

Knowledge-Aware Prompt Tuning for Generalizable Vision-Language Models

Baoshuo Kan^{1*}, Teng Wang^{2,3*}, Wenpeng Lu^{1†}, Xiantong Zhen⁴, Weili Guan⁵, Feng Zheng^{2†}

¹Qilu University of Technology (Shandong Academy of Sciences)

²Southern University of Science and Technology ³The University of Hong Kong

⁴United Imaging Healthcare ⁵Monash University

10431200583@stu.qlu.edu.cn tengwang@connect.hku.hk wenpeng.lu@qlu.edu.cn

zhenxt@gmail.com weili.guan@monash.edu f.zheng@ieee.org

Abstract

Pre-trained vision-language models, e.g., CLIP, working with manually designed prompts have demonstrated great capacity of transfer learning. Recently, learnable prompts achieve state-of-the-art performance, which however are prone to overfit to seen classes, failing to generalize to unseen classes. In this paper, we propose a Knowledge-Aware Prompt Tuning (KAPT) framework for vision-language models. Our approach takes the inspiration from human intelligence in which external knowledge is usually incorporated into recognizing novel categories of objects. Specifically, we design two complementary types of knowledge-aware prompts for the text encoder to leverage the distinctive characteristics of category-related external knowledge. The discrete prompt extracts the key information from descriptions of an object category, and the learned continuous prompt captures overall contexts. We further design an adaptation head for the visual encoder to aggregate salient attentive visual cues, which establishes discriminative and task-aware visual representations. We conduct extensive experiments on 11 widely-used benchmark datasets and the results verify the effectiveness in few-shot image classification, especially in generalizing to unseen categories. Compared with the state-of-the-art Co-CoOp method, KAPT exhibits favorable performance and achieves an absolute gain of 3.22% on new classes and 2.57% in terms of harmonic mean.

1. Introduction

Recently, large-scale pre-trained vision-language models, e.g., CLIP [29], ALIGN [15], and FLIP [44], have demonstrated remarkable performance in zero/few-shot learning tasks. Unlike traditional vision-only frameworks

* Equal contribution. † Corresponding author. Work done when Baoshuo Kan visited to Feng Zheng Lab in SUSTech.

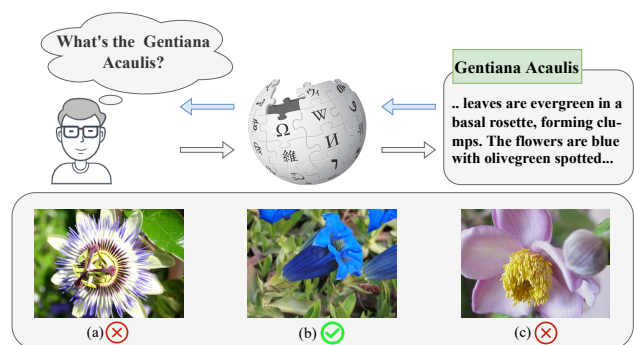


Figure 1: **A motivating example.** The textual description of the “Gentiana Acaulis” conduces to the recognition of the corresponding image (b) of Gentiana Acaulis.

that are trained mainly by a closed set of single-modal data, vision-language models train two uni-modal encoders on massive amounts of image-text pairs to exploit cross-modal alignments in the semantic space. By leveraging large-scale web-scale image-text data, pre-trained vision-language models are endowed with the ability to solve zero/few-shot downstream tasks and even recognize open-set visual concepts [29, 15, 44]. Expressly, when a new classification task arrives, the CLIP text-encoder encodes manually designed textual prompt (e.g., “a photo of a [label].”), and then cosine similarity between textual features and image features is computed. However, identifying appropriate manually designed prompts is an art that requires both domain expertise and laborious prompt engineering.

To avoid the manual prompt design, some recent research (e.g., CoOp [49]) on visual representations are mainly inspired by prompt tuning approaches [47, 18, 21] in Natural Language Processing (NLP), like learnable prompts. By optimizing their models with learnable prompts in closed datasets, these methods achieve outstanding performance in seen classes. However, the learned prompts are usually prone to overfit to the seen classes







Prompts	Dataset	Acc.	Dataset	Acc.	Dataset	Acc.
Original Prompts		42.4 ☹️		24.2 ☹️		83.7 ☹️
Knowledge Prompts	DTD	42.8 ☺️	EuroSAT	25.2 ☺️	OxfordPets	86.4 ☺️
Original Prompts		60.9 ☹️		74.5 ☹️		86.0 ☺️
Knowledge Prompts	Flowers101	67.3 ☺️	Food101	75.9 ☺️	Caltech	83.4 ☹️

Figure 2: **External knowledge improves the generalizability of CLIP.** We perform zero-shot classification with discrete prompts on six image classification datasets. The original prompt from CLIP [29] is *a photo of a [label]*. The proposed knowledge prompt concatenates category-related external knowledge from the Wikipedia Encyclopedia with the original prompt.

and suffer from insufficient generalization ability to unseen classes under the same task.

Recently, CoCoOp [48] was developed to improve the generalizability. The model constructs specific prompts by conditioning them on each instance, which achieves stronger robustness to category shift. Nonetheless, their learnable prompts are shared across all categories, which leads to weak discrimination between distinct characteristics of different categories; meanwhile, the model is incapable of perceiving factual details for class labels due to lack of fine-grained textual information on category, especially for uncommon classes that are rarely encountered during pre-training or have poor relevance to seen classes. Thus, the CoCoOp model still falls short of the generalizability to unseen scenarios.

Inspired by how humans utilize knowledge bases to learn novel visual concepts, we propose to incorporate external knowledge into prompt learning by leveraging accurate descriptions of concepts and their contextual relationships to overcome the aforementioned issues. As a motivating example, we can see in Figure 1 that it is generally hard for humans to imagine the visual appearance of uncommon categories, and even harder to recognize them when seeing a scientific name for the first time (*e.g.*, *Gentiana acaulis*). However, once we learn the key characteristics by reading the textual description from the knowledge base, it becomes much easier to recognize the images and the correspondence between different categories. Considering that textual descriptions of scientific names are unparalleled and contain identifying authentication information, we assign unique knowledge for each category to improve the generalization ability by enhancing the discrimination of prompts.

Specifically, we present a novel, Knowledge-Aware Prompt Tuning (KAPT) framework for vision-language models. We first retrieve encyclopedic knowledge related to category names from Wikipedia Encyclopedia contain-

ing a large number of entity descriptions. To take full advantage of category-related external knowledge, we design two complementary types of prompts: discrete and learnable continuous prompts. Discrete prompts carry the summarized texts that directly describe the visual appearance of the category, and learnable continuous prompts carry contextual information that may cover a broader background of the category. As shown in Figure 2, a preliminary experiment verifies that the proposed discrete knowledge-aware prompt improves performance of CLIP on several image datasets. Meanwhile, to further adapt the visual representation towards a specific task for inhibiting disturbance of task-irrelevant visual concepts, we propose an adaptation head that refines the image features by attending to the salient visual cues relevant to categories of the target task.

The main contributions can be summarized as follows:

- We propose a novel prompt tuning framework for vision-language models by incorporating external knowledge, which greatly improves the generalizability on unseen object categories.
- We design two complementary types of knowledge-aware prompts, which enables the model to fully exploit category-related dense knowledge retrieved from the Wikipedia Encyclopedia.
- We further propose a task-aware visual adaptation head to aggregate the attentive visual features conditioned on linguistic description of categories, which maximally capture task-related visual cues, while suppressing the disturbance caused by task-irrelevant visual concepts.
- Extensive experimental results on 11 popular image datasets demonstrate the effectiveness of the proposed method. Our method significantly outperforms state-of-the-art methods on the overall metric in the base-to-new generalization setting.

2. Related Work

2.1. Vision-Language Pre-training

In recent times, the introduction of extensive image-text data into pre-trained vision-language models has emerged as a prominent trend [29, 15, 44, 19, 20, 5, 39]. A representative work is CLIP [29], which aggregates 400 million image-text pairs from websites, facilitating the vision-language representation learning using a contrastive objective. Contemporary work with CLIP, ALIGN [15] also takes advantage of a large-scale dataset, 1.8 billion noisy image-text pairs, to pre-train a model with contrastive loss. These vision-language models are dual-encoder architectures consisting of an image encoder and a text encoder. Leveraging

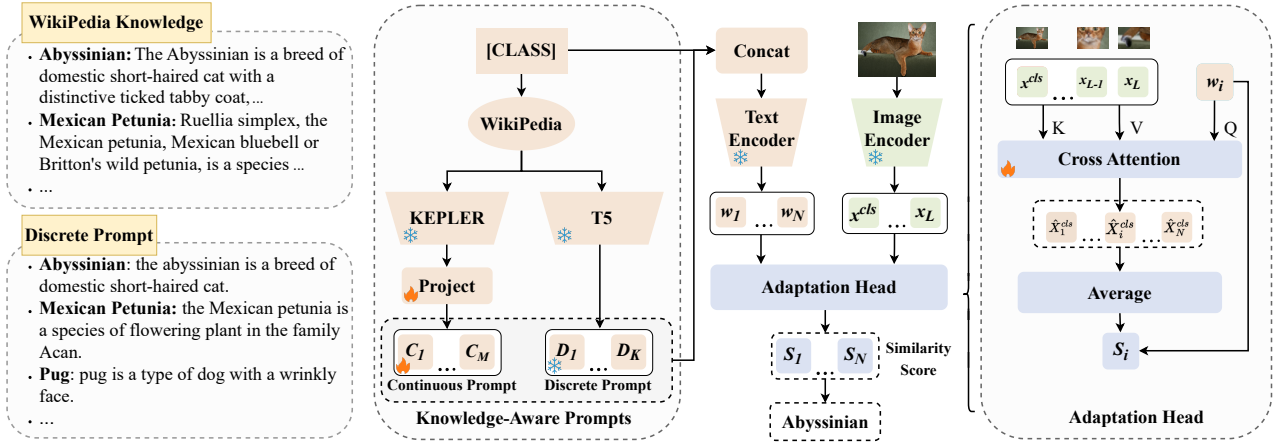


Figure 3: **Overview of our proposed KAPT (Knowledge-Aware Prompt Tuning) for vision-language models.** We first retrieve textual descriptions of task labels from the external knowledge base. Then, we construct the discrete and continuous knowledge prompts to enhance the discrimination of prompts for each category. The discrete, continuous prompts and the class label are concatenated as the input for the text encoder. The output tokens for the image encoder are modulated by an adaptation head, which aggregates task-related visual cues conditioned on high-level text features. The final classification confidence is obtained by calculating the cosine similarity between visual and text embeddings. Wikipedia knowledge (top left) is sourced from Wikipedia Encyclopedia, denoted as {category:description}. Discrete prompts (lower left) are summaries of corresponding descriptions in Wikipedia. * and 🔥 represent frozen and tunable weights during tuning, respectively.

extensive image-text pairs and dual-encoder architectures, these approaches showcase remarkable prompt-based zero-shot performance across diverse visual classification tasks, by exploiting alignments between text and image features.

2.2. Prompt Learning

With the continuous parameter scaling of pre-training models like GPT-3 [3] and CLIP, fine-tuning the entire models for downstream tasks becomes daunting because of inefficiency in parameters and potential catastrophic forgetting [28]. Notably, recent works [21, 4, 47] have introduced prompt learning that exclusively fine-tune a limited parameter subset, yielding robust results in NLP tasks. Inspired by the swift proliferation of prompt learning within NLP, the computer vision domain has also delved into prompt tuning for resolving downstream tasks [38, 45, 1, 16, 41, 8]. In this context, manual prompt templates within CLIP (e.g., “a photo of a [label]”) have been employed for image recognition. However, research in NLP [10] reveals that identifying suitable manual prompts demands both domain expertise and laborious prompt engineering. Some works [49, 31] adopt learnable continuous prompts to make the vision-language model recall the task-relevant knowledge. Due to the weak generalizability of simple learnable prompts, Co-CoOp [48] proposes conditional prompts by further learning a lightweight neural network to generate an input-conditional token for each image. Different from previous works, we advance the generalization of prompt learning by incorporating external knowledge, achieved by integrating accurate descriptions of concepts and their relationships.

2.3. External Knowledge Bases

In natural language processing (NLP), external knowledge bases, such as WordNet [25] and ConceptNet[35], are frequently harnessed to enhance performance [22, 46]. Early attempts [42, 24] in the computer vision community have also verified its effectiveness in visual question answering. Another notable work K-LITE[33], distinctively employs external knowledge in the vision-language pre-training phase, yielding visual models with better transferability and sample efficiency. However, within the scope of prompt tuning for vision-language models, there is little attention on harnessing external knowledge for model generalization. Moreover, our knowledge base and task-related knowledge extraction methods, derived from visual entity understanding, distinctly set our work apart from previous NLP research.

3. Methodology

KAPT is a prompt tuning method that incorporates category-related external knowledge into vision-language pre-training models. Our method builds upon CLIP [29] to effectively leverage its strong zero/few-shot transferability (Section 3.1). To construct knowledge-aware prompt tuning, we propose two variants of prompts to take full advantage of category-related external knowledge and a task-aware visual adaptation head to adapt the visual representation toward a specific task (Section 3.2). The overall framework is illustrated in Figure 3.

3.1. Preliminary: Prompting for CLIP

The CLIP model [29] is a typical dual-encoder architecture consisting of an image encoder and a text encoder. For each image-text pair, an image and the paired text are transformed into high-dimensional embeddings by image encoder (e.g., ResNet [11]) and text encoder (e.g., Transformers [36]), respectively. The training objective is to align the uni-modal embeddings by contrastive learning, where the model pulls paired image-text together and pushes the unpaired ones away in latent space. By pre-training on 400M large-scale web datasets, the learned visual representations are discriminative and transferable to zero/few-shot downstream tasks.

For the zero-shot transfer to image classification with N classes, CLIP constructs a simple prompt “a photo of a [label]”, fills *label* with each class name C_i , and hence obtains the corresponding textual embedding \mathbf{w}_i by the text encoder. Meanwhile, the image I is fed into the image encoder to generate image embedding \mathbf{x} . The prediction probability is acquired by calculating the cosine similarity between the image embedding and N text embeddings:

$$p(y = i|\mathbf{x}) = \frac{\exp(\text{sim}(\mathbf{x}, \mathbf{w}_i)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{x}, \mathbf{w}_j)/\tau)}, \quad (1)$$

where $\text{sim}(\cdot, \cdot)$ represents the computation of cosine similarity, and τ represents the temperature ratio.

3.2. Knowledge-Aware Prompt Tuning (KAPT)

The manually designed prompts of CLIP show inferior performance compared with learnable prompts trained on few-shot datasets. However, learnable prompts usually overfit to seen classes, suffering from the weak generalizability problem in unseen classes. To improve the generalizability to novel concepts under the same task, we present a novel prompt tuning method called Knowledge-Aware Prompt Tuning (KAPT). Specifically, we design two complementary knowledge-aware prompts to take full advantage of category-related external knowledge, and we propose a task-aware visual adaptation head to capture task-related visual cues.

Summarized Knowledge. Category-related knowledge is retrieved from the open-source external knowledge base. Note that from existing external knowledge bases [25, 35], it is difficult to retrieve related knowledge for rare or fine-grained concepts, thus resulting in a lower coverage of categories in mainstream vision datasets. For a high coverage rate of object categories of downstream datasets, we use textual descriptions sourced from Wikipedia Encyclopedia to form a category knowledge base \mathcal{K} for a specific downstream task. However, the resultant textual knowledge is usually expatiatory and contains irrelevant information for visual recognition. To remove redundant informa-

tion from category-related external knowledge, we feed the knowledge into T5 [30], an end-to-end pre-training model which takes text as input and is expected to output modified text summarizing the description of each category. The short text summary is considered as the automatic discrete prompt D for capturing the key description of categories.

Contextual Knowledge. In simplified text descriptions, some category-related information is inevitably filtered as redundant information. In order to make prompts carry contextual information that may cover a broader background of the category, we feed category-related external knowledge retrieved from \mathcal{K} into a pre-trained KEPLER [40] model to generate continuous features. Furthermore, we map obtained continuous features to the multi-modal embedding space by employing a lightweight projector which is composed of two linear layers with a bottleneck structure. The features are then combined with the context vectors to construct the learnable continuous prompt C for the corresponding category. Note that the context vectors are learnable parameters with random initialization.

Knowledge-Aware Prompts. To take full advantage of category-related external knowledge and considering learning-based continuous prompts have a higher risk of overfitting towards seen classes than discrete prompts, we concatenate learnable continuous prompt C_{y_i} , y_i , and automatic discrete prompt D_{y_i} to build knowledge-aware prompt for category y_i . Feeding knowledge-aware prompts into the text encoder $f^T(\cdot)$, we obtain the text features $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^N$ which are used to calculate similarity with visual representations and provided to adaptation head as the salient cues relevant to categories.

Adaptation Head. Given an image I , the image encoder $f^I(\cdot)$ transforms I into a set of visual feature vectors $\mathbf{X} = [\mathbf{x}^{cls}, \mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_L]$. To further adapt the visual representation towards a specific task to reduce disturbance of task-irrelevant visual concepts, we construct the task-aware visual adaptation head to focus the attentive visual features by attending to the salient cues relevant to categories. For each text features \mathbf{w}_i , we take it as query vector, and image features \mathbf{X} as the key and value vector.

$$\hat{\mathbf{X}}_i^{cls} = LN(\mathbf{x}^{cls} + \text{CrossAttention}(\mathbf{w}_i; \mathbf{X})), \quad (2)$$

where $\hat{\mathbf{X}}_i^{cls}$ represents the enhanced image features, $\text{CrossAttention}(\cdot)$ refers to cross attention, LN is Layer Normalization and x^{cls} denotes the [CLS] token of the input image. To converge all information of $\hat{\mathbf{X}}_i^{cls}$, we obtain the mean image features by computing the average of all enhanced image features $[\hat{\mathbf{X}}_i^{cls}]_{i=1}^N$.

$$\bar{\mathbf{X}} = \text{AvgPool}([\hat{\mathbf{X}}_i^{cls}]_{i=1}^N) = \frac{1}{N} \sum_{i=1}^N \hat{\mathbf{X}}_i^{cls}, \quad (3)$$

Table 1: **Evaluation on the base-to-new generalization setting.** Prompt-based methods learn their prompts from the base classes with 16 shots. We report the accuracy on base and new classes. H (Harmonic mean) evaluates overall performance.

(a) Average over 11 datasets.				(b) ImageNet.				(c) Caltech101.			
	Base	New	H		Base	New	H		Base	New	H
CLIP	64.83	70.27	67.44	CLIP	67.50	64.00	65.70	CLIP	93.70	94.00	93.84
CoOp	79.88	59.39	68.12	CoOp	71.30	62.40	66.55	CoOp	97.30	89.10	93.01
CoCoOp	76.70	67.30	71.69	CoCoOp	70.90	66.66	68.71	CoCoOp	97.13	93.06	95.05
KAPT	78.41	70.52	74.26	KAPT	71.10	65.20	68.02	KAPT	97.10	93.53	95.28
(d) OxfordPets.				(e) StanfordCars.				(f) Flowers102.			
	Base	New	H		Base	New	H		Base	New	H
CLIP	86.80	96.30	91.30	CLIP	60.90	69.90	65.09	CLIP	68.70	72.40	70.50
CoOp	91.20	89.73	90.46	CoOp	75.16	50.63	60.50	CoOp	96.46	54.23	69.43
CoCoOp	93.43	94.66	94.04	CoCoOp	65.86	66.03	65.94	CoCoOp	90.43	63.66	74.72
KAPT	93.13	96.53	94.80	KAPT	69.47	66.20	67.79	KAPT	95.00	71.20	81.40
(g) Food101.				(h) FGVCaircraft.				(i) SUN397.			
	Base	New	H		Base	New	H		Base	New	H
CLIP	84.50	84.10	84.29	CLIP	20.10	28.10	23.43	CLIP	69.80	73.10	71.41
CoOp	82.80	80.30	81.53	CoOp	34.10	18.96	24.37	CoOp	78.40	62.20	69.36
CoCoOp	86.20	87.00	86.59	CoCoOp	27.13	25.00	26.02	CoCoOp	77.23	74.73	75.96
KAPT	86.13	87.06	86.59	KAPT	29.67	28.73	29.19	KAPT	79.40	74.33	76.78
(j) DTD.				(k) EuroSAT.				(l) UCF101.			
	Base	New	H		Base	New	H		Base	New	H
CLIP	53.90	58.10	55.92	CLIP	43.40	61.40	50.85	CLIP	63.90	71.60	67.53
CoOp	77.13	41.43	53.90	CoOp	92.13	51.83	66.34	CoOp	82.76	52.53	64.27
CoCoOp	73.56	52.70	61.40	CoCoOp	82.33	50.00	62.21	CoCoOp	79.50	66.76	72.57
KAPT	75.97	58.30	65.97	KAPT	84.80	67.57	75.21	KAPT	80.83	67.10	73.33

where $\bar{\mathbf{X}}$ represents average-pooled image features. We then compute the cosine similarity between $\bar{\mathbf{X}}$ and text features \mathbf{W} ,

$$p(y = i | \bar{\mathbf{X}}) = \frac{\exp(\text{sim}(\bar{\mathbf{X}}, \mathbf{w}_i) / \tau)}{\sum_{j=1}^N \exp(\text{sim}(\bar{\mathbf{X}}, \mathbf{w}_j) / \tau)}, \quad (4)$$

where τ is a learned temperature parameter and $\text{sim}(\cdot)$ denotes cosine similarity.

Training Objective. During training, we update the gradients of knowledge-aware prompts and adaptation head while keeping the parameters of CLIP frozen. The training objective is to minimize the cross-entropy loss:

$$\mathcal{L}_{ce} = - \sum_i \log p(y = i | \bar{\mathbf{X}}), 1 \leq i \leq N. \quad (5)$$

4. Experiments

4.1. Experimental Setup

Datasets. For evaluation, we perform extensive experiments on 11 image classification datasets: Flowers102 [26], OxfordPets [27], Food101 [2], StanfordCars [17], FGVCaircraft [23], SUN397 [43], DTD [6], EuroSAT [12], UCF101 [34], Caltech101 [9], and ImageNet [7]. These datasets cover a variety of fine-grained classification tasks, building an all-around benchmark, including species of plants or animals, satellite imagery of traffic, and diverse general objects. Meanwhile, we use Wikidata5m [40], the main source of external knowledge, which is a large-scale knowledge graph dataset with aligned text descriptions from the corresponding Wikipedia pages.

Training Details. We adopt ViT-B/32 as the backbone network for all experiments. For constructing automatic

Table 2: **Model comparison in the one-shot setting.** We fine-tune KAPT and other models mentioned above on 11 datasets with only one sample in each category, where KAPT performs best in the average.

(a) Comparison to CoOp and CoCoOp in one-shot setting.													(b) Model ablations.	
Method	Image Net	Caltech 101	Oxford Pets	Stanford Cars	Flowers 102	Food 101	FGVC Aircraft	SUN 397	DTD	Euro SAT	UCF 101	Avg.	Method	Avg.
CoOp	60.20	91.16	86.43	60.56	71.06	74.73	15.23	63.90	47.66	55.46	66.23	62.96	Baseline	62.96
CoCoOp	64.50	92.43	87.43	60.60	68.30	79.46	10.26	65.66	45.46	48.60	67.56	62.75	w/ knowledge	63.47
KAPT	62.90	89.63	87.60	60.73	74.17	78.07	22.13	64.50	50.93	46.50	65.90	63.91	w/ adaptation	63.20
													KAPT	63.91

discrete prompts, we set the maximum length of discrete prompts and beam search as 20 and 6, respectively. We fix the number of context tokens for automatic continuous prompts to 16. The adaptation head is configured with the dropout rate of 0.1 and 8 attention heads. Throughout training, the initial learning rates for both the knowledge-aware prompts and adaptation head are established at 0.002 and 0.005, respectively, with adjustment governed by the cosine annealing rule. Our optimization approach employs SGD, and we set the maximum epoch count to 50. To ensure fair comparisons with previous works, we average the three scores with different seeds. All experiments are conducted on a single Tesla V100 GPU with 32GB memory, within the PyTorch framework.

Compared Methods. We compare KAPT with existing representative prompting methods based on CLIP. CoOp [49] is the landmark prompt tuning method, which learns the context prompt to make CLIP recall the task-relevant knowledge for downstream image recognition. To overcome the deficiency of learnable prompts on generalization ability, CoCoOp [48] proposes conditional prompts with a network module to improve the generalization on unseen classes. Note that the results of baseline models are obtained by the released official codes.

4.2. Comparison with State-of-the-Art Methods

The performance of KAPT and two baselines are shown in Table 1. KAPT achieves outstanding performance and establishes state-of-the-art results on the overall accuracy (evaluated by harmonic mean). Compared to CoOp, KAPT demonstrates an overall performance improvement across all datasets, with a notable increase of 11.13% particularly on unseen classes. As the generalization to unseen classes is an essential capability of models, CoCoOp proposes conditional prompts to improve the generalizability. Although CoCoOp achieves a better performance than CoOp, KAPT still makes considerable progress compared with CoCoOp in average scores of all metrics. When contrasting KAPT with CoCoOp, improvements of 1.71%, 3.22%, and 2.57% are observed in base classes, new classes, and harmonic mean, respectively. Importantly, KAPT outperforms CoCoOp across 10 out of 11 datasets. Zhou et al. [48]

claim that CLIP is a strong competitor in unseen classes due to learning-based prompts easily overfitting to base classes than manual prompts. Compared with zero-shot CLIP, KAPT achieves an absolute gain of 0.25% on new classes and outperforms CLIP on 7 out of 11 datasets, including ImageNet, DTD, EuroSAT, OxfordPets, Food101, SUN397, and AirCraft101. However, it’s worth noting that CoCoOp’s accuracy on new classes outperforms zero-Shot CLIP only on ImageNet and SUN397.

4.3. One-shot Classification Performance

KAPT performs excellently on the generalization test, demonstrating outstanding overall performance on the base-to-new setting. Meantime, the anti-overfitting ability is also essential for KAPT. Here, we train KAPT and other baseline methods in the one-shot setting to evaluate anti-overfitting ability. Table 2a shows the comparisons of KAPT with other models on 11 datasets. Overall, our KAPT shows its superiority over baseline models on average performance on one-shot settings. It is well known that CoOp trained on closed datasets has strong performance on seen classes in closed datasets. However, our KAPT still beats CoOp on 8 out of 11 datasets under closed datasets and one-shot setting. In Table 2b, the integration of the knowledge-aware prompt and the adaptation head has demonstrated remarkable achievement. The success of this approach can be attributed to the fact that when the downstream task has limited samples, category-related external knowledge is able to make up the lack of visual information to a certain extent and assist the adaptation head in filtering out irrelevant information, which is not associated with the category.

4.4. Distribution Shift Robustness

We systematically evaluated the robustness of KAPT under distribution shifts, *i.e.*, cross-dataset transfer and domain generalization scenarios. In the cross-dataset transfer scenario, we train two models separately using ImageNet and SUN397 datasets. Subsequently, our approach’s performance was assessed across 9 diverse datasets. As evident in Table 3a, when utilizing ImageNet as the source dataset, KAPT exhibits a modest improvement in transferability compared to CoCoOp, achieving an average accuracy of 61.50%, surpassing CoCoOp by 0.25%. Similarly,

Table 3: **Robustness evaluation to distribution shift.** All models are trained with 16-shot samples.

(a) Cross-dataset transfer (source: ImageNet or SUN397).											(b) Domain generalization.		
Method	Caltech	Oxford	Stanford	Flowers	Food	FGVC	DTD	Euro	UCF	Avg.	Source dataset: ImageNet		
	101	Pets	Cars	102	101	Aircraft		SAT	101		Target	CoCoOp	KAPT
Source dataset: ImageNet													
CoCoOp	92.15	88.90	60.30	65.80	80.65	17.50	40.05	41.70	64.20	61.25	INV2	58.40	58.10
KAPT	88.90	89.40	58.15	68.00	79.95	17.95	44.80	41.35	65.05	61.50	IN-S	42.00	42.30
Source dataset: SUN397													
CoCoOp	90.95	80.20	52.00	57.70	76.70	12.15	37.90	43.60	66.35	57.50	IN-A	31.60	31.10
KAPT	90.20	85.45	53.85	61.80	78.10	14.55	42.95	33.30	59.90	57.78	IN-R	66.30	66.60

Table 4: **Ablation study over KAPT components.** Average accuracy (%) over 11 datasets is reported.

Method	Base	New	H
Baseline	79.41	64.02	70.88
w/ knowledge	77.70	67.24	72.09
w/ adaptation head	80.09	65.58	72.11
KAPT	78.41	70.52	74.26

when SUN397 is employed as the source dataset, KAPT showcases slightly better performance relative to CoCoOp, outperforming CoCoOp across 6 out of 9 datasets and yielding an average accuracy gain of 0.28% over CoCoOp.

Furthermore, in order to study the robustness of our method to domain generalization, we evaluate the transferability of the model trained on ImageNet (IN) to various out-of-domain datasets, *i.e.*, ImageNetV2 (INV2) [32], ImageNet-Sketch (IN-S) [37], ImageNet-A (IN-A) [14] and ImageNet-R (IN-R) [13]. By observing Table 3b, it becomes evident that KAPT and CoCoOp yield comparable performance in the domain generalization. We conclude that the proposed knowledge enhances the transferability to new categories, while robust to the domain change on seen categories.

4.5. Ablation Study

We investigate the importance of the critical components of KAPT for its excellent generalization ability through a series of ablation experiments. Initially, we assess the performances of two variants, namely, ablated knowledge-aware prompt or adaptation head, in base-to-new settings to ascertain their necessity. Subsequently, we quantitatively examine the essential role of discrete and learnable continuous prompts in enhancing KAPT’s generalization ability under unseen setting. Lastly, additional experiments offer further hyperparameter analysis.

Effectiveness of Proposed Components. In our framework, knowledge-aware prompts and adaptation head are two core components. To investigate the effectiveness of

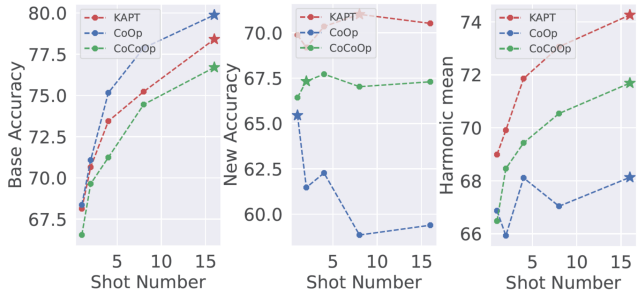


Figure 4: **Performance comparisons with different shot numbers.** KAPT and CoOp/CoCoOp are trained on 11 benchmark datasets with varying shots on the base-to-new generalization setting. KAPT outperforms other models from 1-shot to 16-shot settings.

each component, we conduct ablation experiments to reveal how the combination of two modules improves the overall performance of KAPT, especially on unseen classes. Specifically, we carry out the experiments by adding them one by one to observe the change in overall performance. We take CoOp as the baseline. The experimental results are shown in Table 4. We can find that adding the adaptation head helps to improve the accuracy on base, new and harmonic by a margin of 0.68%, 1.56%, and 1.23% compared with the baseline method. The result demonstrates that the adaptation head not only adopts visual representation towards a specific task but also helps improve the model’s generalizability on unseen classes in the same task. Moreover, we also find that adding knowledge-aware prompts can improve the baseline by 3.22% and 1.21% on new accuracy and harmonic mean. Although the base accuracy of adding knowledge-aware prompts drops below the baseline model, the gains on unseen classes are far outweighed by the losses on seen classes. The result demonstrates that knowledge-aware prompts are able to relieve the weak generalizability problem greatly. Compared to the baseline model, KAPT constructed by knowledge-aware prompts and adaptation head shows an improvement of 6.50% and 3.38% on classification accuracy on unseen classes and harmonic mean.

Table 5: **Ablation study over types of prompts.** We report the accuracy of KAPT (discrete + continuous) and other variant models (with continuous only or with discrete only) across the 11 datasets in the unseen setting.

Method	Image Net	Caltech 101	Oxford Pets	Stanford Cars	Flowers 102	Food 101	FGVC Aircraft	SUN 397	DTD	Euro SAT	UCF 101	Average
discrete + continuous	65.20	93.53	96.53	66.20	71.20	87.06	28.73	74.33	58.30	67.57	67.10	70.52
w/ continuous only	66.06	93.46	96.10	66.96	66.63	87.46	11.30	74.66	51.76	58.23	68.73	67.53
w/ discrete only	64.36	92.93	96.33	63.80	66.93	85.00	23.70	71.10	57.86	64.60	66.33	68.86

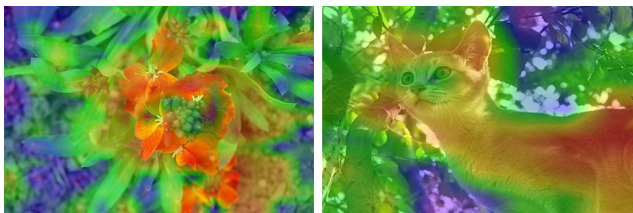


Figure 5: **Attention heatmaps of adaptation head.** Two images are from Flowers102 and OxfordPets, respectively.

Generalizability of Knowledge Prompts. The construction of automatic prompts is an essential part of KAPT. Although we analyze the importance of knowledge-aware prompts in Section 4.5, we still need to explore the independent role and mutual influence of two different automatic prompts. We consider two variant models: i) KAPT (with continuous only) and ii) KAPT (with discrete only). As shown in Table 5, KAPT outperforms two variant models on most datasets and obtains the best average accuracy on unseen classes. This demonstrates that the cooperation of the two kinds of automatic prompts can promote the improvement of the generalization ability of the model. Meanwhile, we discover that the average results of KAPT (with continuous only) and KAPT (with discrete only) outperform KAPT (with adaptation head) in Table 4 by 1.95% and 3.28% on unseen classes, respectively. The above findings prove that any automatic prompts sourced from related-category external knowledge help the adaptation head to remove task-irrelevant concepts to improve its generalization ability. However, we also notice that our KAPT underperforms KAPT (with continuous only) on 5 out of 11 datasets. This is probably because some noise information is still retained in discrete prompts and the qualities of textual descriptions between the different domains in the Wikipedia Encyclopedia have significant differences. Thus, we leave more sophisticated data processing as future work.

Shot Number. Section 4.2 reports the performance of KAPT, CoOp, and CoCoOp in 16-shot setting. Here, we want to investigate further the effect of different shot settings on base-to-new generalization setting. Therefore, we conduct experiments with varying shot values on KAPT, *i.e.*, 1, 2, 4, 8, 16. By examining the outcomes depicted in Figure 4, it becomes evident that the performance of both

Table 6: **Performance comparisons using different vision backbones from CLIP.** KAPT outperforms CoOp and CoCoOp in three vision backbones overall and even achieves better performance in all settings than CoCoOp. Δ denotes absolute gains of KAPT over CoCoOp.

Method	ResNet-50			ViT-B/32			ViT-B/16		
	Base	New	H	Base	New	H	Base	New	H
CoOp	77.54	57.48	66.02	79.88	59.39	68.12	82.83	62.82	71.45
CoCoOp	75.26	64.36	69.39	76.70	67.30	71.69	79.53	71.49	75.29
KAPT	75.39	64.71	69.94	78.41	70.52	74.26	81.10	72.24	76.41
Δ	+0.13	+0.35	+0.55	+1.71	+3.22	+2.57	+1.57	+0.75	+1.12

KAPT and CoCoOp progressively improves with the augmentation of shot numbers in the new setting and overall performance. In contrast, the performance of CoOp exhibits instability with increasing sample numbers. This erratic behavior in CoOp’s results might arise from its tendency to overfit to seen classes, resulting in diminished accuracy on unseen classes. In contrast, KAPT achieves an optimal overall performance by striking a skillful balance between seen and unseen classes.

Visualization of the Adaptation Head. We conducted experiments to analyze the adaptation head’s ability to filter out task-irrelevant information. For this analysis, we utilized images from OxfordPets and Flowers102 datasets as examples. As depicted in Figure 5, the model pays more attention to task-related content, like flowers and cats. Therefore, the adaptation head aligns the visual representation towards the specific task, mitigating the interference from task-irrelated visual concepts.

Backbone Models. CLIP provides a variety of vision backbones, such as ResNet-50, ViT-B/32 and ViT-B/16. We assess the performance of KAPT not only with ViT-B/32 as our backbone but also with ResNet-50 and ViT-B/16. Table 6 presents the averaged performance across 11 datasets using different backbones. As anticipated, KAPT consistently outperforms both CoOp and CoCoOp across various backbone architectures.

Parameter Number. Considering the introduction of trainable parameters through the adaptation head, we empirically investigated their influence on model performance.

Table 7: **Parameter Comparison.** † means CoCoOP with more prompt tokens and larger dimension of Meta-Net.

Method	Parameters	Base	New	H
CoCoOp	0.4M	76.70	67.30	71.69
CoCoOp†	1.5M	76.79	68.62	72.47
KAPT	1.3M	78.41	70.52	74.26

To ensure fair evaluation, we devised CoCoOp† to increase the parameter number of CoCoOp to the same level as our models. As presented in Table 7, the results verify the notable advantages of our method, even when parameters in both approaches are matched in scale.

5. Conclusion

To overcome potential overfitting towards seen classes and underperforming generalizability in unseen scenarios under the same task, we propose a knowledge-aware prompt tuning (KAPT) for vision-language models. Specifically, we identify the importance of exploring category-related external knowledge by designing two types of knowledge-aware prompts for text. Further, we also present an adaptation head to adapt the visual representation toward a specific task. Extensive experiments validate KAPT’s superiority over state-of-the-art approaches on standard benchmark datasets. However, since KAPT builds upon the CLIP backbone, inherent biases and fairness concerns from the original model may persist during prompt learning. While our model exhibits enhanced performance, further refinements are possible. To achieve broader coverage of visual concepts, Wikipedia Encyclopedia and other external knowledge bases could be jointly used as the knowledge source. Meanwhile, existing external knowledge bases are generally diverse in the open domain, often lacking task-specific expertise. Constructing multi-source knowledge bases with specialized expertise remains future investigation.

Acknowledgments. This work was supported by the National Key R&D Program of China (Grant NO. 2022YFF1202903) and the National Natural Science Foundation of China (Grant NO. 62122035).

References

[1] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 3

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 5

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakan-

tan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 3

[4] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web conference 2022*, pages 2778–2788, 2022. 3

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020. 2

[6] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 5

[8] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. *arXiv preprint arXiv:2307.16525*, 2023. 3

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178. IEEE, 2004. 5

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, 2021. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, pages 770–778, 2016. 4

[12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 7

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 7

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom

- Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2
- [16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3
- [17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 5
- [18] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1
- [19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2
- [20] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 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, 2020. 2
- [21] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 1, 3
- [22] Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13507–13515, 2021. 3
- [23] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5
- [24] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 3
- [25] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3, 4
- [26] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 5
- [27] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 5
- [28] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 3
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4
- [30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020. 4
- [31] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. 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, 2022. 3
- [32] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 7
- [33] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. In *NeurIPS*, 2022. 3
- [34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5
- [35] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on Artificial Intelligence*, 2017. 3, 4
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [37] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 7
- [38] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3
- [39] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022. 2
- [40] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A

- unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021. 4, 5
- [41] Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. π -tuning: Transferring multimodal foundation models with optimal multi-task interpolation. In *International Conference on Machine Learning*, pages 37713–37727. PMLR, 2023. 3
- [42] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Motlaghi. Multi-modal answer validation for knowledge-based VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2712–2721, 2022. 3
- [43] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 5
- [44] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *International Conference on Learning Representations*, 2021. 1, 2
- [45] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 3
- [46] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, 2021. 3
- [47] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, 2021. 1, 3
- [48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2, 3, 6
- [49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pages 1–12, 2022. 1, 3, 6