[Raffel et al., 2020] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

[Rao et al., 2022a] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. (2022a). Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Rao et al., 2022b] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., and Lu, J. (2022b). Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091.

[Ravi and Larochelle, 2016] Ravi, S. and Larochelle, H. (2016). Optimization as a model for few-shot learning.

[Roediger and Butler, 2011] Roediger, H. L. and Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1):20–27.

[Ruis et al., 2020] Ruis, L., Andreas, J., Baroni, M., Bouchacourt, D., and Lake, B. M. (2020). A benchmark for systematic generalization in grounded language understanding. *Advances in neural information processing systems*, 33:19861–19872.

[Schuhmann et al., 2021] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. (2021). Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

[SHAO et al., 2023] SHAO, N., Cai, Z., xu, H., Liao, C., Zheng, Y., and Yang, Z. (2023). Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.

[Sharma et al., 2018] Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

[Snell et al., 2017a] Snell, J., Swersky, K., and Zemel, R. (2017a). Prototypical networks for few-shot learning. *Advances in neural information processing systems*, pages 4077–4087.

[Snell et al., 2017b] Snell, J., Swersky, K., and Zemel, R. S. (2017b). Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

[Song et al., 2022] Song, H., Dong, L., Zhang, W.-N., Liu, T., and Wei, F. (2022). Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

[Su et al., 2020] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2020). Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

[Sun et al., 2023] Sun, S., Liu, Y., Iter, D., Zhu, C., and Iyyer, M. (2023). How does in-context learning help prompt tuning? *arXiv preprint arXiv:2302.11521*.

[Sung et al., 2018a] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018a). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

[Sung et al., 2018b] Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., and Hospedales, T. M. (2018b). Learning to compare: Relation network for few-shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

[Surís et al., 2020] Surís, D., Epstein, D., Ji, H., Chang, S., and Vondrick, C. (2020). Learning to learn words from visual scenes. *European Conference on Computer Vision (ECCV)*.

[Tan and Bansal, 2019] Tan, H. and Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

[Thornberg et al., 2014] Thornberg, R., Perhamus, L., and Charmaz, K. (2014). Grounded theory. *Handbook of research methods in early childhood education: Research methodologies*, 1:405–439.

[Thrush et al., 2022] Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

[Touvron et al., 2023] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[Tsai et al., 2020] Tsai, Y.-Y., Chen, P.-Y., and Ho, T.-Y. (2020). Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*, pages 9614–9624. PMLR.

[Tsimpoukelli et al., 2021] Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

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

[Vaswani et al., 2017a] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017a). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[Vaswani et al., 2017b] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017b). Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

[Velickovic et al., 2017] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. (2017). Graph attention networks. *stat*, 1050(20):10–48550.

[Vinyals et al., 2016] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. *Advances in neural information processing systems*, pages 3630–3638.

[Vinyals et al., 2015] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

[Wang et al., 2021] Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., and Tu, K. (2021). Improving named entity recognition by external context retrieving and cooperative learning. *arXiv preprint arXiv:2105.03654*.

[Wang et al., 2019] Wang, X., Yu, F., Wang, R., Darrell, T., and Gonzalez, J. E. (2019). Tafe-net: Task-aware feature embeddings for low shot learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840.

[Wei et al., 2022] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

[Wei et al., 2019] Wei, K., Yang, M., Wang, H., Deng, C., and Liu, X. (2019). Adversarial fine-grained composition learning for unseen attribute-object recognition. pages 3741–3749.

[Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

[Xian et al., 2018a] Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. (2018a). Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551.

[Xian et al., 2018b] Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. (2018b). Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551.

[Xu et al., 2024] Xu, G., Chai, J., and Kordjamshidi, P. (2024). Gipcol: Graph-injected soft prompting for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5774–5783.

[Xu et al., 2022] Xu, G., Kordjamshidi, P., and Chai, J. (2022). Prompting large pre-trained vision-language models for compositional concept learning. *arXiv:2211.05077*.

[Xu et al., 2021] Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. (2021). Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

[Yang et al., 2022] Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., and Wang, L. (2022). An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

[Ye et al., 2023] Ye, J., Wu, Z., Feng, J., Yu, T., and Kong, L. (2023). Compositional exemplars for in-context learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

[Yin et al., 2019] Yin, M., Tucker, G., Zhou, M., Levine, S., and Finn, C. (2019). Meta-learning without memorization. *arXiv preprint arXiv:1912.03820*.

[Yu and Grauman, 2014] Yu, A. and Grauman, K. (2014). Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 192–199.

[Yu and Grauman, 2017] Yu, A. and Grauman, K. (2017). Semantic jitter: Dense supervision for visual comparisons via synthetic images. *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579.

[Yu et al., 2016] Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. (2016). Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer.

[Zellers et al., 2019] Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

[Zhang et al., 2022] Zhang, Y., Feng, S., and Tan, C. (2022). Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*.

[Zhang et al., 2023] Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., et al. (2023). Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*.

[Zhang and Sabuncu, 2018] Zhang, Z. and Sabuncu, M. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, pages 8778–8788.

[Zhou et al., 2022a] Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022a). Conditional prompt learning for vision-language models. In *CVPR*.

[Zhou et al., 2022b] Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhou et al., 2020] Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., and Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

[Zhu et al., 2023] Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

[Zhuge et al., 2021] Zhuge, M., Gao, D., Fan, D.-P., Jin, L., Chen, B., Zhou, H., Qiu, M., and Shao, L. (2021). Kaleido-bert: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12647–12657.