LoGoPrompt: Synthetic Text Images Can Be Good Visual Prompts for Vision-Language Models

Cheng Shi, Sibei Yang[†]

School of Information Science and Technology, ShanghaiTech University

{shicheng2022,yangsb}@shanghaitech.edu.cn

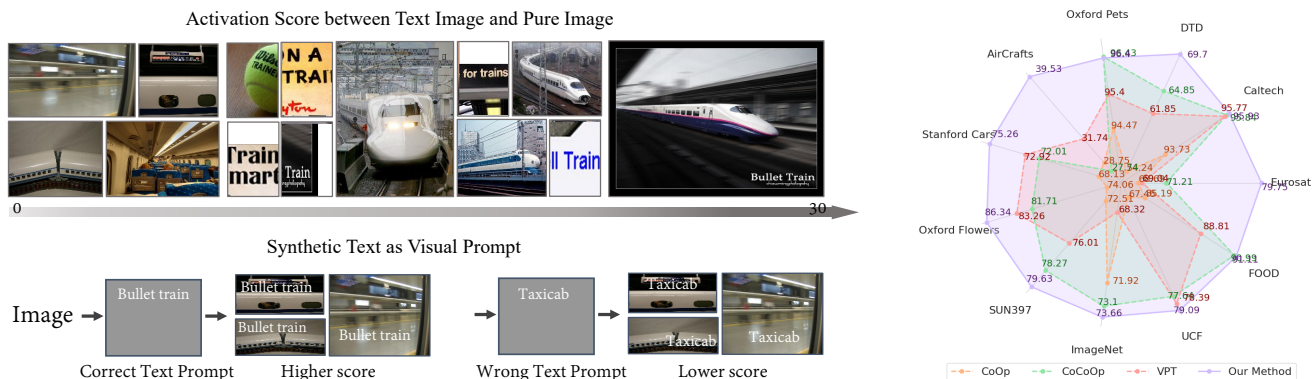Project Page: https://chengshiest.github.io/logo †: Corresponding author

Figure 1: Why can synthetic text images be good visual prompts for CLIP: (a) Images are sorted in ascending order of their zero-shot CLIP scores with the text *"A photo of train"*. (b) Significant improvement of zero-shot CLIP scores for images with synthetic text images as visual prompts. (c) Comparison of base-to-new generalization among prompt-tuning methods.

Abstract

Prompt engineering is a powerful tool used to enhance the performance of pre-trained models on downstream tasks. For example, providing the prompt "Let's think step by step" improved GPT-3's reasoning accuracy to 63% on MutiArith while prompting "a photo of" filled with a class name enables CLIP to achieve 80% zero-shot accuracy on ImageNet. While previous research has explored prompt learning for the visual modality, analyzing what constitutes a good visual prompt specifically for image recognition is limited. In addition, existing visual prompt tuning methods' generalization ability is worse than text-only prompting tuning. This paper explores our key insight: synthetic text images are good visual prompts for vision-language models! To achieve that, we propose our LoGoPrompt, which reformulates the classification objective to the visual prompt selection and addresses the chicken-and-egg challenge of first adding synthetic text images as class-wise visual prompts or predicting the class first. Without any trainable visual prompt parameters, experimental results on 16 datasets demonstrate that our method consistently outperforms state-of-the-art methods in few-shot learning,

base-to-new generalization, and domain generalization.

1. Introduction

Large-scale contrastive vision-language models (VLMs) like CLIP [34] and ALIGN [18], pretrained on large-scale image-text pairs via contrastive learning, encode general knowledge about the alignment of visual concepts and textual sequences. VLMs with dual-encoder separately encode images and texts into vectors in joint embedding space, enabling the transfer for downstream image classification by treating image vectors as "image features" and text vectors corresponding to classes as "classification weights", respectively. The class-specific weights can encoded from handcraft *prompts* [34, 18, 10, 24, 25], such as "a photo of a [*class*]", where the [*class*] token is filled with real class names. Thanks to exploring open-set visual concepts with high-capacity text encoder, VLMs with handcraft prompts show impressive potential in zero-shot transfer to image classification tasks.

Recently, CoOp [52] and CoCoOp [53] utilize prompt tuning [21, 27, 25] to learn a continuous text prompt with only a few shots of images and improve zero-shot VLMs' performance. However, the text prompt tuning can only

change the "classification weights" but keeps the "image features" unchanged. It is a sub-optimal solution for adaption, making it hard to classify images with different classes but close image features. Therefore, some works [48, 1, 19] introduce visual prompts to VLMs for simultaneously adjusting the "image features" and "classification weights". The visual prompts of VPT [19] and DPT [48] are specific to ViT architecture [8], and they adapt to downstream tasks by tuning additional image patch token embeddings. Considering the generality to different families of visual backbones, the recent work [1] proposes to learn image perturbation to perform data-space adaption. Although it explores a new visual prompting paradigm, the performance improvement is limited, especially in the same few-shot setting as the above-mentioned methods. Furthermore, these visual prompts [48, 1, 19] seem to affect VLMs' generalization ability, whose base-to-new generalization performance is lower than zero-shot VLMs and approaches with text-only prompting [52, 53] (see Figure 1).

Therefore, we propose to explore visual prompting for VLMs' adaption, which can (1) work for different backbone families such as CNNs and Transformers, (2) effectively adapt to downstream classification tasks with few shots of images, and (3) preserve the VLMs' generalization ability. To achieve these goals, we follow [1] to perform data-space adaption (*i.e.*, adapting on image pixels) and then analyze how VLMs understand different images, especially the classification accuracy gaps between images with the same class. We are surprised to observe that the images with class name text have very high confidence to classify to the class, even only lower than the simplest cases, as shown in Figure 1. The empirical study [12] about multimodal neurons of CLIP validates our observation. For a class, both images with the class name text and generic natural images of that class can activate the same neurons important for classifying that class.

The observation motivates our key insight: we can use synthetic images with class name text as the visual prompts for VLMs! However, it is not trivial to effectively utilize synthetic images because simply treating the synthetic images as extra image data cannot achieve consistent performance improvement across different datasets and with different image shots. In this paper, we propose to use synthetic images with class name text (*i.e.*, the class-wise visual prompts) to modify images in the training set so that class-wise synthetic parts can help VLMs perceive class-relevant content in the original images for classification. In the training phase, given a training image and its ground-truth class, we can easily transform it with its class-wise visual prompt, such as randomly replacing one of the original image's pixel blocks with the synthetic one, as shown in Figure 1. However, in the testing stage, since the test image's class is unknown, it remains unclear which class-wise visual prompt

should be used to benefit the classification prediction of the test image, which is a chicken-and-egg problem. Therefore, we reformulate the downstream classification objective as the visual prompt selection: select the synthetic images of the correct class for the original image as its visual prompts. Specifically, for a training image, we transform it into multiple images with different class-wise visual prompts and maximize the similarity of the image with the ground-truth visual prompt to the text features of that ground-truth class while minimizing the similarities of other images.

Furthermore, to develop the visual prompt selection to be effective and efficient, we propose a min-max contrastive learning objective and introduce hard negative mining [37, 17]. The min-max contrastive learning first groups an original image and its transformed image with the class-wise visual prompt as a group. Then, it maximizes the minimal similarity to the text features in one ground-truth group while minimizing the maximal similarity in negative groups to preserve the ability to classify the original image.

Despite the simplicity, our novel insights of using synthetic images with class name text as visual prompts and visual prompt selection learning strategy are particularly effective. Without requiring any tuning of visual prompts, our method significantly outperforms state-of-the-art methods [53] on base-to-new generalization, especially compared to previous visual prompting approaches [48, 1, 19], as shown in Figure 1. Notice that our visual prompts have no parameters to learn, but we still have text prompts for tuning following existing works [52, 48, 53]. To verify whether our method can benefit from visual prompt tuning, we extend a tuning version for visual prompts, further boosting the performance on few-shot classification. We name our proposed method LoGoPrompt, since we incorporate synthetic text onto images in a manner akin to the application of logos onto visuals. We evaluate and compare LoGoPrompt with state-of-the-art methods on 16 datasets that cover diverse image classification tasks. In summary, we make the following contributions:

- We are the first to propose using synthetic images with the text class name as visual prompts for VLMs. Our visual prompts can work for different backbone families, *e.g.*, CNNs and Transformers.
- We reformulate the classification objective to visual prompt selection to address the chicken-and-egg challenge of adding class-wise visual prompts and achieve the selection via min-max contrastive learning.
- Despite the simplicity, experiments on 16 datasets demonstrate our novel insight and learning strategy particularly effectively, consistently outperforming state-of-the-art methods in base-to-new generalization, few-shot learning, and domain generalization.

2. Related Work

Vision-Language Pre-training Models, pre-trained on image-text corpora for modeling vision and language, have shown great potential for transferable to downstream vision and language tasks. The pre-training approaches mainly contain the BERT-like masked-language and masked-region modeling methods [28, 40, 41, 4], contrastive learning for learning a joint embedding space of vision and language [34, 18, 22, 50], and vision-language multimodal autoregressive techniques [5, 35]. In this paper, we focus on the contrastive vision-language models (VLMs) that adopt a dual encoder to encode images and texts into the joint embedding space and use contrastive learning to align the visual and textual representations. The VLMs, particularly CLIP [34] and ALIGN [18], leverage hundreds of millions and even billions of image-text pairs to learn transferable visual representation from textual supervision and shows impressive zero-shot performance for various image classification tasks. CLIPPO [42] also treats render images as text to achieve similar results compared with CLIP without a text-specific tower or embedding. In this work, we devise our LoGoPrompt based on the CLIP.

Data-efficient Fine-tuning for VLMs learns to adapt the VLMs to downstream tasks with a few shots of samples. Following [52, 53], we focus on the image classification task. Linear Probe [34] straightforwardly fixes the visual backbone and learns a classifier for classification. CLIP-Adapter [11] and Tip-Adapter [51] introduce additional feature adapters to improve the few-shot classification of CLIP. Unlike the feature adapters, prompt tuning methods [52, 53, 54] learn input prompts to keep a closer objective form to the pre-training task and can achieve better generalization ability. CoOp [52] proposes a context optimization method to learn task-specific text prompts. Co-CoOp [53] extends the static text prompts of CoOp by learning image-conditional dynamic prompts to improve the base-to-new generalization. ProGrad [54] addresses the issue of forgetting general knowledge by regularizing the tuning step not to conflict with the zero-shot CLIP's prediction. ProDA [29] prompt distribution learning to handle the varying visual representation. In this work, we focus on prompt-based tuning and propose to take synthetic images with class name text as visual prompts to improve both base-to-new generalization and few-shot classification.

Visual Prompt Tuning. Inspired by the success of prompt tuning in the NLP community [3, 23, 26], recent works [19, 49, 1, 45, 38] explore visual prompt tuning for vision and vision-language models. CPT [49] learns color-based prompts to adapt the BERT-like VLMs to the visual grounding task. VPT [19] and VPT for generative transfer [38] learn visual tokens for Transformer-based vision models for better transfer to visual recognition and image synthesis tasks, respectively. Unlike adapting to computer vision tasks, VPT for text classification [45] deploys the VLMs in text classification via visual prompt tuning. The recently published DPT [48] and the unpublished VP [1] are the most relevant works to ours, and both of them adapt the contrastive VLMs to downstream image classification. DPT extends VPT by dynamically generating visual prompts via the cross-attention between text prompts features and visual image patch token embeddings. But, its visual prompts are specific to Transformer-based visual backbones. In contrast, VP learns the visual prompts in pixel space for generalizable to different backbones. However, it cannot obtain satisfied few-shot performance, demonstrating that simply learning pixel-space perturbations for images cannot achieve data-efficient tuning for VLMs. Unlike previous methods, our class-wise visual prompts are synthetic images with the class name text, which naturally work for different backbone families and have no trainable parameters, and can improve state-of-the-art prompt tuning for VLMs.

3. Method

Figure 2 shows an overview of our proposed LoGo-Prompt. We first briefly review the CLIP [34] and CoOp [52] in Section 3.1. Next, we introduce LoGoPrompt, including the visual prompt generation, min-max contrastive learning, and extension to tunable visual prompts, in Section 3.2. Like CoOp, our LoGoPrompt is applicable to other CLIP-like contrastive VLMs.

3.1. Preliminary

Contrastive Language-Image Pre-training (CLIP) [34] has a pair of image and text encoders, where the image encoder $f(\cdot)$ can be either a CNN (e.g., ResNet [13]) or a Vision Transformer (e.g., ViT [8]), and the text encoder $g(\cdot)$ is a Transformer [43]. During training, CLIP adopts the dual encoder to separately encode images and text into vectors in joint embedding space and utilizes contrastive loss to maximize the cosine similarities of the real image-text vector pairs while minimizing the cosine similarities of incorrect pairs. After being pre-trained on highly-diversified hundreds of millions of image-text pairs, CLIP is available for computing the text-image similarity and can be generalized to downstream image recognition without fine-tuning.

Let $\boldsymbol{x}$ be an input image and $\{[class]_c\}_{c=1}^{C}$ be the C categories for classification, where $[class]_c$ represents the class name of the c-th class. With a handcraft prompt like "a photo of a $[class]_c$.", the prediction probabilities are:

$$p(\hat{y} = c|\boldsymbol{x}) = \frac{\exp(\cos(f(\boldsymbol{x}), g(\boldsymbol{l}_c))/\tau)}{\sum_{i=1}^{C} \exp(\cos(f(\boldsymbol{x}), g(\boldsymbol{l}_i))/\tau)} \quad (1)$$

where $\boldsymbol{l}_c$ is the sequence embeddings of the handcraft prompt for c-th class, τ is a learned temperature coefficient, $\cos(\cdot, \cdot)$ represents the cosine similarity, and $f(\cdot)$ and $g(\cdot)$ are the image encoder and text encoder, respectively.
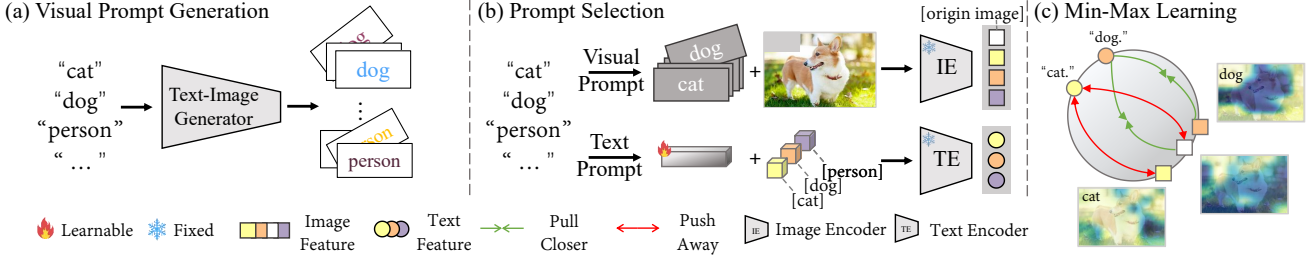
Figure 2: **Overview of LoGoPrompt**, which (a) generates class-wise visual prompts as synthetic images with text class names and (b) reformulates the classification objective to visual prompt selection to address the chicken-and-egg challenge by (c) the proposed min-max contrastive learning.

Context Optimization (CoOp) [52] addresses time-consuming and unstable issues of prompt engineering by replacing the fixed handcraft prompt with the tunable prompt that can be learned from data. The tunable prompt is composed of M learnable continues context vectors $\boldsymbol{u} = [\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M]$ that has same dimension as the word embedding and the word embedding of class names. For c-th class, the prompt $\boldsymbol{p}_c$ is $[\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M, \boldsymbol{e}_c]$, where $\boldsymbol{e}_c$ is the word embedding of the c-th class's class name $[class]_c$. CoOp optimizes the context vectors $\boldsymbol{u}$ by minimizing the cross-entropy loss between the ground truth and prediction probability as follows,

$$p(\hat{y} = c | \boldsymbol{x}) = \frac{\exp(\cos(f(\boldsymbol{x}), g(\boldsymbol{p}_c))/\tau)}{\sum_{i=1}^{C} \exp(\cos(f(\boldsymbol{x}), g(\boldsymbol{p}_i))/\tau)}, \quad (2)$$
$$\mathcal{L}_{ce} = -\log p(\hat{y} = y | \boldsymbol{x})$$

where y-th class is the ground-truth class of the image $\boldsymbol{x}$. Note that only the context vectors $\boldsymbol{u}$ are updated during the tuning, while CLIP's image encoder $f(\cdot)$ and the text encoder $g(\cdot)$ are frozen. We also follow the same training protocol as CoOp.

3.2. Synthetic Text Images Can Be Good Visual Prompts for VLMs

We now present the proposed LoGoPrompt. As illustrated in Figure 2, our LoGoPrompt first generates visual prompts as synthetic images with class name texts (Section 3.2.1). Then, it reformulates the learning objective of image classification to the visual prompt selection by utilizing min-max contrastive learning (Section 3.2.2). Moreover, LoGoPrompt can easily extend the frozen visual prompts to be tunable to boost the performance further (Section 3.2.3).

3.2.1 Visual Prompt Generation

Motivated by the observation that images with class name text easily activate the same classification neurons as general natural images of the same class (presented in the Introduction 1), we propose to use synthetic images with class

name text as visual prompts. As the class name text is independent and different for different classes, the visual prompts are naturally class-wise. The class-wise visual prompts are expected to help VLMs perceive class-relevant content in general natural images for better classification.

Class-wise Visual Prompts are defined as class-specific text images, *i.e.*, the synthetic pixel blocks of the text classes rendered on the empty background, where the colors of the text classes and the background are randomly generated. Figure 2a shows class-wise visual prompts of two different categories of dog: "Saluki" and "Otterhound". Formally, for the c-th class, we denote its visual prompt as $V_c \in \mathbb{R}^{h \times w \times 3}$, where the h and w are the height and width of the pixel block, respectively. The class-wise visual prompts $\{V_c\}_{c=1}^{C}$ have the same size of $h \times w$ across different classes.

Class-conditional Images. Given an input image $\boldsymbol{x}$, we use the class-wise visual prompt V_c to transform it into a class-conditional image $\boldsymbol{x}_c$ of the c-th class. There are different ways to perform the transform, and we simply randomly replace a $h \times w$ pixel block of the image $\boldsymbol{x}$ with the c-th class's visual prompt V_c for simplicity. The class-conditional image $\boldsymbol{x}_c$ can be treated as an augmented image of the original image $\boldsymbol{x}$ on the specific class c. For example, as shown in Figure 2c, the visual prompt of "dog" enhances the original image with the class "dog".

3.2.2 Min-Max Contrastive Learning for Visual Prompt Selection

Problem Formulation: Visual Prompt Selection. Although class-conditional images enhance the original images by applying class-wise visual prompts, it remains unclear how a class-wise visual prompt should be selected to enhance the image in the test phase. As the test image's class is unknown, it becomes a chicken-and-egg problem: should the class of the test image be predicted first to obtain the class-wise visual prompt or should the test image be augmented with the class-wise visual prompt for better prediction? To overcome the challenge, we reformulate the

classification objective as the visual prompt selection: learn to select candidate class-wise visual prompts for the image during the training phase so that the same selection strategy can be applied in the test phase.

Specifically, we construct the real and negative class-conditional images for a training image by applying ground truth and other class-wise visual prompts, respectively. Then, we utilize contrastive loss to maximize the prediction probability for the real image while minimizing that of the negative image. More importantly, to ensure that original images can be classified, we construct the groups of class-conditional and original images and improve the traditional contrastive loss to min-max one for applicable to groups of images.

Sample Construction for Synthetic and Original Image-Class Pairs constructs real and negative image-class pairs for class-conditional and original images. Given the input image x and its ground-truth class y, we construct a group of real image-class pairs and K groups of negative image-class pairs. Each group has two image-class pairs, one for the original image and one for the class-conditional image. Our construction rules are as follows:

- Group of real image-class pairs is $[(x, y)^+, (x_y, y)^+]$. The reason is that the ground-truth label y is the correct class for both the original image x and the class-conditional image x_y of the class y. For example, the pairs of (image, ground-truth class "dog") and (image with visual prompt "dog", ground-truth class "dog") are correct, as shown in Figure 2c.
- The K groups of negative image-text pairs are $\{[(x, c_k)^-, (x_{c_k}, c_k)^-]\}_{k=1}^K$, where class $c_k \neq y$. The class c_k is incorrect class for original image x and all class-conditional images $\{x_c\}_{c=1}^C$, including the class-conditional image x_{c_k} of class c_k. Since class y is the correct class for image x, the visual prompt V_{c_k} of class c_k cannot activate visual concepts relevant to class c_k from the image x of class y, resulting in $(x_{c_k}, c_k)^-$ is a negative pair. The class-wise visual prompts are expected to help VLMs to perceive class-relevant visual concepts in the image and should not change the original inherent visual semantics of the image. For example, an image of "a dog" should not be recognized as "a cat" by adding "cat" visual prompt.

Min-Max Contrastive Loss maximizes the similarity of real pairs $[(x, y)^+, (x_y, y)^+]$ while minimizing that of negative pairs $\{[(x, c_k)^-, (x_{c_k}, c_k)^-]\}_{k=1}^K$ to optimize the context vectors u of text prompts $\{p_i\}_{i=1}^C$ (see CoOp in Section 3.1). Our visual prompts are synthetic images with class name text, which have no parameters to learn. Therefore, we only need to optimize continuous context vectors u following CoOp. Formally, we extend the InfoNCE loss [32] on the groups of real and negative pairs as follows,

$$p(\hat{y} = c | x) = \frac{\exp(\cos(f(x), g(p_c))/\tau)}{\sum_{i=1}^C \exp(\cos(f(x), g(p_i))/\tau)},$$
$$\mathcal{L}_N = -\log \frac{\min(p(y|x), p(y|x_y))}{\sum_{k=1}^K \max(p(c_k|x), p(c_k|x_{c_k}))} \quad (3)$$

where the $\min(\cdot, \cdot)$ and $\max(\cdot, \cdot)$ operations obtain the minimum and maximum matching probabilities for the real and negative groups, respectively. Note that the min-max contrastive loss maximizes the minimal probability within the real group while minimizing the maximal probability within negative groups to preserve the ability to classify the original image.

During tuning, we adopt the hard negative mining to select the K classes $\{c_k\}_{k=1}^K$ with the top-K prediction probabilities $p(\hat{y} = c_k|x)$ for the original image x other than the ground-truth class y. During inference, we also first obtain the top-K prediction classes for the original image, and then select the class with the highest prediction probability for the original or the corresponding class-conditional image as the predicted class, *i.e.*, $\hat{y} = \text{argmax}_{c_k}{}_{k=1}^K \max(p(c_k|x), p(c_k|x_{c_k}))$.

3.2.3 Extension to Tunable Prompts

Visual Prompt: We use the synthetic pixel blocks of the text classes to initialize the visual prompts before tuning. Next, we optimize both the visual prompts $\{V_c\}_{c=1}^C$ and textual context vectors u during the tuning via the proposed contrastive prompt learning in Section 3.2.2.

Text Prompt: Following [52], we adopt a class-specific text prompt and optimized it in a two-stage strategy. In the first stage, inspired by the Meta-Net of CoCoOp [53], we use a lightweight MLP to directly generate class-specific text prompts from handcraft prompts' text vectors extracted by CLIP's text encoder. In the second stage, we fine-tune the learned prompt from stage one directly on seen classes.

4. Experiments

We evaluate the proposed LoGoPrompt on three problem settings: (1) generalization from base classes to new classes (Section 4.1), (2) few-shot classification (Section 4.2), (3) domain generalization (Section 4.3).

Datasets. For the first two settings, we follow CLIP [34] and CoOp [52] to evaluate on 11 image classification datasets, *i.e.*, ImageNet (Img) [7], Caltech101 (Cal) [9], OxfordPets (Pet) [33], StanfordCars (Car) [20], Flowers102 (Flo) [31], Food101 (Foo) [2], FGVCAircraft (FGV) [30], SUN397 (SUN) [47], DTD [6], EuroSAT (Eur) [14], and UCF101 (UCF) [39]. For domain generalization, we adopt ImageNet [7] as the source and ImageNetV2 (V2) [36], ImageNet-Sketch (Sketch) [44], ImageNet-A (A) [16] and

Table 1: **Accuracy comparison for base-to-new generalization setting**. The prompts are learned from the base classes (16 shots) for prompt-based tuning methods). Following CoCoOp, ViT-B/16 of CLIP is used as the vision backbone. H: Harmonic mean [46]. The latter three methods employ visual prompts. The superior performance of LoGoPrompt on both base and new classes shows its strong generalizability.

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

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
VPT	80.81	70.36	74.68
DPT	81.29	29.23	44.54
Ours	**84.47**	**74.24**	**79.03**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
VPT	70.93	65.90	68.32
DPT	75.15	25.43	38.00
Ours	**76.74**	**70.83**	**73.66**

(c) Caltech101.

	Base	New	H
CLIP	96.84	**94.00**	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
VPT	97.86	93.76	95.77
DPT	97.79	45.21	61.83
Ours	**98.19**	93.78	**95.93**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	**97.69**	**96.43**
VPT	94.81	96.00	95.40
DPT	95.11	54.55	69.33
Ours	**96.07**	96.31	96.18

(e) StanfordCars.

	Base	New	H
CLIP	63.37	**74.89**	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
VPT	72.46	73.38	72.92
DPT	73.65	34.35	46.84
Ours	**78.36**	72.39	**75.26**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
VPT	95.39	73.87	83.26
DPT	95.20	17.22	57.50
Ours	**99.05**	76.52	**86.34**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
VPT	89.88	87.76	88.81
DPT	88.87	26.81	41.19
Ours	**90.82**	**91.41**	**91.11**

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	31.09
CoOp	40.44	22.30	28.75
CoCoOp	33.41	23.71	27.74
VPT	33.10	30.49	31.74
DPT	34.27	2.24	4.20
Ours	**45.98**	34.67	**39.53**

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
CoCoOp	79.74	76.86	78.27
VPT	79.66	72.68	76.01
DPT	80.42	25.22	38.39
Ours	**81.20**	**78.12**	**79.63**

(j) DTD.

	Base	New	H
CLIP	53.24	59.90	56.37
CoOp	79.44	41.18	54.24
CoCoOp	77.01	56.00	64.85
VPT	79.15	50.76	61.85
DPT	79.53	25.54	38.66
Ours	**82.87**	**60.14**	**69.70**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
CoCoOp	87.49	60.04	71.21
VPT	93.01	54.89	69.04
DPT	93.07	28.61	43.76
Ours	**93.67**	**69.44**	**79.75**

(l) UCF101.

	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	84.69	56.05	67.46
CoCoOp	82.33	73.45	77.64
VPT	82.67	74.54	78.39
DPT	81.09	36.35	50.19
Ours	**86.19**	73.07	**79.09**

ImageNet-R (R) [15] as target datasets following Co-CoOp [53].

Training details. For all three problem settings, we follow the experimental settings of CoOp [52] and CoCoOp [53] for a fair comparison, including dataset splits, data augmentation, training schedule, shots of samples, backbones, length of context tokens (*i.e.*, M is 16 for few-shot classification and 4 for other problems), *etc*. The K is set to 5 for all the experiments. Please refer to the Supplementary for more details.

4.1. Base-to-New Generalization

To evaluate the base-to-new generalization ability, we follow CoCoOp to train LoGoPrompt on 16 images per base class in the training set and report the performance on both base and new classes of the test set. As shown in Table 1, the average performance of LoGoPrompt consistently surpasses other methods in terms of all three metrics, demonstrating that our LoGoPrompt has not only impressive few-

shot learning ability but also has strong generalizability.

Specifically, the Harmonic mean [46] indicates the generalization trade-off between base classes and new classes. We evaluate DPT [48] and VPT based on their official implementations. Compared with DPT and VPT that also use visual prompts, LoGoPrompt shows impressive generalization ability, which improves the average accuracy on new classes by 3.88%. Moreover, LoGoPrompt shows a clear gain of 3.20% over the previous best-performing CoCoOp and surpasses CoCoOp on 10 out of 11 datasets. For the accuracy on base classes, LoGoPrompt outperforms the strong few-shot baseline CoOp by nearly 2% on average accuracy and beats all the methods on all the datasets. For the accuracy of unseen new classes, LoGoPrompt improves the average accuracy of learning-based CoOp and CoCoOp by 11.02% and 2.55%, respectively. Moreover, LoGoPrompt even achieves the average accuracy of zero-shot CLIP and outperforms CLIP by 2.69%, 2.77% and 5.39% on ImageNet, SUN397 and EuroSAT datasets, respectively. The
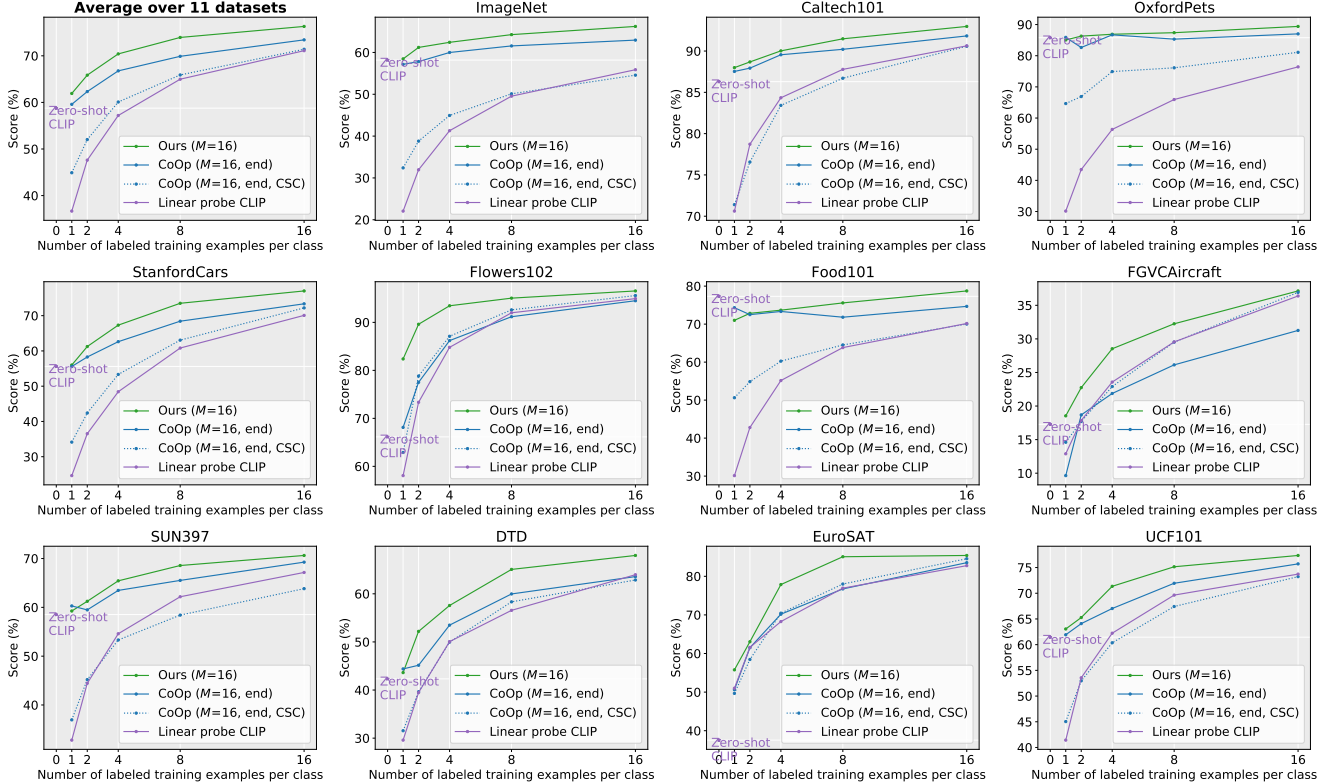
Figure 3: Accuracy comparison of few-shot classification. LoGoPrompt consistently outperforms compared methods on all the 11 datasets. Following CoOp, ResNet-50 of CLIP is used as the vision backbone.

results demonstrate that our class-specific visual and text prompts tuning method does not hurt the generalization ability of CLIP even if it learns the prompts based on the base classes during tuning.

4.2. Few-Shot Classification

Figure 3 summarizes accuracy (%) comparison with 1, 2, 4, 8 and 16 shots on 11 datasets. Our LoGoPrompt consistently outperforms prompt-tuning models and zero-shot CLIP [34] on all the datasets with different shots, which demonstrates LoGoPrompt's data effectiveness and generalization ability on various types of datasets. Specifically, our LoGoPrompt outperforms CoOp [52] by 3.28% and 3.58% on average accuracy given 2 and 4 shots, respectively. The results show that LoGoPrompt can achieve superior performance even with limited training samples. Moreover, LoGoPrompt improves the average accuracy over all the shots by 7.91%, 6.32%, and 3.04% on Flowers102, FGVCAircraft and StanfordCars datasets respectively, expressing its effectiveness on fine-grained image recognition.

Although we focus on the line of prompt-based tuning methods, we also compare LoGoPrompt to other fine-tuning approaches, *i.e.*, Linear Probe [34], CLIP-Adapter [11] and Tip-Adapter [51]. The results illustrated in Table 2 show

that LoGoPrompt outperforms other fine-tuning methods on various datasets. Besides, our method can benefit from the feature adapter. With an adapter, the performance of our LoGoPrompt further improves and surpasses state-of-the-art methods on all the datasets. Note that we add a feature adapter following Tip-Adapter but do not search for the optimal hyper-parameter on thousands of sets of hyper-parameters like Tip-Adapter.

Moreover, we evaluate the few-shot performance of Visual Prompting [1], the most relevant visual prompting work to ours, using the officially released code. As shown in Table 2, our LoGoPrompt surpasses Visual Prompting [1] by large margins, which demonstrates the effectiveness of our class-specific visual prompt tuning. Please refer to the Supplementary for detailed results and more comparisons.

4.3. Domain Generalization

Following CoOp and CoCoOp, we validate the generalization of our LoGoPrompt to out-of-distribution data. We evaluate the accuracy by transferring LoGoPrompt trained on ImageNet to four target benchmarks, and the results are shown in Table 3. Our LoGoPrompt consistently outperforms compared models on all four target datasets, and the average improvement on the source and target datasets are

Table 2: **Comparison of LoGoPrompt, CoOp and other fine-tuning methods in the few-shot classification setting** (16 shots). Following CoOp, ResNet-50 of CLIP is used as the vision backbone for all the models. The performance of LoGo-Prompt surpasses both prompt-based learning and other fine-tuning methods. VP and TP are abbreviation for Visual prompt and Text prompt. Best results are bold.

	Prompt Adapter	Ima [7]	Cal [9]	Pet [33]	Car [20]	Flo [31]	Foo [2]	FGV [30]	SUN [47]	DTD [6]	Eur [14]	UCF [39]	Average
CoOp [52]	TP	62.95	91.83	87.01	73.36	94.51	74.67	31.26	69.26	63.58	83.53	75.71	73.42
Vis Prompt [1]	VP+TP	50.65	77.02	69.64	56.99	89.97	59.99	23.01	57.01	55.67	73.19	67.11	61.84
Ours	VP+TP	**66.34**	**92.98**	**89.40**	**77.01**	**96.55**	**78.73**	**37.14**	**70.63**	**67.97**	**85.40**	**77.35**	**76.31**$_{\pm\,0.52}$
Tip [51]	TP ✓	62.01	90.18	88.14	66.77	89.89	77.83	29.76	66.85	60.93	70.54	70.58	70.32
Tip-F [51]	TP ✓	65.51	92.86	89.70	75.74	94.80	79.43	35.55	71.47	66.55	84.54	78.03	75.83
Ours	VP+TP ✓	**67.34**	**93.23**	**90.35**	**77.64**	**96.83**	**79.77**	**38.70**	**72.13**	**70.07**	**85.74**	**78.32**	**77.28**$_{\pm\,0.43}$

	Source	Target				Avg
	ImageNet	V2	Sketch	A	R	
CLIP [34]	66.73	60.83	46.15	47.77	73.96	59.08
CoOp [52]	71.51	64.20	47.99	49.71	75.21	61.72
CoCoOp [53]	71.02	64.07	48.75	50.63	76.18	62.13
VPT [48]	70.72	58.22	44.67	43.00	71.86	57.69
DPT [48]	69.95	15.93	13.41	9.10	32.34	28.14
Ours	**75.27**	**66.65**	**48.99**	**51.36**	**76.85**	**63.82**

Table 3: **Accuracy comparison of domain generalization**. Learning-based methods are trained on ImageNet with 1,000 classes and 16 images per class. Following CoCoOp, ViT-B/16 of CLIP is used as the vision backbone. LoGo-Prompt is more domain-generalizable than others.

(a) Visual-Pro Type		(b) Visual-Pro Size	
text augmentation	63.12	ratio = 1/14	66.02
paired-text-image	63.54	ratio = 1/7	**66.34**
text-image-selection	**66.34**	ratio = 2/7	66.17
(c) Text-Pro Length		(d) Visual-Pro Location	
$M = 2$	65.58	top	65.98
$M = 8$	66.14	bottom	66.01
$M = 16$	**66.34**	rand	**66.34**$_{\pm\,0.21}$

Table 4: **Further analysis and ablation study on ImageNet dataset,** which reports the few-shot (16-shot) accuracy with ResNet-50 as the vision backbone.

2.10% and 1.69% compared to CoOp and CoCoOp, respectively. Furthermore, since the instance-conditional prompt learning of CoCoOp cannot be optimized in parallel, it requires dozens of times training time than ours. The results give evidence that our visual and text prompts are more domain-generalizable and efficient. We also evaluate visual prompting methods, *i.e.* VPT and DPT. And our method significantly outperforms them by 6.13%.

4.4. Further Analysis

Visual Prompt. The variants of the visual prompt shown in Table 4a contains: (1) By directly treating synthetic text images as additional one-shot training images, the model's accuracy exhibits slight changes. (2) By replacing the text augmentation with the paired text-image, which means the standard classification loss is utilized, the model's accuracy achieves a mild gain by 0.42%. Notice that paired text-image is also another solution for the chicken-and-egg problem. Despite the performance improvement from (1), it yields only marginal improvements compared with CoOp [52]. (3) We further introduce our visual prompt selection strategy and min-max contrastive loss. The model's accuracy further improves to 66.34%. The results demonstrate not only the success and effectiveness of our strategy but also that it is non-trivial to leverage synthetic images to adapt VLM to downstream tasks better.

Besides, Table 4b and 4d illustrate the effects of the size and location of the visual prompt. Table 4c illustrates the effects of the length of the text prompt. We observe that visual prompts that are too large would dominate the class-conditioned images and ignore the original image information, which will lead to a 0.17% performance drop. Our random location strategy on the location of the visual prompt mitigates the overfitting problem and improves 0.33% compared with the fixed location.

5. Conclusion

To summarize, we present a new method that employs synthetic images with text class names as visual prompts for VLMs. The proposed LoGoPrompt reformulates the classification objective through min-max contrastive learning, overcoming the challenge of adding class-specific visual prompts. Experiment results on 16 datasets demonstrate the efficacy of our approach.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 3, 7, 8

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 5, 8

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020. 3

[5] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 3

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5, 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5, 8

[10] Andreas Fürst, Elisabeth Rumetshofer, Viet Tran, Hubert Ramsauer, Fei Tang, Johannes Lehner, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto-Nemling, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *arXiv preprint arXiv:2110.11316*, 2021. 1

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3, 7

[12] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 8

[15] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 5

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 2

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 3

[19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *ECCV*, 2022. 2, 3

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5, 8

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *EMNLP*, 2021. 1

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[23] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[24] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 1

[26] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ACL*, 2022. 3

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

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NIPS*, 32, 2019. 3

[29] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5206–5215, June 2022. 3

[30] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5, 8

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 5, 8

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[33] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5, 8

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 3, 5, 7, 8

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[36] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 5

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[38] Kihyuk Sohn, Yuan Hao, José Lezama, Luisa Polania, Huiwen Chang, Han Zhang, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. *arXiv preprint arXiv:2210.00990*, 2022. 3

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5, 8

[40] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *ICLR*, 2020. 3

[41] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *EMNLP*, 2019. 3

[42] Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11006–11017, June 2023. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[44] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[45] Jingyuan Wen, Yutian Luo, Nanyi Fei, Guoxing Yang, Zhiwu Lu, Hao Jiang, Jie Jiang, and Zhao Cao. Visual prompt tuning for few-shot text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5560–5570, 2022. 3

[46] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017. 6

[47] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 8

[48] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022. 2, 3, 6, 8

[49] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 3

[50] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[51] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *ECCV*, 2022. 3, 7, 8

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 2, 3, 5, 6, 8

[54] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 3