Concept-Guided Prompt Learning for Generalization in Vision-Language Models

Yi Zhang[1,2], **Ce Zhang**[3], **Ke Yu**[2], **Yushun Tang**[2], **Zhihai He**[2,4*]
[1]Harbin Institute of Technology
[2]Southern University of Science and Technology
[3]Carnegie Mellon University
[4]Pengcheng Laboratory
zhangyi2021@mail.sustech.edu.cn, cezhang@cs.cmu.edu
{yuk2020, tangys2022}@mail.sustech.edu.cn, hezh@sustech.edu.cn

Abstract

Contrastive Language-Image Pretraining (CLIP) model has exhibited remarkable efficacy in establishing cross-modal connections between texts and images, yielding impressive performance across a broad spectrum of downstream applications through fine-tuning. However, for generalization tasks, the current fine-tuning methods for CLIP, such as CoOp and CoCoOp, demonstrate relatively low performance on some fine-grained datasets. We recognize the underlying reason is that these previous methods only projected global features into the prompt, neglecting the various visual concepts, such as colors, shapes, and sizes, which are naturally transferable across domains and play a crucial role in generalization tasks. To address this issue, in this work, we propose **C**oncept-Guided **P**rompt **L**earning (CPL) for vision-language models. Specifically, we leverage the well-learned knowledge of CLIP to create a visual concept cache to enable concept-guided prompting. In order to refine the text features, we further develop a projector that transforms multi-level visual features into text features. We observe that this concept-guided prompt learning approach is able to achieve enhanced consistency between visual and linguistic modalities. Extensive experimental results demonstrate that our CPL method significantly improves generalization capabilities compared to the current state-of-the-art methods.

Introduction

Recent studies in pre-trained Vision-Language Models (VLMs), such as CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021), highlight a promising direction for foundation models in performing a variety of open-vocabulary tasks. By understanding various visual concepts learned from extensive image-text pairs, these models exhibit impressive capabilities across a broad spectrum of downstream tasks in a zero/few-shot manner (Radford et al. 2021; Alayrac et al. 2022; Yu et al. 2022).

Although the zero-shot CLIP model demonstrates competitive performance in various visual tasks, its nature as a pre-trained model hinders its ability to generalize to unseen domains. Therefore, several works focus on fine-tuning these pre-trained VLMs for downstream tasks through designing learnable prompts derived from training instances.

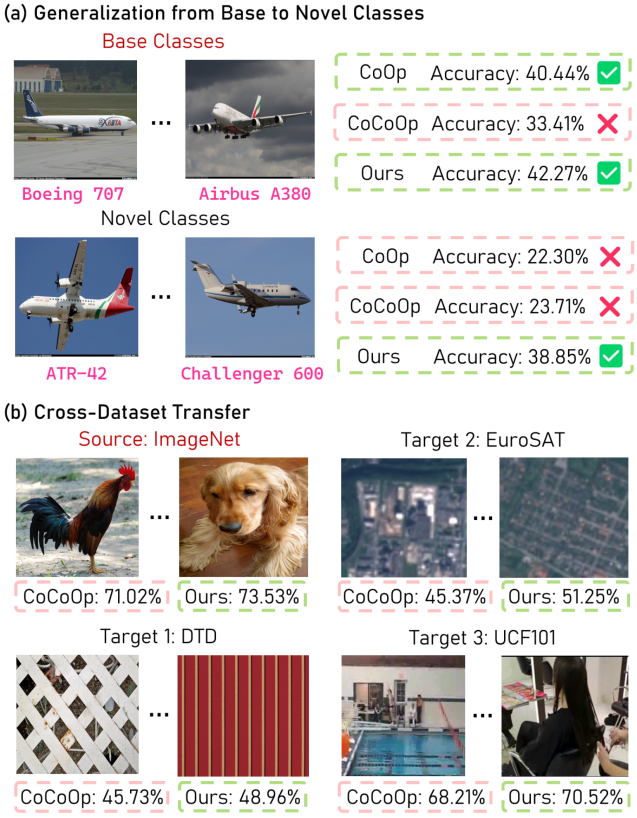

Figure 1: Examples and performance comparisons on base-to-novel generalization and cross-dataset transfer tasks. Our proposed CPL exhibits remarkable generalization capabilities in comparison to other state-of-the-art methods.

For example, CoOp (Zhou et al. 2022b) firstly introduces learnable prompts to distill task-relevant knowledge; CoCoOp (Zhou et al. 2022a) suggests adjusting the prompt based on each individual image; and TaskRes (Yu et al. 2023) proposes to incorporate a prior-independent task residual that doesn't undermine the well-learned knowledge of CLIP. For clarification, we provide an overview of each of the aforementioned methods in Figure 2.

For the generalization tasks as shown in Figure 1, we ob-

*Corresponding author.

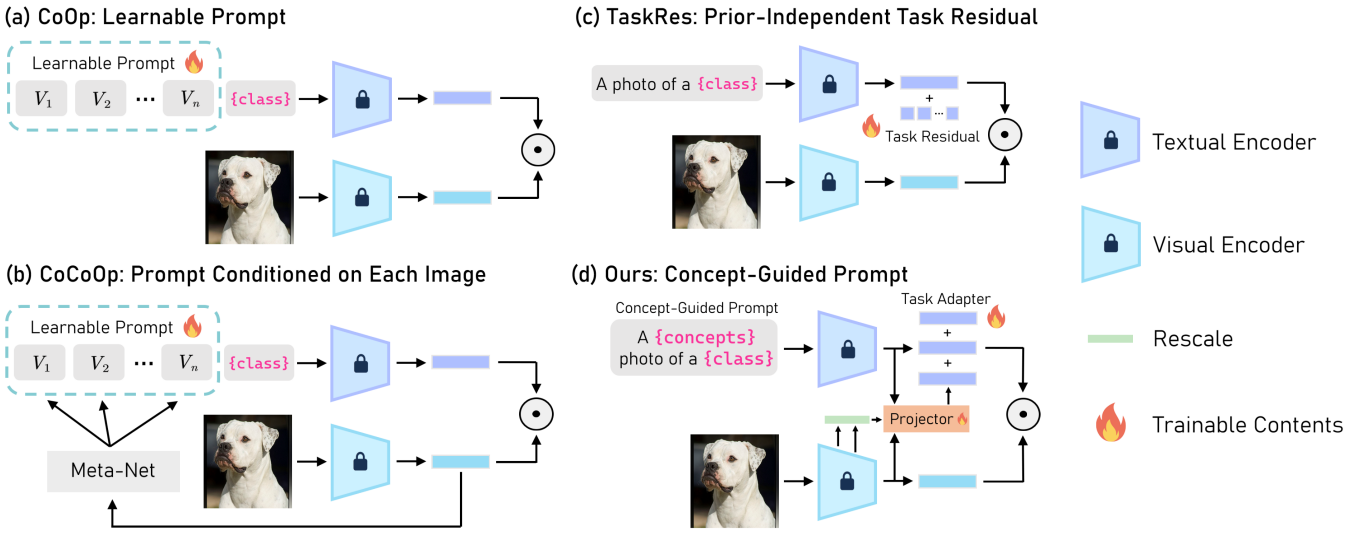

Figure 2: An illustration comparing our proposed proposed CPL approach with related baselines. We include CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a) and TaskRes (Yu et al. 2023) for comparison.

serve that the current fine-tuning methods for CLIP, such as CoOp and CoCoOp, demonstrate relatively low performance on some difficult fine-grained datasets such as DTD (texture recognition), FGVC Aircraft (fine-grained classification), EuroSAT (satellite image recognition), and UCF101 (action recognition). We recognize that this issue may arise from CoOp and CoCoOp's direct tuning of the input text prompts to the text encoder, which can potentially undermine the previously well-learned knowledge of VLMs. To address this issue, TaskRes attempts to incorporate prior-independent learnable contexts to preserve this knowledge. To further explore the potential of prompt tuning methods, we ask: *is it possible to utilize the prior knowledge of VLMs during the fine-tuning process without destroying it?*

We also observe that the previous fine-tuning methods only considered adapting to a specific task using supervised loss, which is not fully effective in generalizing to unseen domains. This limitation stems from the fact that they primarily consider class-specific features and overlook low-level visual concepts, such as colors, shapes, and materials. However, these low-level concepts are naturally transferable across domains and are therefore essential for enabling vision-language models to generalize. As a result, we are prompted to ask: *is it possible to incorporate visual concepts into the fine-tuning process for VLMs to enhance their transfer capabilities?*

To address the problems above, in this work, we propose Concept-Guided Prompt Learning (CPL) for vision-language models. Specifically, we leverage the well-learned knowledge of CLIP to create a visual concept cache to enable concept-guided prompting. In order to refine the text features, we further develop a projector that projects multi-level visual features into text features. We observe that this concept-guided prompt learning approach is able to achieve enhanced consistency between visual and linguistic modalities, leading to improved generalization capability. We conducted extensive experiments to evaluate the proposed CPL approach on base-to-novel generalization, cross-dataset transfer, and domain generalization tasks. Our comprehensive empirical results demonstrate the significantly superior performance of CPL compared to existing state-of-the-art methods.

Related Work

Vision-Language Models

In recent years, vision-language models have attracted significant attention from researchers, emerging as a novel paradigm for performing visual tasks. Specifically, large-scale VLMs have been utilized to acquire general visual representations guided by natural language supervision (Radford et al. 2021). Current studies highlight that these models, pre-trained on vast image-text pairs available online, are capable of understanding both the semantics of images paired with their respective textual descriptions (Radford et al. 2021; Yu et al. 2022). Recent studies (Zhang et al. 2021; Zhou et al. 2022b) have showcased that with a robust comprehension of open-vocabulary concepts, VLMs are able to tackle various downstream visual tasks, including image retrieval (Duan et al. 2022), depth estimation (Hu et al. 2023), visual grounding (Li et al. 2022), visual question answering (Duan et al. 2022).

Fine-Tuning VLMs

Fine-tuning is crucial in adapting VLMs to downstream tasks (Duan et al. 2022). Among various fine-tuning methods for VLMs, two primary approaches stand out: prompt tuning methods and adapter-based methods, respectively.

Prompt Tuning Methods. Prompt tuning methods transform prompts into continuous vector representations for

end-to-end objective function optimization, distilling task-relevant information from prior knowledge of VLMs (Zhou et al. 2022a,b). As the foundational work in this field, CoOp (Zhou et al. 2022b) optimizes the prompt context by a continuous set of learnable vectors. Further, CoCoOp (Zhou et al. 2022a) recognizes the generalization issue not addressed by CoOp and proposes to generate prompts on each individual image. MaPLe (Khattak et al. 2023) tunes both vision and language branches via a vision-language coupling function in order to induce cross-modal synergy.

Adapter-Based Methods. Another series of works directly transforms the features extracted by encoders of CLIP to perform adaptation to downstream tasks. These methods are referred to as adapter-based methods. For example, CLIP-Adapter (Gao et al. 2023), one of the pioneering works, leverages an extra feature adapter to enhance traditional fine-tuning outcomes. Following CLIP-Adapter, Tip-Adapter (Zhang et al. 2022) introduces a training-free approach by constructing a key-value cache model based on few-shot samples. CCLI (Zhang et al. 2023b) proposes to enable concept-level image representation to perform downstream tasks. BDC-Adapter (Zhang et al. 2023a) enhances vision-language reasoning by providing a more robust metric for measuring similarity between features. In addition, Zhu et al. (2023b) proposes APE, which harnesses the prior knowledge of VLMs by a prior cache model, and explores the trilateral relationships among test images, the prior cache model, and textual representations.

Visual Concept Learning

Existing literature has suggested two major approaches to visual concept learning. The first approach typically uses hand-crafted semantic concept annotations (*e.g.*, colors, textures, and fabric) for the training images (Patterson and Hays 2012, 2016; Pham et al. 2021), which is labor-intensive in practice. To address this issue, researchers propose the second approach, which aims at designing data-driven concepts through unsupervised learning (Fei-Fei and Perona 2005; Liu, Kuipers, and Savarese 2011; Huang, Loy, and Tang 2016). While these acquired concepts might initially appear sensible, they can often carry inherent biases, ultimately constraining their overall performance. Empowered by CLIP (Radford et al. 2021), in this work, we design an unsupervised concept mining-and-cache technique that is capable of discovering a large set of visual concepts with semantics corresponding to pre-defined text concepts.

Method

Background

CLIP. CLIP (Radford et al. 2021) stands out as a foundational model that constructs an shared embedding space through the fusion of visual and semantic understanding. This architecture is composed of two encoders: a visual encoder denoted as E_v responsible for handling image input x, and a text encoder referred to as E_t designed to process the corresponding textual prompt t_c built as "`a photo of` $[CLS]_c$", where $[CLS]_c$ represents the word embedding for the class c. During training, CLIP learns to optimize the resemblance between the image feature and the prompt embeddings associated with the true label.

CoOp and CoCoOp. CoOp (Zhou et al. 2022b) replaces manual prompt construction by introducing an alternate method that involves learned prompts. This method utilizes a collection of n adaptable context vectors $\{[V_1], [V_2], \cdots, [V_n]\}$, each having the same dimension as word embeddings. These vectors are iteratively updated through gradient descent. For a specific class c, the respective prompt can be represented as $t_c = \{[V_1], [V_2], \cdots, [V_n], [CLS]_c\}$. CoCoOp (Zhou et al. 2022a) integrates visual features into prompt learning by utilizing a meta-network $h_\theta(x)$ that generates a meta-token π, denoted as $\pi = h_\theta(x)$. The meta-token, combined with the context vectors, form the textual prompts $t_c = \{[V_1(x)], [V_2(x)], \cdots, [V_n(x)], [CLS]_c\}$, where $V_n(x) = V_n + \pi$ represents the n^{th} text token.

Concept-Guided Prompt Learning

Overview. In Figure 3, we present an overview of our proposed CPL method. Figure 3 (a) shows the visual concept cache establishing process. We first construct a list of text concepts Ψ_t that describe major visual concepts. Then we leverage CLIP's robust text-image correlation capability to discover the image feature v_j with the highest similarity score for each text concept feature $c_t^i \in C_t$. These "matched" features are stored in the visual concepts cache as keys, with their corresponding text concepts $\psi_i \in \Psi_t$ as values. Figure 3 (b) shows the concept-guided discovery process: we first extract the image feature v by E_v, then use the image feature as the query to find Top-K similar keys using cosine distance, and finally we utilize the corresponding values to generate a concept-guided prompt. Figure 3 (c) presents the training pipeline for CPL. We first extract the visual features for a given image x using the visual encoder, then we can obtain the concatenated outputs of different layers $\hat{E}(x)$ as the multi-level features. Next, we follow (b) to generate the concept-guided prompt and extract text features by E_t. These features are used as input for the projector, which is a transformer decoder for mapping multi-level visual features into the textual feature space, providing the multi-level visual context. Combined with the multi-level visual context and a task adapter, refined text features work as a classifier for final prediction.

Visual Concept Cache. In Figure 3 (a), following Zhang et al. (2023b), we start by constructing a comprehensive list Ψ_t comprising I text concepts that describe major visual concepts. This list Ψ_t incorporates 2000 common text descriptions for visual concepts gathered from established visual concept datasets (Zhao et al. 2019; Pham et al. 2021). The descriptions encompass words representing materials, colors, shapes, *etc.* Illustrations of these terms can be found in Figure 4. The dictionary is represented as $\Psi_t \triangleq \{\psi_i\}_{i=1}^I$. Adhering to CLIP's zero-shot setup, we begin by appending ψ_i to a manually designed prompt $\phi =$ "`The photo is ...`" to form a concept-specific textual input $\{\phi; \psi_i\}$. Conse-

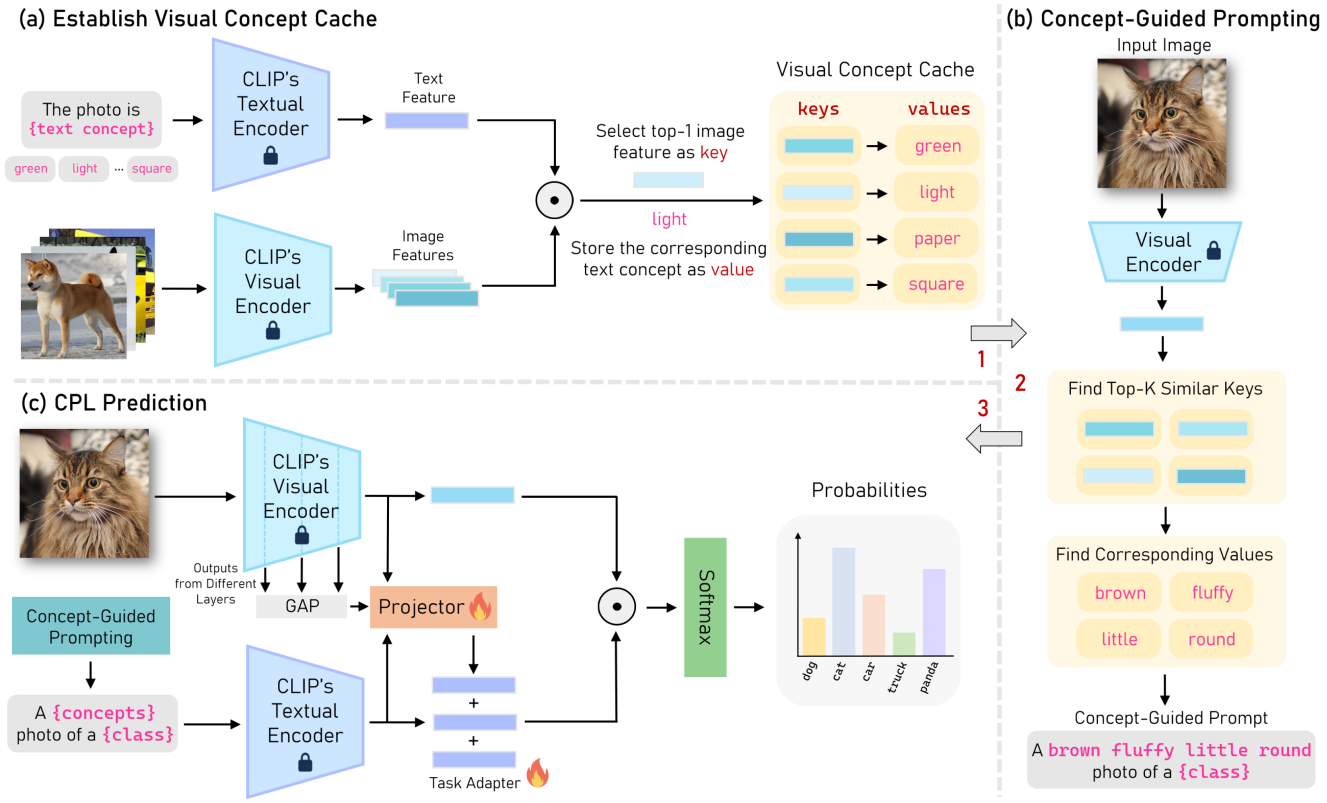

Figure 3: An overview of our proposed Concept-Guided Prompt Learning (CPL) method. Subfigure (a) shows the visual concept cache-establishing process. Subfigure (b) shows the concept-guided prompt discovery process. Subfigure (c) presents the training pipeline of our proposed CPL, where the projector and task adapter are learnable.

quently, utilizing the text encoder E_t, we generate text concept features $C_t \triangleq \{c_t^i\}_{i=1}^{I}$, denoted as $c_t^i = E_t(\phi; \psi_i)$.

Within CPL, the visual concepts are discovered by leveraging the text concept features C_t and the CLIP model, derived from the training images. In the scenario of N-shot D-class few-shot learning, where there exist N labeled images within each of the D classes, the training set is denoted as $T_r \triangleq \{x_j\}_{j=1}^{ND}$. Utilizing the CLIP visual encoder E_v, we generate their respective image features $V \triangleq \{v_j\}_{j=1}^{ND}$, expressed as $v_j = E_v(x_j)$. For every text concept feature $c_t \in C_t$, the similarity score S_t is calculated against all visual features in V using the formula $S_t = \text{sim}(c_t, v_j) = c_t v_j$, where both c_t and v_j are normalized. Subsequently, we identify the image feature with the highest similarity score as the key and its corresponding text concept word ψ as the associated value, stored within the visual concept cache.

Projector for Vision-to-Language Prompting. Incorporating depictions of rich visual semantics can enhance the precision of the textual content. Multi-level visual features provide richer visual semantics than only high-level (class-specific) features. Therefore, we explore how to utilize multi-level features to optimize the text features. Obviously, we can use a projector to transform multi-level features into the space of text features. Transformer decoder (Vaswani et al. 2017; Rao et al. 2022; Lu et al. 2021) can model the interactions between vision and language by adopting cross-attention mechanism. Hence, we use the Transformer decoder as the projector. Several studies (Lin et al. 2017; Wang et al. 2021) have already demonstrated that in deep neural networks, the features generated by the earlier layers differ from those produced by the subsequent layers in level. Typically, the earlier layers yield low-level features, such as edges and colors, whereas the later layers produce high-level features, referred to as class-specific features.

Inspired by Singha et al. (2023), our aim is to incorporate the multi-level visual features from E_v into the projector $\mathbf{P}$. To achieve this, we utilize global average pooling to reduce the spatial dimensions of individual channels. This produces $\hat{E}_v^q(x) \in \mathbb{R}^{C\times 1}$, where $E_v^q \in \mathbb{R}^{W\times H\times C}$ signifies the output derived from the q^{th} layer, where W, H, C are the width, height, and number of channels of the feature maps. Incorporating this method, we formulate $\hat{E}(x)$ as the concatenation of multi-level features acquired from all Q encoder layers within E_v, denoted as $[\hat{E}_v^1(x); \cdots; \hat{E}_v^Q(x)]$. We subsequently pass $\hat{E}(x)$ and $\mathbf{f_t}$ through the projector $\mathbf{P}$, which generates multi-level visual context $\mathbf{f_{tv}}$:

$$\mathbf{f_{tv}} = \mathbf{P}(\mathbf{f_t}, \hat{E}(x)), \tag{1}$$

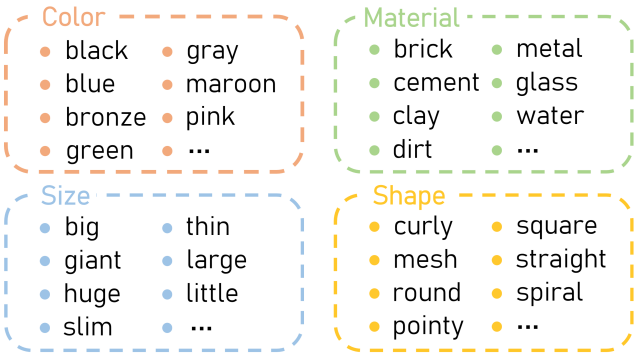

Figure 4: Example text concepts collected from existing visual attribute datasets. Here we present several instances of terms that illustrate `color`, `material`, `size`, and `shape` within our dictionary of text concepts.

where $\mathbf{f_{tv}}$ is the extracted visual contexts, and $\mathbf{f_t}$ is the text feature generated by the CLIP text encoder. This implementation fosters the exploration of text features to identify the most relevant visual cues.

Task Adapter. As illustrated in Figure 3, we append a learnable matrix (*i.e.* task adapter), denoted by A, to text features $\mathbf{f_t}$ generated by the text encoder E_t. A is task-specific and updated by gradient descent during the training process. In this way, it performs directly on the text-based classifier and explicitly decouples the inherent knowledge of the pre-trained models and the new knowledge for a target task. Therefore, we can preserve the prior knowledge of concept-guided prompts and assimilate knowledge from new tasks, improving the adaptability of the proposed model.

CPL Training and Inference

CPL Training. In the training phase, we utilize the supervised contrastive loss, represented as $\mathcal{L}_{ce}$, as the loss function for our approach. This cross-entropy loss guarantees an appropriate alignment between visual and textual feature representations. Given an image x, we first generate its visual features by $E_v(x)$, denoted as $\mathbf{f_v}$, we follow the concept-guided prompt discovery process to find the concept-guided prompt, denoted as P_c, then we can obtain the text features by $E_t(P_c)$, denoted as $\mathbf{f_t}$. According to Equation (1), we get $\mathbf{f_{tv}}$. Finally, we can calculate the refined text features $\widetilde{\mathbf{f_t}}$ by,

$$\widetilde{\mathbf{f_t}} = \mathbf{f_t} + \boldsymbol{\alpha}\mathbf{f_{tv}} + \boldsymbol{\beta}A, \tag{2}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are learnable parameters to control the scaling of the residual and text features are updated through a residual connection. The values assigned to parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ upon initialization are exceptionally diminutive (*e.g.*, 10^{-4}). This choice aims to uphold the language priors extensively within the original text features. The prediction probability for x pertaining to label i is represented as

$$p(y=i|x) = \frac{\exp(\text{sim}(\widetilde{\mathbf{f_t^i}}, \mathbf{f_v})/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(\widetilde{\mathbf{f_t^j}}, \mathbf{f_v})/\tau))}, \tag{3}$$

where τ is a temperature coefficient and 'sim' represents the cosine similarity. The cross-entropy loss is calculated by

$$\mathcal{L}_{ce} = -\arg\min_{\theta_{\mathbf{P}},\, A} \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}_{tr}} \sum_{k=1}^{\mathcal{Y}_{tr}} y_k \log(p(y_k|x)), \tag{4}$$

where $\theta_{\mathbf{P}}$ is the parameter weights of the projector $\mathbf{P}$, A is the learnable matrix of the task adapter, and $\mathcal{Y}_{tr}$ are the class labels for the training dataset.

CPL Inference. During the inference phase for the test dataset $\mathcal{D}_{te}$, where $\mathcal{Y}_{te}$ signifies the labels of this dataset, we calculate the cosine similarity between the images x_{te} and the prompt embeddings for all classes within the test dataset $\mathcal{Y}_{te}$. The class exhibiting a higher probability value is subsequently chosen:

$$\hat{y}_{te} = \arg\max_{y\in\mathcal{Y}_{te}} p(y|x_{te}). \tag{5}$$

Experiments

Benchmark Settings

Task Settings. We follow previous work to evaluate our proposed approach on four challenging task settings:

- **Generalization from Base to Novel Classes.** We evaluate the generalization capability of our method in a zero-shot scenario by dividing the datasets into base and novel classes. We train our model with few-shot images on the base classes and then evaluate on unseen novel classes.
- **Cross-Dataset Transfer.** We conduct a direct evaluation of our ImageNet-trained model across various other datasets. Following previous work, we train our model on all 1,000 ImageNet classes in a few-shot setting.
- **Domain Generalization.** We evaluate the robustness of our method on OOD datasets. Similarly, we evaluate our model trained on ImageNet directly on four ImageNet variants that encompass different types of shifts.

Datasets. For base-to-novel generalization, cross-dataset transfer tasks, we follow previous work (Radford et al. 2021; Zhou et al. 2022b,a) to conduct the experiments on 11 representative image classification datasets, including ImageNet (Deng et al. 2009) and Caltech101 (Fei-Fei, Fergus, and Perona 2004) for generic object classification; OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Van Gool 2014), and FGVCAircraft (Maji et al. 2013) for fine-grained classification; SUN397 (Xiao et al. 2010) for scene recognition; UCF101 (Soomro, Zamir, and Shah 2012) for action recognition; DTD (Cimpoi et al. 2014) for texture classification; and EuroSAT (Helber et al. 2019) for satellite image recognition. For domain generalization, we utilize ImageNet as the source dataset and four ImageNet variants as target datasets including ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), ImageNet-V2 (Recht et al. 2019), ImageNet-Sketch (Wang et al. 2019).

(a) **Average over 11 datasets.**

Method	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
ProGrad	82.48	70.75	76.16
KgCoOp	80.73	73.60	77.00
Ours	**84.38**	**78.03**	**81.08**
	+1.69	+2.89	+2.53

(b) ImageNet.

Method	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
ProGrad	77.02	66.66	71.46
KgCoOp	75.83	69.96	72.78
Ours	**78.74**	**72.03**	**75.24**
	+1.72	+1.49	+1.77

(c) Caltech101.

Method	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
ProGrad	98.02	93.89	95.91
KgCoOp	97.72	94.39	96.03
Ours	**98.35**	**95.13**	**96.71**
	+0.33	+0.74	+0.68

(d) OxfordPets.

Method	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
ProGrad	95.07	97.63	96.33
KgCoOp	94.65	97.76	96.18
Ours	**95.86**	**98.21**	**97.02**
	+0.43	+0.45	+0.44

(e) StanfordCars.

Method	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
ProGrad	77.68	68.63	72.88
KgCoOp	71.76	75.04	73.36
Ours	**79.31**	**76.65**	**77.96**
	+1.19	+1.61	+4.49

(f) Flowers102.

Method	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
ProGrad	95.54	71.87	82.03
KgCoOp	95.00	74.73	83.65
Ours	**98.07**	**80.43**	**88.38**
	+0.47	+2.63	+4.73

(g) Food101.

Method	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
MaPLe	90.71	92.05	91.38
ProGrad	90.37	89.59	89.98
KgCoOp	90.05	91.70	91.09
Ours	**91.92**	**93.87**	**92.88**
	+1.21	+1.82	+1.50

(h) FGVCAircraft.

Method	Base	Novel	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	36.50
ProGrad	40.54	27.57	32.82
KgCoOp	36.21	33.55	34.83
Ours	**42.27**	**38.85**	**40.49**
	+1.73	+2.56	+3.99

(i) DTD.

Method	Base	Novel	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
MaPLe	80.36	59.18	68.16
ProGrad	77.35	52.35	62.45
KgCoOp	77.55	54.99	64.35
Ours	**80.92**	**62.27**	**70.38**
	+0.56	+2.37	+2.22

(j) SUN397.

Method	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
MaPLe	80.82	78.70	79.75
ProGrad	81.26	74.17	77.55
KgCoOp	80.29	76.53	78.36
Ours	**81.88**	**79.65**	**80.75**
	+0.62	+0.95	+1.00

(k) EuroSAT.

Method	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
MaPLe	94.07	73.23	82.35
ProGrad	90.11	60.89	72.67
KgCoOp	85.64	64.34	73.48
Ours	**94.18**	**81.05**	**87.12**
	+0.11	+7.82	+4.77

(l) UCF101.

Method	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
MaPLe	83.00	78.66	80.77
ProGrad	84.33	74.94	79.35
KgCoOp	82.89	76.67	79.65
Ours	**86.73**	**80.17**	**83.32**
	+2.04	+1.51	+2.55

Table 1: Comparison with state-of-the-art methods on base-to-novel generalization (on ViT-B/16 backbone). Our proposed method learns local concepts and demonstrates strong generalization results over existing methods on 11 recognition datasets. The best results are in bold and the second-best results are underlined.

Implementation Details. For a fair comparison, we use the ViT-B/16 CLIP model for base-to-novel generalization and cross-dataset transfer and the ResNet-50 CLIP model for domain generalization. Throughout the training process, both the visual and textual encoders remain fixed. We adhere to the data pre-processing protocol outlined in CLIP, which involves resizing and applying random cropping operations, *etc.*. We conduct training for 70 epochs on the ImageNet and 50 epochs for other datasets. We designate the number of concepts K as 10. Training involves a batch size of 256 and an initial learning rate set at 10^{-3}. We employ the AdamW optimizer with a cosine annealing scheduler and train the models on a single NVIDIA RTX 3090 GPU. Code will be available at https://github.com/rambo-coder/CPL.

Generalization from Base to Novel Classes

We compare our method with six baselines: zero-shot CLIP (Radford et al. 2021), CoOp (Zhou et al. 2022b), CoCoOp (Zhou et al. 2022a), ProGrad (Zhu et al. 2023a), MaPLe (Khattak et al. 2023), and KgCoOp (Yao, Zhang, and Xu 2023). Table 1 displays results regarding base-to-novel generalization across 11 datasets with 16-shot samples.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
Ours	**73.53**	**95.52**	**91.64**	**66.17**	**73.35**	**87.68**	**27.36**	**68.24**	**48.96**	**51.25**	**70.52**	**68.07**

Table 2: Comparison of our method with existing approaches on cross-dataset evaluation. Overall, our method demonstrates superior generalization capabilities with the highest average accuracy on 10 datasets.

Method	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
CLIP	60.33	53.27	35.44	21.65	56.00
CoOp	63.33	55.40	34.67	23.06	56.60
CoCoOp	62.81	55.72	34.48	23.32	57.74
ProGrad	62.17	54.70	34.40	23.05	56.77
PLOT	63.01	55.11	33.00	21.86	55.61
DeFo	64.00	58.41	33.18	21.68	55.84
TPT	60.74	54.70	35.09	26.67	59.11
Ours	**66.92**	**58.67**	**37.64**	**31.05**	**60.08**

Table 3: Comparison with other methods on robustness (%) to natural distribution shifts. The best results are in bold and the second-best results are underlined.

Performance Evaluation on Base Classes. CoOp demonstrates remarkable performance on base classes among previous methods. However, it exhibits an overfitting problem for its excessive dependence on a single learnable prompt component, as argued by CoCoOp. Our method remarkably surpasses CoOp by an average accuracy gain of 1.69% without a generalizability depletion, and achieves the best performance on base classes for all datasets, as illustrated in Table 1. Our method's efficacy indicates its substantial capability to adapt to downstream tasks.

Generalization to Unseen Classes. Although CoCoOp improves CoOp's limited generalizability by conditioning prompts on image instances, it has an average degradation of -2.22% on base classes. As a comparison, MaPLe obtains balanced performance on both base and novel classes. Remarkably, our CPL method achieves the highest performance in terms of novel classes and harmonic mean (**HM**) on all 11 datasets, with an accuracy improvement of 2.89% and 2.53%, respectively. With visual concepts extracted from prior knowledge, our CPL can better generalize to novel categories. The exceptional performance demonstrates the enhanced generalizability of CPL to unseen classes without sacrificing performance in base classes.

Method	1	2	4	8	16
CLIP	60.33	60.33	60.33	60.33	60.33
+ CGP	61.06	61.59	62.65	63.17	64.38
+ CGP + P	62.32	62.88	63.80	64.83	66.35
+ CGP + P + TA	**63.02**	**63.37**	**64.36**	**65.31**	**66.92**

Table 4: Effectiveness of different components in our method. CGP and P represent concept-guided prompting and the projector, respectively, and TA is the task adapter.

Cross-Dataset Transfer

Cross-dataset transfer is a much more challenging generalization task compared to base-to-novel generalization, since the latter only transfers within a single dataset while the former transfers across different datasets, *e.g.*, from object recognition to scene classification. We test the cross-dataset generalization ability of our method on 1000 ImageNet classes and then transfer it directly to the remaining 10 datasets. The comparison results with CoOp, CoCoOp, and MaPLe are presented in Table 2. Overall, our CPL method marks the best performance on both source and target datasets with a target average of 68.07%, and outperforms MaPLe by 1.77%. Notably, our method surpasses MaPLe by 3.2% on EuroSAT, a satellite image dataset whose fundamentals are distinctive from ImageNet. This suggests that concept-guided prompting in our method facilitates better generalization, as illustrated in Figure 1.

Domain Generalization

In Table 3, we provide the classification accuracy across the source domain and target domains, as well as the average accuracy within target domains (OOD Average). In addition to the methods mentioned earlier, we also compare our approach with PLOT (Chen et al. 2023), DeFo (Wang et al. 2023), TPT (Shu et al. 2022). Our approach surpasses other methods in all scenarios, indicating the remarkable robustness of our model against distribution shifts.

Ablation Studies

Contributions of major algorithm components. From Table 4, we can see that all three components contribute significantly to the enhanced performance. Among them,

Value of K	6	8	**10**	12	14
Accuracy	65.33	66.28	**66.92**	66.56	66.31
Value of I	1000	1500	**2000**	2500	3000
Accuracy	63.67	65.88	**66.92**	66.71	66.23

Table 5: Number K of concepts selected and total size I of concept set. Experiments are conducted on 16-shot ImageNet. Here for I, we fix the number of concept categories and vary the number of concepts in each category.

Method	Epochs	Time	Accuracy	Gain
Zero-shot CLIP	0	0	60.33	0
Linear Probe CLIP	-	13m	56.13	-4.20
CoOp	200	14h 40m	62.26	+1.93
ProGrad	200	17h	63.45	+3.12
Ours	70	50min	**66.92**	**+6.59**

Table 6: Comparison on the number of training epochs and time on 16-shot ImageNet.

concept-guided prompting brings the largest performance improvement, for example, a 4.05% improvement in 16-shot accuracy. This shows that a more accurate and specific text description leads to better classification results.

The number K of concepts selected and the size I of text concepts set. We investigate the impact of K by varying the number of concepts selected and show the results in Table 5. We find that our method achieves the best performance when $K = 10$. The results also show that our method achieves the best performance when $I = 2000$. When the size is too large, the performance decreases since the different text concepts might match the same visual concept.

Comparison on the number of training epochs and time. As shown in Table 6, our proposed CPL outperforms other methods by a large margin with only 50 minutes, while CoOp and ProGrad need more than 14 hours. This demonstrates the remarkable efficiency of our method.

Conclusion

In this work, we introduce Concept-Guided Prompt Learning (CPL) for vision-language models. By utilizing the profound knowledge embedded in CLIP, we form a visual concept cache that facilitates concept-guided prompting. To further refine text features, we design a projector that projects multi-level visual features into corresponding textual features. Our proposed CPL method exhibits great effectiveness in diverse applications such as base-to-novel generalization, cross-dataset transfer, and domain generalization tasks. Supported by thorough experimental analysis, we demonstrate that our proposed CPL achieves remarkable performance improvements, and also surpasses existing leading-edge methods by substantial margins.

References

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. In *Advances in Neural Information Processing Systems*, volume 35, 23716–23736.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, 446–461.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2023. PLOT: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Duan, J.; Chen, L.; Tran, S.; Yang, J.; Xu, Y.; Zeng, B.; and Chilimbi, T. 2022. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15651–15660.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 178.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 2, 524–531.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2023. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 1–15.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. 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.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271.

Hu, X.; Zhang, C.; Zhang, Y.; Hai, B.; Yu, K.; and He, Z. 2023. Learning to adapt CLIP for few-shot monocular depth estimation. *arXiv preprint arXiv:2311.01034*.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Unsupervised learning of discriminative attributes and visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5175–5184.

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*, 4904–4916.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 554–561.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Liu, J.; Kuipers, B.; and Savarese, S. 2011. Recognizing human actions by attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3337–3344.

Lu, Z.; He, S.; Zhu, X.; Zhang, L.; Song, Y.-Z.; and Xiang, T. 2021. Simpler is better: Few-shot semantic segmentation with classifier weight transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8741–8750.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3498–3505.

Patterson, G.; and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2751–2758.

Patterson, G.; and Hays, J. 2016. Coco attributes: Attributes for people, animals, and objects. In *European Conference on Computer Vision*, 85–100.

Pham, K.; Kafle, K.; Lin, Z.; Ding, Z.; Cohen, S.; Tran, Q.; and Shrivastava, A. 2021. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13018–13028.

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*, 8748–8763. PMLR.

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

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, 14274–14289.

Singha, M.; Jha, A.; Solanki, B.; Bose, S.; and Banerjee, B. 2023. APPLeNet: Visual attention parameterized promptlearning for few-Shot remote sensing image generalization using clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024–2034.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 6000–6010.

Wang, F.; Li, M.; Lin, X.; Lv, H.; Schwing, A.; and Ji, H. 2023. Learning to decompose visual features with latent textual prompts. In *International Conference on Learning Representations*.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 10506–10518.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3485–3492.

Yao, H.; Zhang, R.; and Xu, C. 2023. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6757–6767.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*.

Yu, T.; Lu, Z.; Jin, X.; Chen, Z.; and Wang, X. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10899–10909.

Zhang, R.; Qiu, L.; Zhang, W.; and Zeng, Z. 2021. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*.

Zhang, R.; Zhang, W.; Fang, R.; Gao, P.; Li, K.; Dai, J.; Qiao, Y.; and Li, H. 2022. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, 493–510. Springer.

Zhang, Y.; Zhang, C.; Liao, Z.; Tang, Y.; and He, Z. 2023a. BDC-Adapter: Brownian distance covariance for better vision-language reasoning. In *British Machine Vision Conference*.

Zhang, Y.; Zhang, C.; Tang, Y.; and He, Z. 2023b. Cross-Modal Concept Learning and Inference for Vision-Language Models. *arXiv preprint arXiv:2307.15460*.

Zhao, B.; Fu, Y.; Liang, R.; Wu, J.; Wang, Y.; and Wang, Y. 2019. A large-scale attribute dataset for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 398–407.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023a. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15659–15669.

Zhu, X.; Zhang, R.; He, B.; Zhou, A.; Wang, D.; Zhao, B.; and Gao, P. 2023b. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2605–2615.