Conceptual Codebook Learning for Vision-Language Models

Yi Zhang[1,2], Ke Yu[3], Siqi Wu[4], and Zhihai He[2]*

[1] Harbin Institute of Technology
[2] Southern University of Science and Technology
zhangyi2021@mail.sustech.edu.cn hezh@sustech.edu.cn
[3] University of California San Diego
key022@ucsd.edu
[4] University of Missouri
siqiwu@missouri.edu

Abstract. In this paper, we propose Conceptual Codebook Learning (CoCoLe), a novel fine-tuning method for vision-language models (VLMs). CoCoLe aims to address the challenge of enhancing the generalization capability of VLMs while adapting them to downstream tasks in a few-shot setting. We recognize that visual concepts like shapes, colors, and textures are inherently transferable across different domains and are essential for generalization tasks. Motivated by this critical finding, we learn a conceptual codebook consisting of visual concepts as keys and conceptual prompts as values, which serves as a link between the image encoder's outputs and the text encoder's inputs. Specifically, for a given image, we leverage the codebook to identify the most relevant conceptual prompts associated with the class embeddings to perform the classification. Additionally, we incorporate a handcrafted concept cache as a regularization to alleviate the overfitting issues in low-shot scenarios. This conceptual codebook learning method has been shown to improve the alignment between visual and linguistic modalities. Extensive experimental results demonstrate that our CoCoLe method remarkably outperforms the existing state-of-the-art methods across various evaluation settings, including base-to-new generalization, cross-dataset evaluation, and domain generalization tasks. Detailed ablation studies further confirm the efficacy of each component in CoCoLe.

Keywords: Vision-Language · Generalization · Concept Learning

1 Introduction

Large-scale pre-trained Vision-Language Models (VLMs), *e.g.*, ALIGN [15] and CLIP [26], have achieved exceptional zero-shot performance in various downstream tasks. These VLMs, trained on massive datasets of image-text pairs with

* Corresponding author

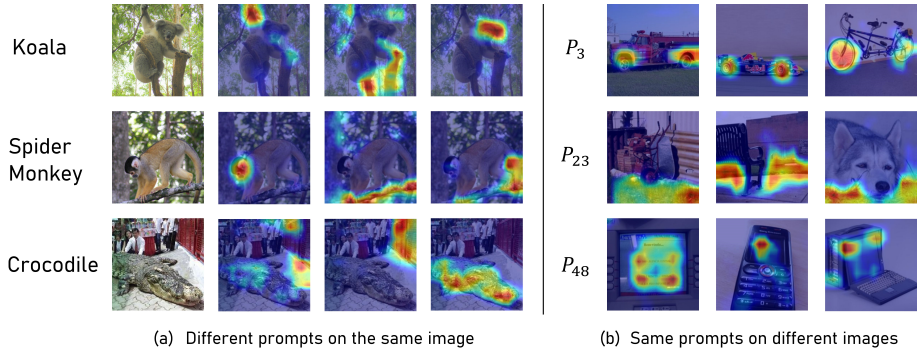


Fig. 1: (a) Visualization of the chosen prompts of the same image. (b) Visualization of the same prompts on different images. Grad-CAM [30] is used for the visualization.

contrastive optimization objectives, effectively align and embed different modalities into a shared vector space. Despite their impressive performance, adapting these models to diverse downstream tasks remains challenging due to their substantial size. As a result, recent research has concentrated on improving the ability of pre-trained VLMs to adapt to downstream tasks by fine-tuning supplementary parameters, while keeping VLMs frozen. Prompt-tuning methods, *e.g.* CoOp [42] and ProGrad [43], replace manual prompts with learnable ones to obtain task-specific knowledge, while adapter-based methods utilize extra modules directly on the top of VLMs, such as Clip-adapter [9] and Tip-adapter [38]. These methods have made significant advancements with limited labeled data.

However, we find that existing fine-tuning methods for CLIP, including CoOp [42] and CPL [40], exhibit relatively low performance on fine-grained datasets like FGVCAircraft [21] (aircraft classification), and UCF101 [31] (action classification). To address the challenge of enhancing the generalization capability of VLMs in a few-shot settings, in this paper, we propose a novel fine-tuning method called Conceptual Codebook Learning (CoCoLe). Our idea stems from the observation that visual concepts are naturally transferable across domains. As illustrated in Fig. 1, within a single image, there exist multiple distinct visual concepts focusing on different regions. For example, the selected prompts highlight the claws, ears of the koala, and the branches where the koala stands. Moreover, there are similar concepts in images from different classes; for example, the "firetruck", "racer", and "bicycle" classes possess the compound concept of "wheel" in common.

Motivated by this interesting finding, we propose to learn a conceptual codebook consisting of visual concepts as keys and conceptual prompts as values, which serve as a link between the image encoder's outputs and the text encoder's inputs. Specifically, for a given image, we leverage the codebook to identify the most relevant conceptual prompts associated with the class embeddings to perform the classification. Additionally, we incorporate a handcrafted concept cache as a regularization to alleviate the overfitting issues in low-shot scenarios.

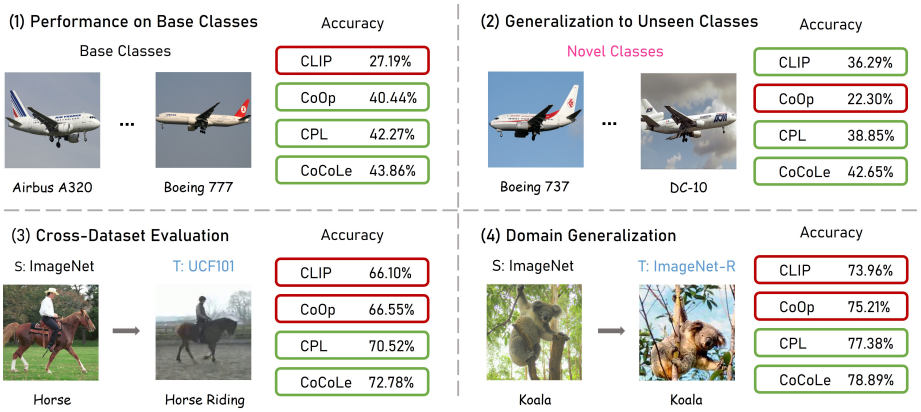


Fig. 2: Illustrations and accuracy comparisons on base-to-novel generalization, cross-dataset transfer and domain generalization tasks. S and T represent source and target datasets respectively.

As shown in Fig. 2, we observe that this conceptual codebook learning method can achieve enhanced alignment between visual and linguistic modalities. Our contributions could be summarized as:

- We proposed a novel fine-tuning method named CoCoLe for VLMs to solve the problem of performance degradation on generalization tasks.
- CoCoLe introduces a conceptual codebook to adaptively learn visual concepts and their corresponding conceptual prompts with regularization to further guarantee the generalization capability.
- Extensive experimental results demonstrate the outstanding performance of CoCoLe compared to existing state-of-the-art methods in base-to-novel generalization, domain generalization and cross-dataset transfer tasks.

2 Related Work

2.1 Vision-Language Models

Pre-trained vision-language models (VLMs) have recently emerged as a notable trend [5,19,26,29]. Capitalizing on tremendous image-text data, these large-scale models can effectively acquire visual representations using contrastive loss, enabling them to grasp both visual and textual semantics and achieve successful modality alignment. Current studies [37, 42] have showcased that by harnessing extensive sets of image-text pairs, VLMs exhibit outstanding performance across a range of downstream visual tasks [6,13,16]. For example, Derived through contrastive learning on 400 million online image-text pairs, CLIP [26] demonstrated remarkable zero-shot accuracy on classification tasks. Our method aims to utilize the comprehensive capability of CLIP to perform knowledge-guide fine-tuning for better adaptation to downstream tasks.

2.2 Prompt Tuning for VLMs

As text input for pre-trained vision-language models, prompts function as the guidance for the downstream tasks, extracting task-specific information from the existing knowledge within VLMs [41, 42]. Setting a precedent in this field, CoOp [42] exploits a set of learnable vectors to perform end-to-end optimization on the prompt context but fails to generalize to unseen classes. To address this issue, CoCoOp [41] improved CoOp's generalization by generating conditional prompts. Further, KgCoOp [36] enhances the generalization by minimizing the discrepancy between learned and handcrafted prompts, and CoPrompt [28] constrains the trainable models by pre-trained ones to avoid overfitting on the downstream task. Meanwhile, there are methods exploring diverse forms of prompts. MaPLe [17] enhances both vision and language components by employing a coupling function to promote cross-modal synergy, whereas CPL [40] leverages the powerful generalization of CLIP to build a visual concept cache with a projector to capture multi-level visual features.

In our work, we mainly focus on prompt-tuning ways and meticulously manipulate learnable vectors by a learnable codebook with the regularization of a handcrafted concept cache. Among existing methods, the most related to ours are CoOp and CPL. Compared with CoOp, the proposed CoCoLe introduces an adaptive codebook rather than fixed to specific classes or tasks. On the other hand, CoCoLe leverages the transferability of concepts across domains, with optimal handcrafted concept-based prompts as a regularization to prevent the codebook from overfitting.

2.3 Visual Concept Learning

Earlier studies have identified two primary methods for visual concept learning. The first approach generally involves using manual annotations of concepts (e.g., textures, fabrics, and colors) for the training images [24, 25], while the other method utilizes unsupervised learning to design data-driven concepts [8, 14, 20]. However, these approaches may introduce biases into the learned concepts, limiting their overall performance. Recent studies have sought complementary prompting methods to capture unbiased visual concepts [33, 34]. Notably, CPL [40] first utilizes the capabilities of CLIP [26] to design an unsupervised concept cache. Furthermore, in this work, we introduce a learnable codebook supervised by a handcrafted concept cache, which automatically selects conceptual prompts that are aligned with visual concepts.

3 Method

3.1 Background and Overview

CLIP and CoOp. CLIP [26] consists of two main encoders: the visual encoder E_v for processing image inputs X_v, and the text encoder E_t for handling textual prompts T_c, which are formatted as "`a photo of` $[\text{CLS}]_c$", where $[\text{CLS}]_c$ is the

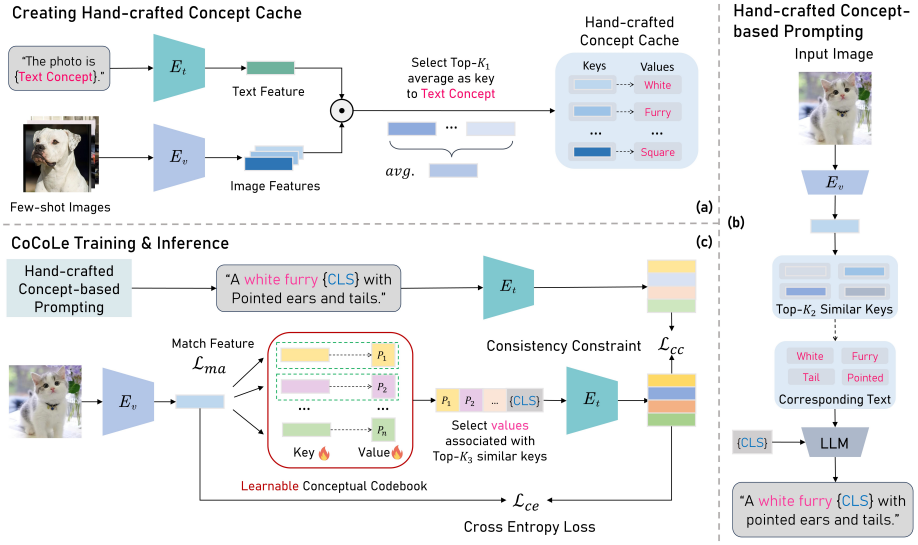


Fig. 3: An overview of the proposed CoCoLe. (a) shows the establishing process of handcrafted concept cache. (b) displays the handcrafted concept-based prompting process. (c) presents the training pipeline for CoCoLe. Within CoCoLe, only the keys and values in the Conceptual Codebook are learnable.

word embedding for class c. CLIP optimizes the similarity between the image features and the prompt embeddings that correspond to the true labels during training. CoOp [42] takes this a step further by replacing manually constructed prompts with learnable ones. It employs a set of n adaptable context vectors $[V_1], [V_2], \cdots, [V_n]$, each with the same dimension as word embeddings. Gradient descent is utilized to optimize the learnable vectors. For a given class c, the corresponding prompt is formulated as $T_c = [V_1], [V_2], \cdots, [V_n], [\text{CLS}]_c$.

Overview of CoCoLe. In Fig. 3, we illustrate an overview of our proposed CoCoLe approach. Figure 3 (a) presents the process of constructing handcrafted concept cache. We begin by compiling a list of text concepts Ω_t that encapsulates key visual concepts. Next, we utilize CLIP's powerful image-text association capability to identify the image feature v_j that has the Top-K_1 highest similarity scores for each text concept feature $c_t^i \in C_t$. The highest scoring Top-K_1 features are averaged and saved as keys in the visual concepts cache, paired with their respective text concepts $\omega_i \in \Omega_t$ as values. In Figure 3 (b), we illustrate the handcrafted concept-based prompting process: initially, we use E_v to obtain the image feature v. This image feature is then used as a query to identify the Top-K_2 similar keys based on cosine distance. Subsequently, the associated values, combined with the class embeddings, serve as input for LLM (*e.g.*, GPT [2]) to generate the optimal handcrafted concept-based prompts represented as $\mathcal{P}^h \triangleq \{P_{h_i}\}_{i=1}^{N^C}$, where N^C denotes the number of class.

Figure 3 (c) shows the CoCoLe training pipeline. We begin by using the visual encoder to obtain the visual features f_v from a given image x. We then follow (b) to generate the handcrafted concept-based prompts $\mathcal{P}^h$ and obtain text features by E_t, denoted as $\mathcal{F}_h = E_t(\mathcal{P}^h)$. The visual concepts (keys) and the conceptual prompts (values) in the conceptual codebook are trainable parameters optimized by four loss functions. The classification loss $\mathcal{L}_{ce}$ is employed to maximize the alignment between the image feature f_v and the related text features f_t. We utilize the loss function $\mathcal{L}_{ma}$ to minimize the distance between the chosen keys (Top-K_3 similar visual concepts) and the image feature f_v, facilitating the learning of generalizable concepts by the keys. $\mathcal{L}_{cc}$ works as a regularization for diminishing the overfitting problem, ensuring the text features produced by the selected learned prompts do not deviate significantly from those generated by the handcrafted concept-based prompts. Finally, $\mathcal{L}_{or}$ ensures that the text features of the prompts are orthogonal to enhance the prompts' diversity.

3.2 Conceptual Codebook Learning (CoCoLe)

Learnable Conceptual Codebook. In the CoOp framework, each class embedding is associated with a single set of prompt vectors. Nevertheless, images belonging to the same class often encompass a variety of concepts. Conflating these varied concepts into a single set of prompts can result in significant knowledge loss. Moreover, the encoded information within CoOp's prompts lacks interclass interaction, since concepts from one class may assist in identifying another class with similar concepts. For instance, when presented with an image of a cat in the tree, the concept of "in the tree" might also apply to images of other animals (e.g., a koala in the tree). We hypothesize that fine-tuning prompts based on image concepts can facilitate the learning of textual descriptions associated with these concepts, thereby improving generalization across datasets.

As such, we propose CoCoLe, as depicted in Fig. 3. The key insight of CoCoLe is a trainable concept codebook, empowering the image to autonomously determine the prompts it should learn based on its inherent concepts. For each training input, only a subset of prompts that align with the current image concepts are chosen and trained individually. The learnable concept codebook stores visual concepts as keys and conceptual prompts as values, comprising N (key, value) pairs, denoted as $\Psi_{cc} \triangleq \{(\mathbf{V}_i, \mathbf{P}_i)\}_{i=1}^{N}$, where Ψ_{cc} denotes the learnable concept codebook, each $\mathbf{V}_i \in \mathbb{R}^D$ shares the same dimensionality as the image feature f_v. Additionally, each $\mathbf{P}_i = [\mathbf{p}_i]_1 \dots [\mathbf{p}_i]_M \in \mathbb{R}^{D \times M}$ consists of M learnable vectors. We represent the set of learnable visual concepts as $\mathcal{V} = \{\mathbf{V}_i\}_{i=1}^{N}$ and the entire set of learnable conceptual prompts as $\mathcal{P} = \{\mathbf{P}_i\}_{i=1}^{N}$. In an optimal scenario, we anticipate that the image itself should determine the prompts to be selected, guided by the concepts it encompasses, in order to steer the prediction process. To achieve this, for an input image $\mathbf{x}_j$, we first extract its image feature $f_{v_j} = E_v(\mathbf{x}_j)$, where j represents image index. Then we calculate the cosine similarity score between f_{v_j} and $\mathbf{V}_i \in \mathcal{V}$, denoted as, $S_c = \frac{f_{v_j} \cdot \mathbf{V}_i}{||f_{v_j}||\ ||\mathbf{V}_i||}$.

Next, we choose the keys with Top-K_3 cosine similarity score to form set $\mathcal{V}_j$, representing the subset of Top-K_3 visual concepts chosen from $\mathcal{V}$ uniquely for the j-th image. Then, we select the conceptual prompts that match these visual concepts, represented as $\mathcal{P}_j = \{\mathbf{P}_{j_i}\}_{i=1}^{K_3}$, where $\mathbf{P}_{j_i}$ denotes the i-th prompt chosen uniquely for $\mathbf{x}_j$. We use these prompts to link to the class name embedding of $\mathbf{x}_j$ as shown in Fig. 3, and the input for text encoder can be represented as, $\mathbf{T}(\mathcal{P}_j) = \text{concat}(\mathbf{P}_{j_1}; \ldots; \mathbf{P}_{j_{K_3}}; [\text{CLS}]_d)$, where $\text{concat}(\cdot)$ signifies concatenation. Thus, for a test image $\mathbf{x}_j$ and prompts $\mathcal{P}_j$ based on the concepts of $\mathbf{x}_j$, the text feature f_{t_j} can be obtained by $f_{t_j} \triangleq E_t(\mathbf{T}(\mathcal{P}_j))$. The likelihood of predicting the image as class y_i is ultimately determined by:

$$p(y_i|\mathbf{x}_j) = \frac{e^{\langle f_{v_j}, f_{t_j} \rangle / \tau}}{\sum_{d=1}^{D} e^{\langle f_{v_j}, f_{t_d} \rangle / \tau}}. \tag{1}$$

From a broader viewpoint, the suggested adaptable concept codebook serves as a link connecting the outcomes of the image encoder and the inputs of the text encoder. The keys are fine-tuned to closely align with the identified image features, which hold abundant high-level information, such as image concepts. Meanwhile, the prompts are refined to encompass textual details associated with the respective image concepts, facilitating improved guidance for the model predictions alongside the class name embeddings.

Handcrafted Concept Cache. In Figure 3 (a), inspired by [39], we construct a comprehensive list Ω_t containing $I = 2000$ descriptive text concepts sourced from established visual concept datasets [39, 40]. These descriptions encompass terms related to texture, colors, brightness, density, etc., categorized into 50 classes. Examples of these terms are depicted in Fig. 4. The dictionary is defined as $\Omega_t \triangleq \{\omega_i\}_{i=1}^{I}$. Following CLIP's zero-shot setup, we start by appending each ω_i to a manually crafted prompt $\phi =$ "`The photo is ...`" to create a concept-specific textual input $\{\pi; \omega_i\}$. Subsequently, using the text encoder E_t, we derive text concept features $C_t \triangleq \{c_t^i\}_{i=1}^{I}$, where each $c_t^i = E_t(\pi; \omega_i)$.

We denote the handcrafted concept cache as $\Phi_{mc} \triangleq \{(key, value)_i\}_{i=1}^{I}$. The key and value are the visual concepts and textual concept words respectively. The visual concepts are identified by exploiting the text concept features C_t extracted from the CLIP model, using training images as a basis. In the context of H-shot D-class few-shot learning, there are H labeled images available for each of the D classes. Using the CLIP visual encoder E_v, we obtain their corresponding image features $V \triangleq \{v_j\}_{j=1}^{HD}$, where each $v_j = E_v(x_j)$. Each text concept feature $c_t \in C_t$ is matched against all visual features in V using similarity score formula $S_t = \text{sim}(c_t, v_j) = c_t v_j$, where both c_t and v_j are normalized. Afterwards, we select the Top-K_1 image features with the highest similarity scores, computing their average as the key, and associating it with the corresponding text concept word ω_t. These pairs are stored within the handcrafted concept cache.

Conceptual Codebook Learning with Regularization. Figure 3 (b) presents the handcrafted concept-based prompting process. Initially, we extract the image feature f_v using E_v. Subsequently, we employ this image feature as the



Fig. 4: Examples of text concepts from established visual concept datasets, including descriptive terms of texture, color, transparency, motion and brightness.

query to retrieve the Top-K_2 most similar keys based on cosine similarity. Ultimately, we obtain the corresponding values (conceptual words). Together with the class name, these concept words are input to an LLM (eg. GPT [2]) to generate optimal handcrafted concept-based prompts. Therefore, our approach addresses the challenge of diminished generalization on downstream tasks by introducing a regularization, ensuring the text features produced by the selected learned prompts do not differ significantly from their counterpart generated by the handcrafted concept-based prompts. We enforce this consistency by utilizing the Euclidean distance as a constraint between the text features generated from the hand-crafted concept-based prompts ($f_{t_d}^h$) and those obtained from selected learned prompts ($f_{t_d}^l$). While alternative measures such as cosine distance could also serve as constraints, our empirical findings suggest that Euclidean distance yields superior performance. We can represent the consistency constraint as:

$$\mathcal{L}_{cc} = \frac{1}{D}\sum_{d=1}^{D} ||f_{t_d}^l - f_{t_d}^h||_2^2, \tag{2}$$

where $||\cdot||$ is the euclidean distance, D is the number of seen classes.

3.3 Training and Inference

Training Objective. According to Equation 1, the loss for image classification is expressed as:

$$\mathcal{L}_{ce} = \mathbb{E}[-\log \frac{e^{\langle f_{v_j}, f_{t_j} \rangle / \tau}}{\sum_{d=1}^{D} e^{\langle f_{v_j}, f_{t_d} \rangle / \tau}}]. \tag{3}$$

Besides $\mathcal{L}_{ce}$, we require a matching loss that brings the matched top-K_3 keys $\mathcal{K}_j$ closer to the image embedding $\mathbf{z}_j$, facilitating the keys to learn diverse concepts from the samples. We employ cosine distance as our matching loss. Nevertheless, alternative metrics such as Euclidean distance can also serve as viable matching loss. Through empirical observation, we find that cosine distance typically delivers optimal performance as a matching loss. Therefore, The matching loss adopted to optimize the keys is defined as:

$$\mathcal{L}_{ma} = \sum_{i=1}^{C}(1 - \frac{f_{v_j} \cdot \mathbf{V}_{j_i}}{||f_{v_j}||\ ||\mathbf{V}_{j_i}||}). \tag{4}$$

Lastly, to enhance the semantic diversity of the learned prompts, we introduce an additional loss that orthogonalizes the embeddings of different prompts, thereby boosting prompt diversity.

$$\mathcal{L}_{or} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} |\langle E_t(\mathbf{P}_i), E_t(\mathbf{P}_j) \rangle|, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity. In this way, the overall optimization objective is defined as:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{ma} + \mathcal{L}_{or} + \mathcal{L}_{cc}, \tag{6}$$

The keys in the conceptual codebook are optimized by $\mathcal{L}_{ma}$, while the prompts are optimized through $\mathcal{L}_{ce}$, $\mathcal{L}_{or}$ and $\mathcal{L}_{cc}$.

Inference. Once visual and textual concepts (keys and values) are learned through training, they can be shipped with CLIP for downstream tasks with a standard zero-shot CLIP inference setup. As shown in Figure 3, we first generate the visual features f_{vt} of the test image using visual Encoder E_v, then we use f_{vt} to choose top-K_3 similar visual concepts(keys) by cosine similarity, next, we concatenate the corresponding prompts(values) of the keys to obtain the conceptual prompt, which is fused with each given name to produce conceptual prompt text features $\{f_{tt}^i\}_{i=1}^{C}$. Finally, zero-shot inference is performed with the conceptual prompted text features and the input image feature f_{vt} to produce classification scores on the test images.

4 Experiments

4.1 Experimental Setup

Benchmark Settings. We follow previous works to extensively evaluate our proposed method on three challenging tasks:

- **Base to Novel Class generalization.** We evaluate the generalization capability of our method in zero-shot scenarios within a dataset. The dataset is evenly divided into base and novel classes. We train our model with few-shot images on the base classes and then evaluate on both base classes and unseen novel classes.
- **Cross-Dataset Evaluation.** The 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 texture classification. In this experiment, we follow previous works to train our model in a few-shot setting on 1000 ImageNet classes and subsequently evaluate its performance on ten other unseen datasets.
- **Domain Generalization.** We assess the performance of our model on out-of-distribution generalization. Likewise, we evaluate our model trained on ImageNet directly on four variants of ImageNet, each containing the same classes but from different distributions.

Table 1: Comparison with existing methods on base-to-novel generalization. The top accuracies are highlighted in bold, with the second best results underlined. The harmonic mean are denoted by HM.

(a) Average

	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
KgCoOp	80.73	73.60	77.00
MaPLe	82.28	75.14	78.55
CoPrompt	84.00	77.23	80.48
CPL	84.38	78.03	81.08
Ours	**85.22**	**80.31**	**82.70**
	+0.84	+2.28	+1.62

(b) ImageNet

	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
Co-CoOp	75.98	70.43	73.10
KgCoOp	75.83	69.96	72.78
MaPLe	76.66	70.54	73.47
CoPrompt	77.67	71.27	74.33
CPL	78.74	72.03	75.24
Ours	**79.25**	**74.58**	**76.84**
	+0.51	+2.55	+1.60

(c) Caltech101

	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
Co-CoOp	97.96	93.81	95.84
KgCoOp	97.72	94.39	96.03
MaPLe	97.74	94.36	96.02
CoPrompt	98.27	94.90	96.55
CPL	**98.35**	95.13	96.71
Ours	98.17	**95.67**	**96.90**
	-0.18	+0.54	+0.19

(d) OxfordPets

	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
Co-CoOp	95.20	97.69	96.43
KgCoOp	94.65	97.76	96.18
MaPLe	95.43	97.76	96.58
CoPrompt	95.67	98.10	96.87
CPL	95.86	98.21	97.02
Ours	**96.21**	**98.55**	**97.37**
	+0.35	+0.34	+0.35

(e) StanfordCars

	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
Co-CoOp	70.49	73.59	72.01
KgCoOp	71.76	75.04	73.36
MaPLe	72.94	74.00	73.47
CoPrompt	76.97	74.40	75.66
CPL	79.31	76.65	77.96
Ours	**80.32**	**78.84**	**79.57**
	+1.01	+2.19	+1.61

(f) Flowers102

	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
Co-CoOp	94.87	71.75	81.71
KgCoOp	95.00	74.73	83.65
MaPLe	95.92	72.46	82.56
CoPrompt	97.27	76.60	85.71
CPL	**98.07**	80.43	88.38
Ours	97.72	**81.04**	**88.60**
	-0.35	+0.61	+0.22

(g) Food101

	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
Co-CoOp	90.70	91.29	90.99
KgCoOp	90.50	91.70	91.09
MaPLe	90.71	92.05	91.38
CoPrompt	90.73	92.07	91.40
CPL	91.92	93.87	92.88
Ours	**92.23**	**94.28**	**93.24**
	+0.31	+0.41	+0.36

(h) FGVCAircraft

	Base	Novel	HM
CLIP	27.19	36.29	31.09
CoOp	40.44	22.30	28.75
Co-CoOp	33.41	23.71	27.74
KgCoOp	36.21	33.55	34.83
MaPLe	37.44	35.61	36.50
CoPrompt	40.20	39.33	39.76
CPL	42.27	38.85	40.49
Ours	**43.86**	**42.65**	**43.25**
	+1.59	+3.32	+2.76

(i) SUN397

	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
Co-CoOp	79.74	76.86	78.27
KgCoOp	80.29	76.53	78.36
MaPLe	80.82	78.70	79.75
CoPrompt	82.63	80.03	81.31
CPL	81.88	79.65	80.75
Ours	**83.97**	**82.24**	**83.10**
	+1.34	+2.21	+1.79

(j) DTD

	Base	Novel	HM
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
Co-CoOp	77.01	56.00	64.85
KgCoOp	77.55	54.99	64.35
MaPLe	80.36	59.18	68.16
CoPrompt	**83.13**	64.73	72.79
CPL	80.92	62.27	70.38
Ours	82.46	**68.38**	**74.76**
	-0.67	+3.65	+1.97

(k) EuroSAT

	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
Co-CoOp	87.49	60.04	71.21
KgCoOp	85.64	64.34	73.48
MaPLe	94.07	73.23	82.35
CoPrompt	94.60	78.57	85.84
CPL	94.18	81.05	87.12
Ours	**95.03**	**84.17**	**89.27**
	+0.43	+3.12	+2.15

(l) UCF101

	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	84.69	56.05	67.46
Co-CoOp	82.33	73.45	77.64
KgCoOp	82.89	76.67	79.65
MaPLe	83.00	78.66	80.77
CoPrompt	86.90	79.57	83.07
CPL	86.73	80.17	83.32
Ours	**88.30**	**83.05**	**85.60**
	+1.40	+2.88	+2.28

Datasets. For conducting experiments on base-to-novel generalization and cross-dataset transfer tasks, we adhere to the setting of prior studies [26,41,42]. Specifically, we evaluate our approach across 11 diverse image classification datasets. These datasets encompass a wide range of tasks, including generic object classification (e.g., ImageNet [4] and Caltech101 [7]), fine-grained classification (e.g., OxfordPets [23], StanfordCars [18], Flowers102 [22], Food101 [1], and FGVCAircraft [21]), scene recognition (SUN397 [35]), action recognition (UCF101 [31]), texture classification (DTD [3]), and satellite image recognition (EuroSAT [10]). For the domain generalization task, we employ ImageNet as the source dataset and evaluate our method's performance on four ImageNet variants: ImageNet-A [12], ImageNet-R [11], ImageNet-V2 [27], and ImageNet-Sketch [32].

Implementation Details. To ensure a fair comparison, we employ the ViT-B/16 CLIP model across all three benchmark tasks. For base-to-novel gener-

alization, we train our proposed CoCoLe with 16-shot images on base classes and subsequently evaluate it both base classes and novel classes. For domain generalization and cross-dataset evaluation, we utilize the model trained with 16-shot ImageNet and test it on each target dataset. Throughout the training, we keep both the visual and textual encoders fixed. Our data preprocessing follows CLIP's protocol, including resizing and random cropping operations, among others. For the base-to-novel generation task, we conduct training for 30 epochs on ImageNet and 20 epochs on the other datasets. K_1 and K_2 are set to 3 and 10 respectively. We set prompt length M to 8, the size N of conceptual codebook to 100, and the number of selected concepts K_3 to 4. The training is performed with a batch size of 8 with an initial learning rate of 10^{-3}. We utilize the AdamW optimizer alongside a cosine annealing scheduler.

4.2 Base-to-Novel Generalization

In this section, We compare our CoCoLe method with seven baselines: zero-shot CLIP [26], CoOp [42], CoCoOp [41], MaPLe [17], KgCoOp [36], CoPrompt [28] and CPL [40]. Table 1 shows the experimental results for base-to-novel generalization across 11 datasets using 16-shot samples. We have bolded the top results and indicated improvements over the second-best performance in blue. As we can see from Table 1a, the average of all 11 datasets shows that our method outperforms all the baselines by a large margin for both base and novel classes. In comparison to CoOp and CoCoOp, which are the pioneering prompt learning methods, the performance gain of our method even reached +11% and +6.9% respectively. Our method outperforms the previous state-of-the-art (CPL) by +2.28% on novel classes and +1.62% on the harmonic mean (HM). These results demonstrate the strong zero-shot generalization capability of our proposed method. Also, our method outperforms CPL on base classes by +0.84%, which shows a strong few-shot learning capability.

For the performance of individual datasets, our method outperforms CPL on all the datasets for novel class and HM. For the base classes, our method achieves superior performance gains compared to CPL on 8 out of 11 datasets. Even for Catech101, Flower102, and DTD, where there is a slight performance drop, it remains marginal. This highlights the enhanced generalization capability of our method towards novel classes without compromising performance on base classes. Notably, aside from CPL, our method outperforms all other methods by a significant margin across all datasets. Compared to the second-best performing baseline, our method surpasses it by up to +3.32%, +3.65%, and +2.55% on FGVCAircraft, DTD, and ImageNet, respectively. These observations indicate that our method can effectively learn diverse and discriminative visual and textual concepts, thereby enhancing CLIP's adaptation for generalization tasks.

4.3 Cross-Dataset Evaluation

Table 2 displays the comparison results with CoOp, CocoOp, MaPLe, CoPrompt, and CPL. Our CoCoLe approach achieves the top performance on both source

Table 2: Comparison with state-of-the-art methods on cross-dataset evaluation.

	Source	**Target**										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [42]	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 [41]	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 [17]	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
CoPrompt [28]	70.80	94.50	90.73	65.67	72.30	86.43	24.00	67.57	47.07	**51.90**	69.73	67.00
CPL [40]	<u>73.53</u>	<u>95.52</u>	<u>91.64</u>	<u>66.17</u>	<u>73.35</u>	<u>87.68</u>	<u>27.36</u>	<u>68.24</u>	<u>48.96</u>	51.25	<u>70.52</u>	<u>68.07</u>
Ours	**73.88**	**95.88**	**91.93**	**67.79**	**74.17**	**87.97**	**28.83**	**68.75**	**49.26**	<u>51.75</u>	**72.78**	**68.91**

Table 3: Comparison with existing methods on domain generalization task. The top results are highlighted in bold with the second-best results underlined.

Method	Source	Target				
	ImageNet	-V2	-Sketch	-A	-R	Ave.
CLIP [26]	66.73	60.83	46.15	47.77	73.96	57.17
CoOp [42]	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp [41]	71.02	64.07	48.75	50.63	76.18	59.90
KgCoOp [36]	71.20	64.10	48.97	50.69	76.70	60.11
MaPLe [17]	70.72	64.07	49.15	50.90	76.98	60.27
CoPrompt [28]	70.80	64.25	49.43	50.50	<u>77.51</u>	60.42
CPL [40]	<u>73.53</u>	<u>65.18</u>	<u>49.92</u>	<u>50.73</u>	77.38	<u>60.80</u>
Ours	**73.88**	**65.86**	**50.89**	**51.75**	**78.89**	**61.85**

and target datasets, averaging 68.91% on the target set, surpassing CPL by 0.84%. Notably, we observe the largest improvement of 2.26% over CPL on UCF101, an action image dataset with distinct characteristics from ImageNet. This underscores that our method's conceptual codebook learning enhances generalization significantly.

4.4 Domain Generalization

We present the classification accuracy for both the source domain and target domains in Table 3, along with the average accuracy across the target domains. Our approach consistently outperforms all baselines on both source and target datasets, setting a new state-of-the-art average accuracy of 61.85% for this domain generalization task. Notably, our method beats CPL [40] by +1.51% on ImageNet-R. This highlights the exceptional robustness of our model against distribution shifts.

4.5 Ablation Studies

To rigorously assess our proposed approach, we conduct an empirical analysis of our design decisions and demonstrate the impact of various components in this

Table 4: The ablation study on each component of CoCoLe.

$\mathcal{L}_{ma}$	$\mathcal{L}_{cc}$	$\mathcal{L}_{or}$	Accuracy(HM)
✓	✓	✓	76.84
✓	✓	✗	76.38
✓	✗	✓	74.86
✗	✓	✓	74.45
✓	✗	✗	74.57
✗	✓	✗	74.26
✗	✗	✗	72.12

Table 5: Ablation study of M, N, and K_3.

Value of M	4	**8**	12	16	20
Accuracy	75.83	**76.84**	76.53	75.97	75.75
Value of N	50	**100**	150	200	250
Accuracy	75.07	**76.84**	76.73	76.51	76.25
Value of K_3	1	2	**4**	6	8
Accuracy	75.30	75.86	**76.84**	76.72	76.33

section. Unless otherwise stated, our experiments are conducted on ImageNet for the Base-to-Novel Task, and we report the harmonic mean.

Contributions of major algorithm components. In Tab. 4, $\mathcal{L}_{ma}$ is used as a matching loss to optimize the visual concepts in the conceptual codebook. $\mathcal{L}_{cc}$ is for consistency constraint, while $\mathcal{L}_{or}$ is employed to orthogonalize the textual features of different prompts. We conduct an ablation study by systematically removing various components of our proposed CoCoLe to assess their individual importance. For evaluation, the first row of the table showcases the overall performance of CoCoLe, achieving a harmonic mean of 76.84%. In the initial ablation experiment, we exclude $\mathcal{L}or$ from CoCoLe, resulting in a performance decrease of 0.46%. This underscores the significance of $\mathcal{L}or$ in CoCoLe. Next, we remove $\mathcal{L}_{cc}$, effectively enforcing consistency between the learnable conceptual prompts and handcrafted concept-based prompts. As a result, there is a decrease in performance by 1.98%, indicating the significance of the regularization strategy. Finally, we remove $\mathcal{L}_{ma}$, which leads to a performance drop of 2.39%. This shows that the learnable conceptual codebook plays a crucial role in CoCoLe. In general, the overall results highlight the significant contribution of all components to enhanced performance.

The value of M***,*** N***, and*** K_3***.*** As defined in Sec. 3.2, M is the length of prompts, N is the size of the conceptual codebook, and K_3 is the number of learned visual concepts (*i.e.*, keys in the codebook) selected for training at the same time. As indicated in Table 5, the model performs optimally with $M = 8$. If the prompt is overly long, it raises both training time and computational costs. Concerning N, the model excels with $N = 100$. Additionally, experimenting with various K_3 values shows that saturating training with excessive keys and prompts simultaneously diminishes model performance. The model achieves its peak performance at $K_3 = 4$.

Comparison on the training time. As shown in Tab. 6, our proposed CoCoLe exhibits a significant performance advantage over other methods. Though our CoCoLe requires more training time compared to CoOp, it still outperforms CoOp by 11%. Furthermore, when compared to CoCoOp and CPL, our method

Table 6: Comparison on the training time. We report the average accuracy across 11 datasets on base-to-novel tasks.

Methods	Prompts	Accuracy			Training-time
		Base	New	H	
CLIP	handcrafted	69.34	74.22	71.70	-
CoOp	textual	82.69	63.22	71.66	6ms/image
CoCoOp	textual+visual	80.47	71.69	75.83	160ms/image
CPL	textual+visual	84.38	78.03	81.08	25ms/image
CoCoLe	textual+visual	**85.22**	**80.31**	**82.70**	**10ms**/image

achieves substantial performance gains with less time. This showcases the efficiency and effectiveness of our proposed CoCoLe.

Visualization. As shown in Fig. 1, in order to confirm that various prompts indeed capture distinct image concepts, we employ Grad-CAM [30] to visually represent the image contents associated with different prompts. Observing Fig. 1(a), it's apparent that various prompts highlight distinct regions within the same image, showcasing the diversity of the acquired prompts. For instance, different prompts applied to the Koala image emphasize different areas such as the head, claws, and tree, illustrating the versatility of the learned prompts. To determine if the learned prompts indeed encapsulate higher-level semantics image concepts, we visualize the content of specific prompts ($\mathbf{P}_3$, $\mathbf{P}_{23}$, $\mathbf{P}_{48}$) across different images in Fig. 1(b). It's evident that $\mathbf{P}_3$ mainly captures the concept "wheels," $\mathbf{P}_{23}$ encapsulates the concept "grass," while $\mathbf{P}_{48}$ focuses on the "screen" of the devices. This highlights the effectiveness of prompts in learning key concepts that can be generalized across images, thereby enhancing performance in generalization tasks. In the Supplemental Materials, we provide additional details of our proposed CoCoLe and experimental results.

5 Conclusion

To address the problem of enhancing the generalization capability of VLMs while adapting them to downstream tasks, we propose Conceptual Codebook Learning (CoCoLe). The learned conceptual codebook consists of visual concepts as keys and conceptual prompts as values, which serves as a link between the image encoder's outputs and the text encoder's inputs. Additionally, we incorporate a handcrafted concept cache as a regularization to alleviate the overfitting issues in low-shot scenarios. Extensive experimental results demonstrate that our CoCoLe method remarkably outperforms the existing state-of-the-art methods.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 62331014) and Project 2021JC02X103.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision. pp. 446–461 (2014)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3606–3613 (2014)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
5. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11162–11173 (2021)
6. Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., Chilimbi, T.: Multi-modal alignment using representation codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15651–15660 (2022)
7. Fei-Fei, L., Fergus, R., Perona, P.: 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. p. 178 (2004)
8. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 524–531 (2005)
9. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021)
10. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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)
11. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., 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. pp. 8340–8349 (2021)
12. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15262–15271 (2021)
13. Hu, X., Zhang, C., Zhang, Y., Hai, B., Yu, K., He, Z.: Learning to adapt clip for few-shot monocular depth estimation. arXiv preprint arXiv:2311.01034 (2023)

14. Huang, C., Loy, C.C., Tang, X.: Unsupervised learning of discriminative attributes and visual representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5175–5184 (2016)
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916 (2021)
16. Kan, B., Wang, T., Lu, W., Zhen, X., Guan, W., Zheng, F.: Knowledge-aware prompt tuning for generalizable vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15670–15680 (2023)
17. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
18. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 554–561 (2013)
19. Lei Ba, J., Swersky, K., Fidler, S., et al.: Predicting deep zero-shot convolutional neural networks using textual descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4247–4255 (2015)
20. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3337–3344 (2011)
21. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
22. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
23. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3498–3505 (2012)
24. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2751–2758 (2012)
25. Patterson, G., Hays, J.: Coco attributes: Attributes for people, animals, and objects. In: European Conference on Computer Vision. pp. 85–100 (2016)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
27. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International Conference on Machine Learning. pp. 5389–5400. PMLR (2019)
28. Roy, S., Etemad, A.: Consistency-guided prompt learning for vision-language models. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (2024)
29. Sariyildiz, M.B., Perez, J., Larlus, D.: Learning visual representations with caption annotations. In: European Conference on Computer Vision. pp. 153–170 (2020)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)

31. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
32. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems. vol. 32, pp. 10506–10518 (2019)
33. Wang, R., Duan, X., Kang, G., Liu, J., Lin, S., Xu, S., Lü, J., Zhang, B.: Attriclip: A non-incremental learner for incremental knowledge learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3654–3663 (2023)
34. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision. pp. 631–648. Springer (2022)
35. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (2010)
36. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6757–6767 (2023)
37. Zhang, R., Qiu, L., Zhang, W., Zeng, Z.: Vt-clip: Enhancing vision-language models with visual-guided texts. arXiv preprint arXiv:2112.02399 (2021)
38. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European Conference on Computer Vision. pp. 493–510. Springer (2022)
39. Zhang, Y., Zhang, C., Tang, Y., He, Z.: Cross-modal concept learning and inference for vision-language models. arXiv preprint arXiv:2307.15460 (2023)
40. Zhang, Y., Zhang, C., Yu, K., Tang, Y., He, Z.: Concept-guided prompt learning for generalization in vision-language models. In: AAAI Conference on Artificial Intelligence (2024)
41. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
42. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
43. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. arXiv preprint arXiv:2205.14865 (2022)