LaViP: Language-Grounded Visual Prompting

Nilakshan Kunananthaseelan[1], Jing Zhang[2], Mehrtash Harandi[1]

[1]Department of Electrical and Computer Systems Engineering, Monash University
[2]College of Engineering and Computer Science, Australian National University
{nilakshan.kunananthaseelan,mehrtash.harandi}@monash.edu, zjnwpu@gmail.com

Abstract

We introduce a language-grounded visual prompting method to adapt the visual encoder of vision-language models for downstream tasks. By capitalizing on language integration, we devise a parameter-efficient strategy to adjust the input of the visual encoder, eliminating the need to modify or add to the model's parameters. Due to this design choice, our algorithm can operate even in black-box scenarios, showcasing adaptability in situations where access to the model's parameters is constrained. We will empirically demonstrate that, compared to prior art, grounding visual prompts with language enhances both the accuracy and speed of adaptation. Moreover, our algorithm excels in base-to-novel class generalization, overcoming limitations of visual prompting and exhibiting the capacity to generalize beyond seen classes. We thoroughly assess and evaluate our method across a variety of image recognition datasets, such as EuroSAT, UCF101, DTD, and CLEVR, spanning different learning situations, including few-shot adaptation, base-to-novel class generalization, and transfer learning.

1 Introduction

Large-scale pretrained models (PTMs) (Brown et al. 2020a; Dosovitskiy et al. 2020; Radford et al. 2021; Touvron et al. 2023; Kirillov et al. 2023) are trained with massive amounts of data and intricate optimization algorithms. This makes designing and developing high-performing PTMs a laborious and costly process. While these models showcase generalization prowess, achieving optimal performance on new tasks necessitates careful finetuning. Nonetheless, the finetuning of PTMs carries inherent challenges, notably the risk of catastrophic knowledge forgetting and vulnerability to overfitting on the downstream tasks (Kumar et al. 2021; Wortsman et al. 2022).

In response to the challenges mentioned earlier, a fresh paradigm called *Model Reprogramming (MR)* has been proposed as a method in the context of transfer learning. The core idea behind MR is to repurpose and harness a high-quality pretrained model, facilitating seamless cross-domain learning *without the need for finetuning* the model. MR introduces a learnable transformation function at the input of the model, along with an output mapping function to achieve this objective. The pioneering work of (Tsai, Chen, and Ho 2020) has demonstrated that through MR, even a CNN initially trained on ImageNet can be swiftly adapted to excel in classifying medical images, interestingly, even outperforming the traditional finetuning approach. Subsequent research efforts have extended the idea of MR into various domains, achieving successful adaption without finetuning (Vinod, Chen, and Das 2020; Yen et al. 2021; Yang, Tsai, and Chen 2021; Neekhara et al. 2022; Chen et al. 2023).

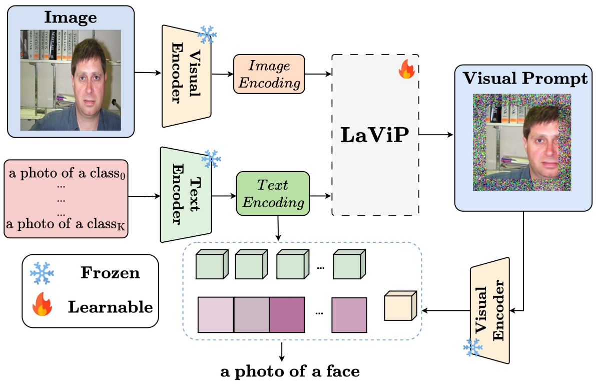

Figure 1: Our key idea is to reprogram the visual encoder of CLIP (Radford et al. 2021) through the generation of language-grounded visual prompts.

The input transformation acquired through MR is commonly conceptualized as a perturbation pattern, which is either added to or concatenated with the input images. By learning the perturbation pattern, also called Visual Prompts (VPs) in vision tasks, the PTM effectively embeds the downstream task samples into a distinct subset of its latent space. As such, MR allows to adeptly repurpose the PTM's capabilities, all while preserving the integrity of the latent space. Despite the promise and rapid progress, several questions remain unanswered in MR;

- **Unimodality in learning VPs.** To the best of our knowledge, in the previous studies focusing on VPs, class semantic information and visual encoding are typically treated separately in many cases, despite human perception being multimodal (*e.g.*, (Gibson 1969; Meltzoff and Borton 1979; Quiroga et al. 2005). This multimodal framework of our cognitive system helps us to learn new concepts with a few examples. This, in AI, will raise a simple question, *if a Vision-Language model is at hand, does language help in designing VPs for MR? If yes, what are the design questions to answer?*

- **Efficient Training.** In practice, learning VPs requires a large number of iterations to achieve quality results. *For example, adapting a PTM to classify 10 classes of satellite images in EuroSAT (Helber et al. 2019), requires 1000 training epochs.* This is because, adapting the visual encoder is challenging due to the complexity of high-dimensional visual input and the asymmetric nature of V-L encoders, compared to its text counterpart. One may wonder whether language can overcome this constraint.
- **Generalizing beyond seen classes.** MR is, by nature, a form of transfer learning. As such, it does not endow an explicit mechanism to generalize beyond what it has seen during adaptation. Recent studies have shown that Vision Language Models (VLMs) have great zero-shot learning capabilities. This would suggest whether one can expect or design an MR algorithm that can benefit from language to generalize beyond its seen classes during adaptation.
- **Adaptation without accessing model parameters** Our method maintains the original foundation model, thus enabling adaptation via APIs and cases where for ethical constraints, accessing the structure and weights of the foundation model is not possible. Furthermore, preserving the foundation model translates into maintaining its generalization capabilities, a virtue, that algorithms such as MaPLe (Khattak et al. 2023) cannot ensure.

Our work takes a stride toward addressing the aforementioned questions. In particular, we propose **Language-Grounded Visual Prompting** (LaViP)[1] , which enables pixel-space input-aware prompting by leveraging the language integration to adapt downstream tasks (Figure 1). In LaViP, we opt for a low-rank solution to generate language grounded visual prompt. This substantially reduces the number of parameters to be learned, a quality particularly advantageous in the context of black-box settings. Furthermore, we develop a mechanism to incorporate novel class knowledge without needing to retrain the VPs, enabling our solution to generalize to novel and unseen classes seamlessly. To contrast and compare our algorithm against previous art, we have performed a thorough set of experiments, ranging over transfer learning, few-shot learning, and generalization beyond seen classes over 12 recognition datasets. Our empirical study shows that our algorithm consistently outperforms state-of-the-art algorithms by a tangible margin by harnessing the multimodal signals in visual prompts.

To summarize, we have made the following contributions to this work. Firstly, to the best of our knowledge, we are pioneering a language-grounded MR solution to adapt a visual encoder to downstream tasks. Secondly, we propose a mechanism effectively extending visual prompts beyond seen classes, a feat largely confined to text prompt adaptation. We extensively evaluate and assess our algorithm on three learning paradigms: few-shot learning, generalization beyond seen classes, and transfer learning.

2 LaViP

Throughout the paper, we denote scalars as x, vectors as $\boldsymbol{x}$, matrices as $\boldsymbol{X}$, and equality by definition as $\triangleq$. The Kronecker product between matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{p\times q}$, denoted by $\boldsymbol{X} \otimes \boldsymbol{Y} \in \mathbb{R}^{mp\times nq}$ is defined as

$$\boldsymbol{X} \otimes \boldsymbol{Y} = \begin{pmatrix} x_{11}\boldsymbol{Y} & \cdots & x_{1n}\boldsymbol{Y} \\ \vdots & \ddots & \vdots \\ x_{m1}\boldsymbol{Y} & \cdots & x_{mn}\boldsymbol{Y} \end{pmatrix}, \tag{1}$$

where a_{ij} represents the element in the i-th row and j-th column of $\boldsymbol{X}$. Below, we describe **LaViP**, our input-dependent visual prompting approach guided by language semantics. In § 2.1, we provide a detailed exposition of the underlying rationale of our algorithm and its design. § 2.2 illustrates how LaViP can be transitioned to base-to-novel generalization tasks.

Problem Statement. Given a training dataset $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1}^m\}$ drawn i.i.d. from distribution $\mathcal{D}$, we seek to learn a model to effectively assign input vectors $\boldsymbol{x}$ to their corresponding class labels $\boldsymbol{y}$, based on the patterns and relationships. We assume $\boldsymbol{x}_i \in \mathbb{R}^{\mathtt{H}\times\mathtt{W}\times\mathtt{C}}$ is an image and $\boldsymbol{y}_i \in \Delta^{K-1}$ is its associated label, with Δ^{K-1} denoting the K-simplex. Furthermore, we assume a pretrained VLM with a visual encoder $\Phi_{\text{vis}} : \mathbb{R}^{\mathtt{H}\times\mathtt{W}\times\mathtt{C}} \to \mathbb{R}^d$ $\Phi_{\text{lan}} : \mathcal{X} \to \mathbb{R}^d$ is at our disposal. Here, $\mathcal{X} \subseteq \mathbb{R}^{d_t}$ denotes the input space of the language encoder, in the case of CLIP, a subset of integers defined by its tokenizer.

To achieve this goal, our objective is to generate padding-style visual prompts with a total of $2\mathtt{pC}(\mathtt{H}+\mathtt{W}-2\mathtt{p})$ parameters, where $\mathtt{C}$ represents channels, $\mathtt{H}$ and $\mathtt{W}$ denote height and width, and $\mathtt{p}$ is the padding size. Unlike previous visual prompting methods such as VP (Bahng et al. 2022), LaViP employs a method where it acquires input-specific prompts that are grounded in language.

2.1 Language Grounded Visual Prompts

Visual Prompts manipulate the pixel space via learnable parameters and steer the PTMS in any desired direction. While VP made the first contribution to this concept in the context of the pretrained vision model and VLMs, they overlooked 1) the multimodal nature of VLMs, and 2) the semantic diversity of images. To address these gaps, we propose LaViP, a novel approach that capitalizes on these two important observations. Figure 2 provides an overview of our method. LaViP synergizes complex semantics in visual inputs and context knowledge, generating visual prompts, that facilitate enhanced modality alignment.

As suggested earlier, the visual prompt for a sample $\boldsymbol{x} \in \mathbb{R}^{\mathtt{H}\times\mathtt{W}\times\mathtt{C}}$ is defined as $\boldsymbol{\nu} \in \mathbb{R}^{2\mathtt{C}(\mathtt{H}+\mathtt{W}-2\mathtt{p})\mathtt{p}}$, which is padded around a resized version of $\boldsymbol{x}$. We mathematically and with a bit of abuse of notation show this process by:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} \oplus \boldsymbol{\nu} \ . \tag{2}$$

[1] https://github.com/NilakshanKunananthaseelan/LaViP

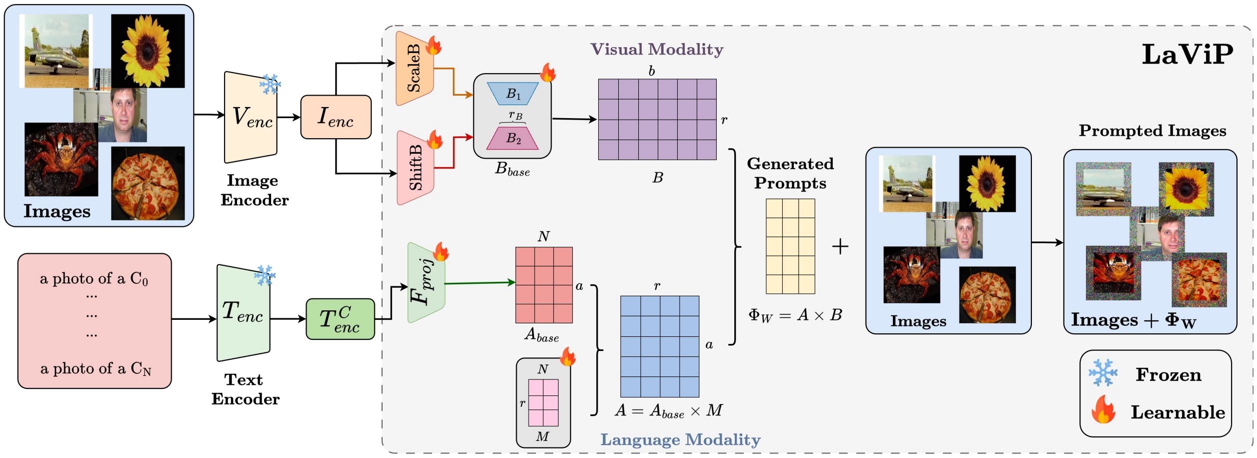

Figure 2: Overview of our proposed Language-Grounded Visual Prompting (LaViP) for VLMs: LaViP utilizes language-grounded input-specific visual programs to reprogram the frozen visual encoder of the CLIP model. LaViP scales and shifts local image encoding and projects global text encoding. The subsequent matrix multiplication of these localized and global projections fosters a mutual synergy between the two modalities, resulting in the generation of adaptive visual prompts.

For a VLM such as CLIP, typical values of $\mathtt{H} = \mathtt{W} = 224$, and $\mathtt{p} = 28$, which results in generating $2\mathtt{C}(\mathtt{H}+\mathtt{W}-2\mathtt{p})\mathtt{p}$ parameters for VPs.

We aim to facilitate the generation of input-specific visual prompts by formulating the process through low-rank matrix decomposition. Specifically, we derive two matrices $\boldsymbol{A} \in \mathbb{R}^{a\times r}$, $\boldsymbol{B} \in \mathbb{R}^{r\times b}$ and $\boldsymbol{M} \in \mathbb{R}^{K\times r}$. Here, $\boldsymbol{A}$ acts as a projection and captures the class semantics of the problem via the language encoder. Furthermore and as we will show shortly, $\boldsymbol{A}$ is obtained from the textual description of all K classes. This implies that after training, our algorithm can only store $\boldsymbol{A}$ and does not need a language encoder to operate in its nominal form.

On the other hand, the $\boldsymbol{B}$ component of the visual prompts is tailored to each image, enabling our method, LaViP, to dynamically adjust its prompts based on the input image it receives. This image dependency aligns with the idea that customized guidance can enhance model performance, as previously discussed. We argue that, despite sharing identical class labels, images often exhibit distinct semantic variations. Relying on universal visual prompts limits the model's capacity to adapt effectively to these variations, especially when extending to unseen classes. The hyper-parameter r controls the rank of $\boldsymbol{A}$ and $\boldsymbol{B}$, and can be considered as a prior in generating visual prompts. Consequently, we represent the visual prompts as $\boldsymbol{\nu} = \mathrm{Vec}(\boldsymbol{AB})$. Here, the notation $\mathrm{Vec}(\cdot)$ denotes the process of reshaping a matrix into a vector. By adopting this formulation, we reduce the complexity of requiring $\boldsymbol{\nu}$ from the initially required $2\mathtt{C}(\mathtt{H}+\mathtt{W}-2\mathtt{p})\mathtt{p}$ parameters to merely $r(a+b)$ parameters for each instance. Learning low-rank decomposition of learnable parameters has proven more effective and efficient than finetuning all parameters (Hu et al. 2021). Below, we provide a detailed explanation of how $\boldsymbol{A}$ and $\boldsymbol{B}$ are generated.

Language Grounded Encoding Following common practice (Radford et al. 2021; Bahng et al. 2022), for all K classes of a downstream, we craft textual descriptions by a template in the form: "$\mathtt{a\ photo\ of\ a}$ $\langle class\rangle$". Then, we obtain the language encoding of the text encoder for all K prompts as $\mathrm{T}_{\mathrm{enc}} \in \mathbb{R}^{K\times d}$ using:

$$\mathrm{T}_{\mathrm{enc}} = \mathrm{VLM}_{\mathrm{TextEncoder}}(\mathrm{prompts})\ . \tag{3}$$

To enrich the representation, we define $\boldsymbol{A} \triangleq \boldsymbol{M}\boldsymbol{A}_{\mathrm{base}}$. Here, $\boldsymbol{A}_{\mathrm{base}}$ is obtained from the semantics $\mathrm{T}_{\mathrm{enc}}$ (see Eq.(3)) as $\boldsymbol{A}_{\mathrm{base}} = f(\mathrm{T}_{\mathrm{enc}})$. The matrix $\boldsymbol{M}$ is learnable, helping the model to gauge the semantics to be incorporated as a part of visual prompts.

Image Dependent Encoding Similar in concept, we formulate the image-dependent part of the visual prompts as a matrix decomposition, albeit with some touch-ups. In particular, we propose the following form for constructing $\boldsymbol{B}$:

$$\boldsymbol{B} \triangleq \boldsymbol{B}_{\mathtt{scale}} \odot \boldsymbol{B}_{\mathtt{base}} + \boldsymbol{B}_{\mathtt{shift}}\ , \tag{4}$$

where $\boldsymbol{B}_{\mathtt{scale}}, \boldsymbol{B}_{\mathtt{shift}} \in \mathbb{R}^{r}$, $\boldsymbol{B}_{\mathtt{base}} \in \mathbb{R}^{b\times r}$ and $\odot$ indicates scaling function. In Eq.(4), $\boldsymbol{B}_{\mathtt{base}}$ is a matrix encoding the visual aspects of the input image and is obtained as:

$$\boldsymbol{B}_{\mathtt{base}} = \boldsymbol{B}_1 \times \boldsymbol{B}_2\ , \tag{5}$$

where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are low-rank decomposition of $\boldsymbol{B}_{\mathtt{base}}$ with rank $r_{\boldsymbol{B}}$. We modulate $\boldsymbol{B}_{\mathtt{base}}$ with $\boldsymbol{B}_{\mathtt{scale}}$ and $\boldsymbol{B}_{\mathtt{shift}}$, which are light-way matrices obtained through simple linear layers. We opt for light-design choices to accelerate image-wise transformation in Eq.(4) without introducing significant computational overhead and provide a convenient way to introduce non-linearity in the process.

Algorithm 1 summarizes the steps involved in our method.

2.2 Generalization from Base to Novel Classes

In the base-to-novel generalization task, the goal is to evaluate the *generalizability of the model to unseen classes* by

Algorithm 1: LaViP algorithm

```
Input: Target dataset: $\mathcal{D}$ with $K$ classes and $X$ images,
Pretrained model: $F$ ,
Prompt learner : $P$ with trainable parameters.
Parameters: $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{M}, F_{proj}, \mathrm{Scale}_{\mathrm{B}}, \mathrm{Shift}_{\mathrm{B}}$
Output: Visual prompt: $\Phi_W$ for the target task.

Initialize Parameters: $\boldsymbol{B}_1, \boldsymbol{B}_2$ and $\boldsymbol{M}$
Create Visual Projection Matrix: $\boldsymbol{B}_{\texttt{base}} = \boldsymbol{B}_1 \times \boldsymbol{B}_2$
Construct K textual prompts: prompts $= \{$ a photo of a $\{i\}\}_{i=1}^{K}$
Encode the image and text prompt:
    $T_{enc}, I_{enc} = F(x, \mathrm{prompts}), x \in X$
Project text encoding : $\boldsymbol{A}_{\texttt{base}} = F_{proj}(T_{enc})$
Control the text knowledge: $\boldsymbol{A} = \boldsymbol{M} \times \boldsymbol{A}_{\texttt{base}}$
Scaling and shifting of image encoding:
    $\boldsymbol{B}_{\texttt{scale}} = \mathrm{Scale}_{\mathrm{B}}(I_{enc})$
    $\boldsymbol{B}_{\texttt{shift}} = \mathrm{Shift}_{\mathrm{B}}(I_{enc})$
Feature-wise modulation:
    $\boldsymbol{B} \triangleq \boldsymbol{B}_{\texttt{scale}} \odot \boldsymbol{B}_{\texttt{base}} + \boldsymbol{B}_{\texttt{shift}}$
Combine the multimodal knowledge: $\Phi_W = \boldsymbol{A} \times \boldsymbol{B}$
```

training on base classes while evaluating on the base and novel classes *separately* (Zhou et al. 2022a).

CoOp (Zhou et al. 2022b) learns text prompts neglecting input differences, therefore failing to generalize well beyond classes in training data. To alleviate such drawbacks, CoCoOp (Zhou et al. 2022a) proposes image-conditioned text prompts to impute novel class knowledge into prompts, and MaPLe (Khattak et al. 2023) injects tokens in both the vision and language branches which efficiently transition the novel class knowledge into prompts. In contrast to these approaches, visual prompt-based techniques lack an efficient means to integrate novel class knowledge.

We address this limitation by embedding novel-class knowledge into the visual prompts on the fly and without the need for retraining.The Kronecker product encapsulates information, eliminating the necessity for additional learning (Gao, Wang, and Ji 2020; Schwartz, Haley, and Tyers 2022; Demir, Lienen, and Ngonga Ngomo 2022; Jin, Kolda, and Ward 2021). The underlying idea is to employ the similarity between novel classes and base classes to refine $\boldsymbol{A}$. Recall that $\boldsymbol{A} = \boldsymbol{M}\boldsymbol{A}_{\texttt{base}}$. This can be understood as $\boldsymbol{A}$ being a linear combination of the semantic information captured in $\boldsymbol{A}_{\texttt{base}}$. In the presence of novel classes, we first encode a notion of similarity between base and novel classes by

$$\mathrm{T}_{\mathrm{enc}}^{\mathrm{K}_{\mathrm{novel}}} \otimes \boldsymbol{A} = \begin{pmatrix} t_{11}\boldsymbol{A} & \cdots & t_{1d}\boldsymbol{A} \\ \vdots & \ddots & \vdots \\ t_{K_{novel}1}\boldsymbol{A} & \cdots & t_{K_{novel}d}\boldsymbol{A} \end{pmatrix}, \tag{6}$$

where $\otimes$ denotes the Kronecker product. The resulting product will exist in $\mathbb{R}^{(aK_{novel})\times(rd)}$, representing the projection of each novel class on all the base classes. To obtain a compact and coherent embedding representation between base and novel classes, we transform this class-wise projection into $\mathbb{R}^{a\times K_{novel}\times r\times d}$. Subsequently, we compute the mean along VLM features(d) and the number of novel classes (K_{novel}). The averaging operation serves to align the base classes with a more unified representation of novel classes.

3 Related Works

In this section, first, we will introduce VLMs and their constraints for adapting to new tasks, then we will discuss existing prompt learning methods, and finally, we will explore how MR is used in repurposing PTMs for diverse domains tasks.

3.1 Pretrained Vision-Language Models

VLMs such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), Flamingo(Alayrac et al. 2022), Flava (Singh et al. 2022) and LiT (Zhai et al. 2022) have demonstrated exceptional performance on a variety of tasks, including few-shot and zero-shot image recognition. These models learn to align the vision-language representations on a web-scale training dataset. Although pretrained models offer a strong foundation for a wide range of tasks, efficiently adapting them to downstream tasks is still a challenging research problem. The difficulty is exacerbated when the downstream task requires specialized context, interpretable representations, or access to the model is forbidden (Mokady, Hertz, and Bermano 2021; Jiang, Liu, and Zheng 2022; Shu et al. 2023; Maus et al. 2023). Furthermore, Kumar *et al.* showed finetuning overparameterized models can yield detrimental results compared to linear probing (i.e., tuning the head while keeping lower layers frozen) when addressing out-of-distribution downstream tasks (Kumar et al. 2021).

3.2 Prompt Learning in VLMs

Standard finetuning and linear probing are common approaches to adapting VLMs to downstream tasks. However, such finetuning causes adverse effects due to the loss of embedded knowledge and poor adaptation techniques (Wortsman et al. 2022). There is a significant body of work in natural language processing (NLP) that focuses on learning effective prompts to adapt a large language model to downstream tasks (Sanh et al. 2021; Houlsby et al. 2019; Brown et al. 2020b; Chen et al. 2022). Inspired by the success of prompt learning in NLP, several recent studies explored prompt learning methods in the context of large-scale VLMs. Visual Prompt Tuning (VPT) learns the prefix prompts in encoder layers or embedding layer (Jia et al. 2022), while (Khattak et al. 2023) proposes injection of learnable tokens in both vision and text encoder layers and couples them with a learnable function. Visual Prompting (VP) investigated input pixel space prompt tuning for pretrained vision and VLMs (Bahng et al. 2022). VP learns a fixed input agnostic perturbation and attaches it to the original images, hence adapting a pretrained model to new tasks without modifying the model parameters. Bar et al. uses the inpainting method as visual prompting. Context Optimization (CoOp) (Zhou et al. 2022b) optimizes a set of context vectors for the text encoder of CLIP, while Conditional Context Optimization (CoCoOp) (Zhou et al. 2022a) generalizes

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	RESISC	CLEVR	UCF	*Avg.*	*Win*
CLIP	89.3	88.9	65.6	70.4	**89.2**	27.1	65.2	46.0	54.1	65.5	23.4	69.8	62.75	1
VP	94.2	90.2	66.9	86.9	81.8	31.8	67.1	61.9	**90.8**	81.4	40.8	74.2	71.26	1
LaViP (Ours)	**95.0**	**91.2**	**77.8**	**96.3**	82.5	**43.2**	**71.1**	**68.8**	86.1	**85.6**	**46.5**	**81.3**	**74.59**	10

Table 1: Comparison with visual prompting method on few-shot transfer learning. LaViP learns language-driven input-aware visual prompts and exhibits robust performance on 10 in 12 recognition datasets with training accelerated by *more than* 3×. *Win* indicates how many cases LaViP outperforms previous methods.

CoOp to unseen classes by conditioning the text prompt optimization on image instances. (Lin et al. 2023) suggests cross-modal adaptation by repurposing class names as one-shot training examples, (Lu et al. 2022) proposes an ensemble of learnable prompts. (Menon and Vondrick 2023; Zhang et al. 2023; Dunlap et al. 2022) showcase how language can be effectively employed to strengthen the adaptation of pretrained vision models to novel domains.

We argue that generating input-agnostic prompts with unimodal knowledge is a suboptimal approach. Considering the large-scale pretraining of VLMs, prompting methods should adeptly utilize the embedded multimodal knowledge to efficiently address new tasks. Further, we underscore the importance of prompting methods being agnostic to the underlying architecture of PTMs. For instance, VPT and MaPLe have successfully adapted ViT encoders through prefix learning. However, these methods lack comprehensive evidence of how their solutions perform across diverse backbone architectures.

3.3 Model Reprogramming

By deriving motivation from adversarial attacks Elsayed, Goodfellow, and Sohl-Dickstein (2018) proposed Adversarial Reprogramming(AR) to repurpose a pretrained model to perform on a new domain. This led to a new learning paradigm called model reprogramming (MR) for transfer learning. We provide some notable examples below. Vinod, Chen, and Das (2020) repurposed a language model to predict biochemical sequences; Tsai, Chen, and Ho (2020) proposed BAR to reprogram an ImageNet model for complex bio-medical under a black-box setting; Yen et al. (2021) used an attention-based RNN speech model for low-resource spoken command recognition; Yang, Tsai, and Chen (2021) reprogrammed a speech model for time-series prediction and Neekhara et al. (2022) reprogrammed a vision model to classify text sentences and DNA sequences. (Oh et al. 2023) extended BAR by generating input-aware visual prompts through an external encoder-decoder model for limited data recognition. To the best of our knowledge, we are pioneering to design of language-grounded visual prompts to reprogram the visual encoder of a VLM. In contrast to previous MR methods that primarily focused on repurposing PTMs using unimodality, our contribution lies in harnessing the power of multimodality to enhance context knowledge during adaptation.

4 Results

We first provide the experimental setup in § 4.1. Next, § 4.2 presents the comparison between LaViP and previous methods. § 4.3 provides the result for the base-to-novel generalization task and § 4.4 provides the results for whole-dataset training.

4.1 Experimental Setup

We extensively evaluate LaViP capability on 12 benchmark datasets (refer Appendix B.3[2]) under three distinct scenarios. First, its transferability in limited data settings is assessed through few-shot learning, where it learns from 16-shots for training and 4-shots for validation. Next, its generalizability is examined by evaluating its ability to learn from base classes and apply that knowledge to unseen novel classes. Finally, we use the full dataset for training, testing and validation. In this paper, we use CLIP ViT-B/16 (Radford et al. 2021) for few-shot learning and base-to-novel generalization, and CLIP ViT-B/32 for transfer learning as the pretrained VL model due to its strong zero-shot generalization capability. More details are provided in Appendix B[2].

4.2 Few-Shot Learning

Table 1 presents the performance of LaViP in a few-shot transfer setting across 12 recognition datasets. We compare our results against CLIP Zero-shot (ZS), and the previous pixel-space reprogramming method. LaViP outperforms ZS on 11 datasets, exhibiting a substantial gain of 11.84%. Additionally, when comparing to VP (Bahng et al. 2022) on 11 datasets, achieving a gain of 3.3% in performance and more than a 3× faster convergence. Furthermore, Table 1 shows that when the domain shifts from generic to rare concepts, LaViP consistently demonstrates higher performance in comparison to CLIP. This highlights the effectiveness of incorporating language guidance in enhancing modality alignment, particularly in cases where concepts can be explicitly described.

4.3 Base-to-Novel Generalization

Table 2 presents the performance of LaViP in the base-to-novel generalization setting, evaluated across 10 recognition datasets. We compare LaViP against a lineup of benchmarks, including CLIP Zero-shot(ZS), CoOP, CoCoOp and MaPLe. Relative to CoCoOp, LaViP exhibits a stronger performance

[2] Appendices available at: https://arxiv.org/abs/2312.10945

	CLIP			CoOp			CoCoOp			**LaViP**		
Dataset	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**	**Base**	**Novel**	**HM**
Caltech101	96.84	94.00	95.40	**98.00**	89.81	93.73	97.96	93.81	**95.84**	97.63	**93.45**	95.49
DTD	53.24	**59.90**	56.37	79.44	41.18	54.24	77.01	56.00	64.85	**80.05**	58.01	**67.27**
EuroSAT	56.48	64.05	60.03	92.19	54.74	68.69	87.49	60.04	71.21	**92.53**	**82.31**	**87.12**
FGVCAircraft	27.19	**36.29**	31.09	**40.44**	22.30	28.75	33.41	23.71	27.74	37.25	34.03	**35.57**
Food101	90.10	91.22	90.66	88.33	82.26	85.19	**90.70**	**91.29**	**90.99**	86.19	91.28	88.66
OxfordPets	91.17	97.26	94.12	93.67	95.29	94.47	**95.20**	**97.69**	**96.43**	92.45	97.22	94.78
SUN397	69.36	75.35	72.23	**80.60**	65.89	72.51	79.74	**76.86**	**78.27**	76.47	73.25	74.82
Flowers102	72.08	**77.80**	74.83	**97.60**	59.67	74.06	94.87	71.75	81.71	96.96	76.34	**85.25**
UCF101	70.53	**77.50**	73.85	**84.69**	56.05	67.46	82.33	73.45	77.64	83.83	76.46	**79.97**
StanfordCars	63.37	**74.89**	68.65	**78.12**	60.40	68.13	70.49	73.59	72.01	73.63	74.63	**74.13**
Average	69.04	74.83	71.72	**83.31**	62.76	70.72	80.92	71.82	75.67	81.7	**75.70**	**78.31**

Table 2: Performance of LaViP on base-to-novel generalization across 10 recognition datasets. LaViP demonstrates competitive generalization performance over CoOp and CoCoOp with an absolute gain of *2.64%*. HM: The harmonic mean of base class acc. and novel class acc.

Method	Pets	Flowers	Food	SUN	DTD	EuroSAT	RESISC	CLEVR	UCF	*Avg.*
ZS	88.3	67.4	85.2	62.6	44.4	42.2	56.6	25.8	65.2	59.75
CLIP + LP	89.2	96.9	**84.6**	**75.0**	**74.6**	95.3	**92.3**	66.0	**83.3**	84.13
VP	85.0	70.3	78.9	60.6	57.1	**96.4**	84.5	**81.4**	66.1	75.72
LaViP (Ours)	**89.6**	96.7	83.2	71.5	72.9	96.3	91.0	67.0	81.9	80.99

Table 3: Performance across 9 recognition dataset using CLIP Zero-shot(ZS), Linear Probe(LP), Visual Prompting(VP) and LaViP with ViT-B/32 backbone.

across both base and novel concepts, yielding absolute gains of 0.78% and 3.88% respectively. With the context-aware knowledge diffused through Kronecker product, LaViP as a strong competitor surpasses CoCoOp in 6/10 datasets and slightly trails in two datasets. When taking into account both base and novel classes, LaViP shows an absolute average gain of with gain of 2.64% compared to CoOp and CoCoOp.

MaPLe (Khattak et al. 2023), the current SOTA has outperformed in many studied datasets. However, unlike MaPLe, other algorithms maintain the original structure of the foundational model, thus enabling adaptation via APIs and cases where for ethical constraints, accessing the structure and weights of the foundation model is not possible. Without any structural adaptation of the pretrained model, LaViP trailed MaPLe by only 0.44% in **HM**, showcasing a competitive performance. It even marginally outperformed in classifying novel classes. In comparison with CLIP on novel classes, CoCoOp improves 3/10 classes, leading to a decrease in the average novel accuracy from 74.83% to 71.82%. LaViP only improves accuracy in 2 out of 10 datasets compared to CLIP for new classes. However, it positively impacts the average accuracy, elevating it from 74.83% to 75.70%. This sustains its position as a robust competitor. CoOp exhibits limited generalization capability to novel classes, a deficiency that CoCoOp strives to address by contextualizing text prompts based on image instances. It shows substantial improvement in novel class recognition. However, it manages to outperform 2/10 base classes with a decrease in average performance of 2.39%. LaViP's language integration exhibits competitive performance in the base class, with only a 1.4% decrease in average performance. Despite marginal improvements compared to CoCoOp, it is important to differentiate between high-dimensional, complex image data and structured textual data. This divergence affects learning speed and effectiveness, especially with limited data. Given the complexities inherent in the visual domain, an approach must be parameter-efficient and context-aware. This dual requirement aligns with the fundamental characteristic of LaViP.

Moreover, from Table 2 we can conclude that as the domain shift increases from the pretraining dataset, LaViP exhibits increasing performance compared to CLIP, CoOp and CoCoOp. This emphasizes the impact of language context in designing visual prompts.

4.4 Transfer Learning

The summarized findings of transfer learning are presented in Table 3. To provide a comprehensive evaluation, we draw comparisons between our results and those of CLIP Zero-shot (ZS), CLIP Linear Probe (LP), and VP, all using the ViT-B/32 CLIP model with the same hyperparameters as those used in few-shot learning. It indicates that LaViP consistently outperforms VP by a notably wider margin across 7/9 recognition datasets. This substantial improvement is

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	DTD	RESISC	UCF	SUN	*Avg.*
ZS(CLIP)	89.3	88.9	**65.6**	70.4	**89.2**	**27.1**	46.0	**65.5**	**69.8**	62.6	67.44
VP w/SPSA-GC	89.4	87.1	56.6	67.0	80.4	23.8	44.5	61.3	64.6	61.2	63.59
BAR	**93.8**	88.6	63.0	**71.2**	84.5	24.5	**47.0**	65.3	64.2	62.4	66.45
BlackVIP	93.7	**89.7**	**65.6**	70.6	86.6	25.0	45.2	64.5	69.1	**64.7**	**67.47**
BlackLaViP (Ours)	92.3	89.3	63.3	68.8	84.6	23.9	46.1	61.8	65.6	62.2	65.77
%	98.5	99.6	96.5	97.5	97.7	95.6	101.2	95.8	95.1	96.1	97.49

Table 4: Comparison with state-of-the-art methods on few-shot transfer learning in a black-box setting. % indicates percentage of BlackVIP score achieved with *more than* $15\times$ faster optimization.

coupled with an optimization process that is three times more efficient. In contrast with the results obtained from CLIP Linear Probe, VP shows enhancement in 2 out of 9 datasets, albeit accompanied by an average accuracy decrease of 8.91%. Comparatively, while LaViP only enhances performance in 1 out of 9 datasets, it achieves an average accuracy drop of 3.14%. These observations highlight the modality alignment enhanced by the LaViP.

5 Ablation Studies

5.1 Learning in Gradient-Free Environment

We adapt our algorithm in a gradient-free environment to understand the effectiveness of language integration. We proceed to evaluate **BlackLaViP**, the gradient-free variant of LaViP, by employing the SPSA algorithm (Spall 1992, 2000). Table 4 displays the average performance of BlackLaViP on 10 recognition datasets, using a few-shot approach. BlackVIP (Oh et al. 2023), the current SOTA uses SPSA with an external model to generate input-aware prompts. Though BlackLaViP doesn't outperform BlackVIP, an intriguing observation emerges from our experiment. Remarkably, BlackLaViP attains 95% of the performance of BlackVIP with a convergence rate that is over $15\times$ faster (Appendix C[2]).

5.2 Impact of Hyperparameters (a, b, r)

VP requires generating a padding of size $\theta = 2\mathtt{p}\mathtt{C}(\mathtt{H} + \mathtt{W} - 2\mathtt{p})$ as the visual prompt. We propose a low-rank formulation to generate the prompt, which efficiently reduces the generator size by a factor of 4 when $r = 32$, 2 when $r = 64$, and 1.2 when $r = 96$. The parameter r (rank of matrices in generating the prompt) can be understood as an inductive bias, regularizing the design. Empirically, we observed that LaViP performed robustly if r was chosen within a reasonable range (not too small, *e.g.* $r \in [16, 96]$). Furthermore, LaVIP robustly performs for varying (a, b) which creates padding of size $p = 20$ to $p = 50$. Additional results are provided in Appendix D[2].

6 Discussion

LaViP, a novel approach to visual prompting, harnesses the power of language to enhance pretrained models without the need for invasive finetuning. By merging textual knowledge into input prompts, LaViP steers models towards desired tasks, surpassing the limitations of previous methods in both accuracy and optimization. The few-shot capability is a direct result of preserving the foundation model. By aligning image and text via visual prompting and without altering the latent space of the foundation model, we capitalize on the generalization capability of the model. Its versatility shines across diverse tasks, requiring no individual finetuning efforts, and its privacy-preserving nature makes it ideal for interacting with APIs and proprietary software. The reprogramming methodology studied in this can work to provide increased user control over bias and fairness issues in pretraining. However, low-resolution images and highly diverse datasets present challenges. We hypothesize that the observed characteristic is due to context tokens failing to grasp semantic content or capture the full spectrum of classes. LaViP's performance is inherently influenced by the context tokens present in the prompt template. This naturally raises the question: *What advantages does learning text prompts offer compared to employing manually crafted templates in LaViP?*. Future research could delve into the direction of learning multimodal prompts with mutual synergy.

7 Conclusion

Adaptating large-scale VLMs(*e.g.* CLIP (Radford et al. 2021)) for new tasks is a challenging research problem due to a large number of tunable parameters. Despite stemming from distinct motivation, prompt learning and model reprogramming provide an efficient and scalable approach to drive VLMs to downstream tasks. To this end, existing visual prompting approaches learn input-agnostic prompts through unimodal knowledge. The perceptual diversity of the image domain makes a difficult to repurpose visual encoders in VLMs compared to text encoders, and these approaches require an external world model to provide context or a large number of iterations. Our work counters these assumptions by leveraging embedded multimodal knowledge within VLMs. Our approach seamlessly integrates these multimodal representations to generate adaptable visual prompts, thereby enhancing performance without compromising. Further, we propose an efficient strategy for generalizing visual prompting methods to unseen classes. Our method improves the few-shot transfer learning, generalization towards novel concepts and full-set transfer learning with varying domain shifts compared to the pretraining dataset.

Acknowledgments

Mehrtash Harandi gratefully acknowledges the support from the Australian Research Council (ARC), project DP230101176.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *ArXiv*, abs/2204.14198.

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

Bar, A.; Gandelsman, Y.; Darrell, T.; Globerson, A.; and Efros, A. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35: 25005–25017.

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

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

Chen, A.; Yao, Y.; Chen, P.-Y.; Zhang, Y.; and Liu, S. 2023. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19133–19143.

Chen, Y.; Liu, Y.; Dong, L.; Wang, S.; Zhu, C.; Zeng, M.; and Zhang, Y. 2022. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*.

Demir, C.; Lienen, J.; and Ngonga Ngomo, A.-C. 2022. Kronecker Decomposition for Knowledge Graph Embeddings. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, HT '22, 1–10. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392334.

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

Dunlap, L.; Mohri, C.; Guillory, D.; Zhang, H.; Darrell, T.; Gonzalez, J. E.; Raghunathan, A.; and Rohrbach, A. 2022. Using language to extend to unseen domains. In *The Eleventh International Conference on Learning Representations*.

Elsayed, G. F.; Goodfellow, I.; and Sohl-Dickstein, J. 2018. Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*.

Gao, H.; Wang, Z.; and Ji, S. 2020. Kronecker attention networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 229–237.

Gibson, E. J. 1969. Principles of perceptual learning and development.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*.

Jiang, J.; Liu, Z.; and Zheng, N. 2022. Finetuning pretrained vision-language models with correlation information bottleneck for robust visual question answering. *arXiv preprint arXiv:2209.06954*.

Jin, R.; Kolda, T. G.; and Ward, R. 2021. Faster johnson–lindenstrauss transforms via kronecker products. *Information and Inference: A Journal of the IMA*, 10(4): 1533–1562.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. MaPLe: Multi-Modal Prompt Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19113–19122.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kumar, A.; Raghunathan, A.; Jones, R. M.; Ma, T.; and Liang, P. 2021. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *International Conference on Learning Representations*.

Lin, Z.; Yu, S.; Kuang, Z.; Pathak, D.; and Ramanan, D. 2023. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19325–19337.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5206–5215.

Maus, N.; Chao, P.; Wong, E.; and Gardner, J. 2023. Adversarial prompting for black box foundation models. *arXiv preprint arXiv:2302.04237*.

Meltzoff, A. N.; and Borton, R. W. 1979. Intermodal matching by human neonates. *Nature*, 282(5737): 403–404.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. *ICLR*.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Neekhara, P.; Hussain, S.; Du, J.; Dubnov, S.; Koushanfar, F.; and McAuley, J. 2022. Cross-modal adversarial reprogramming. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2427–2435.

Oh, C.; Hwang, H.; Lee, H.-y.; Lim, Y.; Jung, G.; Jung, J.; Choi, H.; and Song, K. 2023. BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24224–24235.

Quiroga, R. Q.; Reddy, L.; Kreiman, G.; Koch, C.; and Fried, I. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045): 1102–1107.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Schwartz, L.; Haley, C.; and Tyers, F. 2022. How to encode arbitrarily complex morphology in word embeddings, no corpus needed. In *Proceedings of the first workshop on NLP applications to field linguistics*, 64–76. Gyeongju, Republic of Korea: International Conference on Computational Linguistics.

Shu, Y.; Guo, X.; Wu, J.; Wang, X.; Wang, J.; and Long, M. 2023. CLIPood: Generalizing CLIP to Out-of-Distributions. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 31716–31731. PMLR.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Spall, J. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3): 332–341.

Spall, J. 2000. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10): 1839–1853.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

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

Vinod, R.; Chen, P.-Y.; and Das, P. 2020. Reprogramming language models for molecular representation learning. *arXiv preprint arXiv:2012.03460*.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7959–7971.

Yang, C.-H. H.; Tsai, Y.-Y.; and Chen, P.-Y. 2021. Voice2series: Reprogramming acoustic models for time series classification. In *International conference on machine learning*, 11808–11819. PMLR.

Yen, H.; Ku, P.-J.; Yang, C.-H. H.; Hu, H.; Siniscalchi, S. M.; Chen, P.-Y.; and Tsao, Y. 2021. Neural model reprogramming with similarity based mapping for low-resource spoken command classification. *arXiv preprint arXiv:2110.03894*.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. LiT: Zero-Shot Transfer With Locked-Image Text Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18123–18133.

Zhang, R.; Hu, X.; Li, B.; Huang, S.; Deng, H.; Qiao, Y.; Gao, P.; and Li, H. 2023. Prompt, Generate, Then Cache: Cascade of Foundation Models Makes Strong Few-Shot Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15211–15222.

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

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