BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning

Changdae Oh[1] Hyeji Hwang[1] Hee-young Lee[2†] YongTaek Lim[1] Geunyoung Jung[1]
Jiyoung Jung[1] Hosik Choi[1] Kyungwoo Song[3‡]
[1]University of Seoul [2]Sungkyunkwan University [3]Yonsei University
changdae.oh@uos.ac.kr kyungwoo.song@yonsei.ac.kr

Abstract

With the surge of large-scale pre-trained models (PTMs), fine-tuning these models to numerous downstream tasks becomes a crucial problem. Consequently, parameter efficient transfer learning (PETL) of large models has grasped huge attention. While recent PETL methods showcase impressive performance, they rely on optimistic assumptions: 1) the entire parameter set of a PTM is available, and 2) a sufficiently large memory capacity for the fine-tuning is equipped. However, in most real-world applications, PTMs are served as a black-box API or proprietary software without explicit parameter accessibility. Besides, it is hard to meet a large memory requirement for modern PTMs. In this work, we propose black-box visual prompting (BlackVIP), which efficiently adapts the PTMs without knowledge about model architectures and parameters. BlackVIP has two components; 1) Coordinator and 2) simultaneous perturbation stochastic approximation with gradient correction (SPSA-GC). The Coordinator designs input-dependent image-shaped visual prompts, which improves few-shot adaptation and robustness on distribution/location shift. SPSA-GC efficiently estimates the gradient of a target model to update Coordinator. Extensive experiments on 16 datasets demonstrate that BlackVIP enables robust adaptation to diverse domains without accessing PTMs' parameters, with minimal memory requirements. Code: `https://github.com/changdaeoh/BlackVIP`

1. Introduction

Based on their excellent transferability, large-scale pre-trained models (PTMs) [8, 19, 63] have shown remarkable success on tasks from diverse domains and absorbed increasing attention in machine learning communities. By witnessing PTMs' success, Parameter-Efficient Transfer Learning (PETL) methods that efficiently utilize the PTMs are recently emerging. While the standard fine-tuning (FT) and its advanced variants [42, 84] update the entire or large portion of a PTM [18], PETL methods aim to achieve comparable performance to FT by optimizing a small number of learnable parameters.

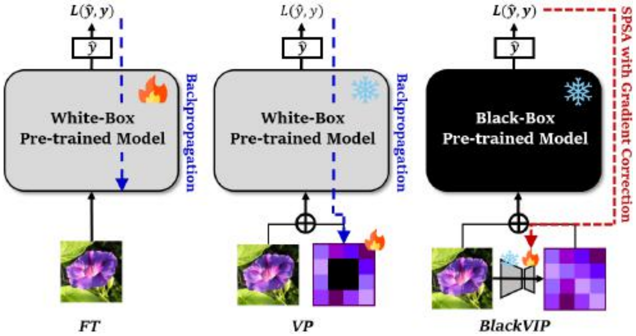


Figure 1. While FT updates the entire model, VP has a small number of parameters in the input pixel space. However, VP still requires a large memory capacity to optimize the parameters through backpropagation. Moreover, FT and VP are only feasible if the PTM's parameters are accessible. Meanwhile, BlackVIP does not assume the parameter-accessibility by adopting a black-box optimization (SPSA-GC) algorithm rather than relying on backpropagation. Besides, BlackVIP reparameterizes the visual prompt with a neural network and optimizes tiny parameters with SPSA-GC. Based on the above properties, BlackVIP can be widely adopted in realistic and resource-limited transfer learning scenarios.

Among them, *prompt-based approaches* [3, 5, 36, 44, 45] have been widely investigated from diverse research areas. For vision PTMs, Visual Prompt Tuning [36] injects a few additional learnable prompt tokens inside of ViT's [19] layers or embedding layer and only optimizes them. Bahng et al. [3] investigate visual prompting (VP), which adopts the learnable parameters on input pixel space as a visual prompt, while no additional modules are inserted into the pre-trained visual model. Besides, prompt learning methods for VLM are also actively studied [38, 90, 93, 94].

While existing PETL methods show impressive performance with few learnable parameters, they rely on two

†Work done at University of Seoul
‡Corresponding author; Work partly done at University of Seoul

optimistic assumptions. First, the previous PETL assumes that the full parameters of the PTM are accessible. However, many real-world AI applications are served as API and proprietary software, and they do not reveal the implementation-level information or full parameters due to commercial issues, e.g., violating model ownership. As a result, exploiting high-performing PTMs to specific downstream tasks not only in the white-box setting but also black-box setting (limited accessibility to the model's detail) is a crucial but unexplored problem. Second, existing methods require a large memory capacity. While PETL approaches have few learnable parameters, they require a large amount of memory for backpropagating the gradient throughout the large-scale PTM parameters to learnable parameters. Therefore, users who want to adopt a large-scale PTM should satisfy large memory requirements despite the small learnable parameters. Besides, if the users entrust PTM fine-tuning to the model owner with their specific data, data-privacy concerns will inevitably arise [85].

To alleviate the above unrealistic assumptions, we are pioneering ***black-box visual prompting*** **(BlackVIP)** approach, which enables the parameter-efficient transfer learning of pre-trained black-box vision models from the low-resource user perspective (illustrated in Figure 1). BlackVIP works based on the following two core components: 1) pixel space input-dependent visual prompting and 2) a stable zeroth-order optimization algorithm.

Firstly, we augment an input image by attaching an visual prompt per pixel. It is noted that input space prompting does not require the accessibility on parts of architecture [40, 90] or the first embedding layer [38, 93, 94] of PTM. While the previous works only introduce a pixel-level prompt to a small fraction of the fixed area, such as outside of the image [3], BlackVIP designs the prompt with the same shape as the original given image to cover the entire image view. Therefore, our prompt has a higher capability and can flexibly change the semantics of the original image. In addition, we reparameterize the prompt with a neural network. Specifically, we propose the ***Coordinator***, an asymmetric autoencoder-style network that receives the original image and produces a corresponding visual prompt for each individual image. As a result, Coordinator automatically designs each prompt conditioned on the input rather than the shared manual design of a previous work [3]. By optimizing the reparameterized model instead of the prompt itself, we greatly reduce the number of parameters (from 69K of VP [3] to 9K) so that suitable for black-box optimization.

Next, unlike other PETL approaches, BlackVIP adopts a zeroth-order optimization (ZOO) that estimates the zeroth-order gradient for the coordinator update to relax the assumption that requires access to the huge PTM parameters to optimize the prompt via backpropagation. Therefore, BlackVIP significantly reduces the required memory for fine-tuning. Besides, we present a new ZOO algorithm, ***Simultaneous Perturbation Stochastic Approximation with Gradient Correction*** **(SPSA-GC)** based on (SPSA) [69]. SPSA-GC first estimates the gradient of the target black-box model based on the output difference of perturbed parameters and then corrects the initial estimates in a momentum-based look-ahead manner. By integrating the Coordinator and SPSA-GC, BlackVIP achieves significant performance improvement over baselines.

Our main contributions are summarized as follows:

- To our best knowledge, this is the first paper that explores the input-dependent visual prompting on black-box settings. For this, we devise Coordinator, which reparameterizes the prompt as an autoencoder to handle the input-dependent prompt with tiny parameters.

- We propose a new ZOO algorithm, SPSA-GC, that gives look-ahead corrections to the SPSA's estimated gradient resulting in boosted performance.

- Based on Coordinator and SPSA-GC, BlackVIP adapts the PTM to downstream tasks without parameter access and large memory capacity. We extensively validate BlackVIP on 16 datasets and demonstrate its effectiveness regarding few-shot adaptability and robustness on distribution/object-location shift.

2. Related Works

2.1. Pre-trained Vision Models

Over the past decade, the pre-train and fine-tune paradigm has become the de-facto standard using deep neural networks. Beyond the label supervision [31, 65], self-supervised learning (SSL) [12, 25, 29, 30, 87, 91] approaches that do not rely on human-annotated labels hit the machine learning community. SSL approaches can roughly be categorized into discriminative and generative approaches. Discriminative SSL methods [12, 25, 30, 91] learn the embeddings by enforcing closeness and/or distantness on the pairwise distance structure among the augmented training samples. Meanwhile, the recently emerging generative SSL methods [4, 29, 87] are based on *masked image modeling*, which supervises the model by encoding and reconstructing the partially masked individual images. SSL approaches are appealing not only due to their label-free training regime but also produce a more transferable representation [9, 29, 47] getting over the pre-defined label category.

Moreover, fuelled by pre-training with rich semantic structures from image-caption pairs of the web-scale dataset, visual-language pre-trained models [35, 63, 66, 89] recently showed surprising performance on the zero-shot transfer and few-shot adaptation. Based on their high transferability, they are being adopted for numerous downstream tasks from diverse domains. Meanwhile, the number of parameters of PTMs has increased continuously, showing the performance improvement proportional to the number of parameters [39]. However, large models require sufficiently large memory capacity in the fine-tuning stage. Besides, it is commonly impossible to access the PTMs' parameters in public. Therefore, we propose a new fine-tuning method that does not require both knowledge about model parameters and a large amount of memory.

2.2. Parameter-Efficient Transfer Learning

To adapt the large-scale PTMs to targeted downstream tasks, Parameter-Efficient Transfer Learning (PETL) methods pursue fine-tuning of a small subset of large PTMs, while achieving competitive performance compared to full fine-tuning. Recently, diverse PETL approaches have emerged in the NLP domain, such as adapter [33, 62] and prompt learning [44, 45].

Motivated by the promising results of PETL in NLP, there have been many efforts to realize PETL in vision or vision-language fields. For the case of adapter-based methods, AdaptFormer [11], and CLIP-Adapter [23] insert a few learnable modules inside of the vision encoder (e.g., ViT [19]) or on top of both the vision and text encoder, respectively. In the case of prompt-based approaches, CoOp [94] introduces the continuous text prompt into the text encoder of a VLM, and Conditional CoOp (CoCoOp) [93] extends CoOp to an input-dependent version. Besides, Jia et al. [36] propose the Visual Prompt Tuning (VPT) that governs learnable visual tokens to the embedding layer (VPT-Shallow) or several encoder layers (VPT-Deep) of ViT. Bahng et al. [3] explore the Visual Prompting (VP) approach, which introduces a learnable prompt in the input space, not into the embedding space or model's building blocks. Then the prompt is attached to the image in the fixed restricted region. Among them, VP is especially attractive because it does not require full accessibility of the model architecture and parameters during the inference phase, which motivates us to investigate the *black-box visual prompting*.

However, we argue that the existing visual prompt approaches can be advanced from three perspectives. (1) *From white-box to black-box*: to be utilized for a variety of real-world applications, PETL should be able to deal with black-box PTMs that are served via API or proprietary software; for this, black-box optimization methods are required rather than backpropagation in previous works. (2) *Towards input-dependent visual prompt*: the visual features of individual images are distinct even though the images share the same class label; therefore, the input-dependent prompt is necessary. (3) *beyond the manual prompt design*: the prompt of VP is manually designed like a frame- or square-shaped and attached to a restricted region; this limited flexibility induces a sub-optimal prompt on challenging generalization scenario (refer to the Sec 5.2). To this end, we propose BlackVIP, which adopts ZOO rather than backpropagation and automatically designs input-dependent prompts over the entire image region. Table 1 summarizes the comparison between the previous methods and ours.

Table 1. Prompt-based PETL. Loc. denotes the prompt location.

Method	Grad-free	Prompt	Loc.	Input-dependent
CLIP ZS [63]	✓	L	*input*	✗
CoOp [94]	✗	L	*emb*	✗
CoCoOp [93]	✗	L	*emb*	✓
VPT [36]	✗	V	*emb*	✗
VP [3]	✗	V	*input*	✗
BAR [79]	✓	V	*input*	✗
BlackVIP (Ours)	✓	V	*input*	✓

2.3. Black-Box Optimization

Numerous high-performing artificial intelligence models have been deployed, and many custom services are based on the API or proprietary software. There are several works on NLP field that fine-tune the large language model via black-box optimization [17, 76, 77]. Besides, black-box adversarial reprogramming (BAR) [79] had been proposed to re-purpose the ImageNet [16] pre-trained vision model to a medical image classifier.

The previous works on black-box attack and optimization utilize ZOO algorithms or derivative-free optimization algorithms for parameter updates. BAR [79] adopts a one-sided approximation gradient estimator, but we find that the one-sided estimator shows inaccurate gradient approximations empirically. BBT [77], and BBTv2 [76] adopt Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [27, 28], and RLPrompt [17] uses reinforcement learning (Soft Q-Learning [26]) to optimize the discrete prompts. However, It has been known that derivative-free optimizations (e.g. evolutionary optimization) are hard to solve large-scale problems and do not guarantee convergence [48]. Besides, reinforcement learning algorithms are notorious for their unstable optimization, and high variance [92].

In this work, we adopt the Simultaneous Perturbation Stochastic Approximation (SPSA) [69] as a ZOO algorithm. It is known that SPSA is efficient at high-dimensional gradient approximation problems [69, 73]. Besides, SPSA theoretically guarantees convergence, and the convergence error is linearly upper bounded by the parameter dimension [69]. While SPSA is designed to estimate high-dimensional gradients efficiently, we found that SPSA-based neural network optimization still requires many queries in practice. Therefore, we propose SPSA with Gradient Correction (SPSA-GC) that corrects the approximated gradients to enhance the convergence speed. To our best knowledge, this is the first work exploring the ZOO-based black-box optimization to large PTMs for general-purpose adaptation (rather than a specific domain [79]).

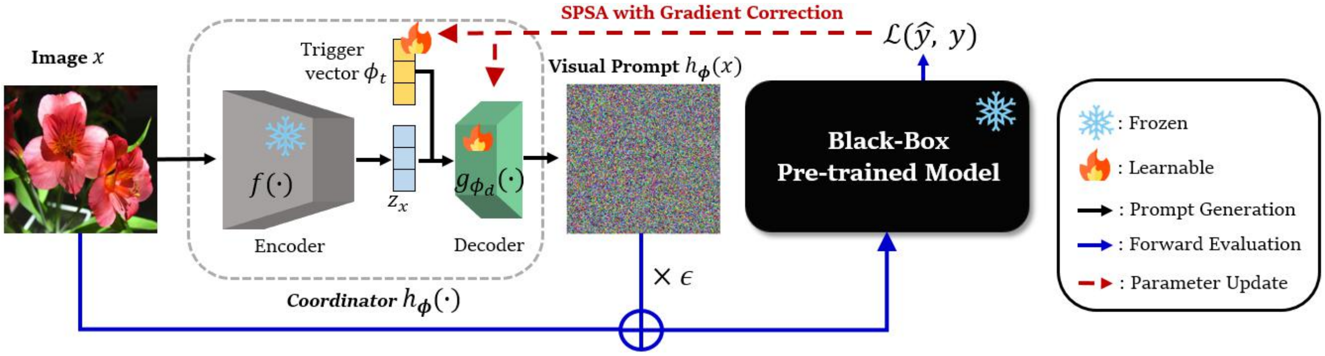


Figure 2. BlackVIP equips an input-dependent prompt designer (Coordinator) and an accurate gradient estimation algorithm (SPSA-GC).

3. Preliminary

We first present an outline of *adversarial reprogramming* and *visual prompting*, that originated from distinct motivations, but closely related topics. Elasyed et al. [20] presented *adversarial reprogramming* (AR) inspired by adversarial attack [24, 59, 88]. The goal of AR is repurposing the pre-trained model to perform a new task. Let $x \in \mathbb{R}^{k\times k\times 3}$ be a downsized image from the adversarial target dataset, and $\tilde{x} \in \mathbb{R}^{n\times n\times 3}$ is a random image from pre-train dataset or a zero-padded tensor that includes x in the center of image $\tilde{x}$, where $k < n$. Given the target class of adversarial task $y_{adv} \in \{1, ..., C_{tar}\}$, the AR is formulated as:

$$\underset{W}{\arg\min}(-\log P_{\theta;W}(h(y_{adv})|\tilde{x}_{adv}) + ||W||_F)$$

Here, θ is the pre-trained model parameters, and the adversarial image is constructed as $\tilde{x}_{adv} = \tilde{x} + \tanh(W \odot M)$, and $W \in \mathbb{R}^{n\times n\times 3}$ is the adversarial program that is optimized, where n is the image width of a pre-train dataset, M is an optional mask for better visualization of the embedded target image, and $\odot$ denotes the element-wise multiplication. Given a pre-defined hard-coded mapping $h(\cdot)$ that maps labels from an adversarial task to labels of a pre-train dataset, AR reprograms the model via learned perturbation without architectural change. The vulnerability of neural networks to adversarial examples has inspired many works that use the AR approach from transfer learning perspectives [10, 52, 53, 79] i.e., *model reprogramming*.

Meanwhile, motivated by the remarkable success of the *prompting* paradigm on NLP, Bahng et al. [3] are the first to explore the input pixel space *visual prompting* (VP) approach for pre-trained vision and vision-language models. By learning pixel-style prompts (i.e., perturbation) attached to the input images, VP adapts the frozen PTM to targeted downstream tasks without modifying on model architecture. Given the input image x and corresponding label y, the learning objective of VP is as follows:

$$\underset{\phi}{\arg\min} -\log P_{\theta;\phi}(y|x+\phi)$$

where θ and ϕ are the PTM parameters and visual prompt, respectively. At the inference phase, VP employs the shared prompt (input-independent) for all images. It is noted that ϕ is attached to the fixed location, e.g., the outer part of the image like a frame by default.

Though AR and VP use different terms and stem from distinct motivations, they share the general idea: adapt a PTM to perform new tasks without modifying the model architecture. This paper aligns with AR and VP, but broadens and improves them for more realistic environments.

4. Methodology

We introduce our novel input-dependent prompt generation module, *Coordinator* (in Section 4.1). Then, we explain the end-to-end framework of BlackVIP with the new ZOO algorithm, *SPSA-GC* (in Section 4.2). Figure 2 illustrates the overall framework of BlackVIP.

4.1. Coordinator: Prompt Reparameterization

Our learning objective is to minimize the downstream task loss by adapting the frozen PTM via input space prompt optimization. Given a frozen prediction model $P_\theta(y|x)$, and perturbed image $\tilde{x}$ with prompt corresponding to the label y, the training objective is formulated as:

$$\underset{\boldsymbol{\phi}}{\arg\min} -\log P_{\theta;\boldsymbol{\phi}}(y|\tilde{x})$$

While VP and AR optimize the input space visual prompt directly, we reparameterize the visual prompt to the prompt generation network $h_{\boldsymbol{\phi}}(\cdot)$ parameterized by $\boldsymbol{\phi} = \{\phi_d, \phi_t\} \in \mathbb{R}^d$. Specifically, we build a novel autoencoder-style network named Coordinator composed of a frozen encoder $f(\cdot)$ which is pre-trained on ImageNet [16] by self-supervised learning (SSL) objective and followed by an extremely light-weight learnable decoder $g_{\phi_d}(\cdot)$. Though the encoder can also be a supervised counterpart or light-weight learnable network, we adopt the SSL pre-trained encoder for the following three reasons: 1) It has been widely substantiated that self-supervised representation contains the multiple discriminative features and spatial information [9, 21, 29, 34, 46, 47, 58], so it is more helpful to use SSL pre-trained encoder than label-supervised encoder for robustly performing on diverse downstream tasks. 2) ImageNet pre-trained encoders are currently well-popularized [1, 61, 82, 83], so they can be easily adopted by local users, and does not hurt our realistic experimental setting. 3) By using the frozen pre-trained encoder, we significantly reduce the number of learnable parameters. The reduced low-dimensional parameters encourage efficient gradient approximation. Consequently, the image equipped with a prompt (prompted image) is constructed as follows:

$$\tilde{x} = \text{clip}(x + \epsilon h_{\boldsymbol{\phi}}(x))$$
$$h_{\boldsymbol{\phi}}(x) = g_{\phi_d}(z_x, \phi_t)$$

where $z_x = f(x)$ is the feature vector of x from the frozen SSL encoder $f(\cdot)$, and $\epsilon \in [0, 1]$ is a hyperparameter that controls the intensity of visual prompt. Here, ϕ_t is a task-specific *prompt trigger vector* that is jointly optimized with decoder parameter ϕ_d. We concatenate it with z_x and then reshape them into a 3D feature map to feed into the convolutional decoder. As a result, the instance-specific rich semantic representation z_x and the task-specific prompt trigger vector ϕ_t are merged to design a valid visual prompt $h_{\phi}(x)$ for a given image. Similar to [53], the prompted image is bounded to a valid RGB scale via pixel-wise clipping.

Unlike previous visual prompts (e.g., VP) or adversarial programs (e.g., BAR), our BlackVIP automatically designs the input-dependent prompts with the same shape as the original images; therefore, it has a higher capability to change the semantics of images if necessary. Thanks to this flexibility, BlackVIP can cover more diverse tasks and be robust to challenging scenarios, e.g., distribution shift.

4.2. End-to-End Black-Box Visual Prompting

Unlike other PETL approaches that assume the accessibility to the architecture and/or parameters of the PTM, we consider the PTM as a black-box predictor that gives only a prediction output (i.e. logit) for a given input image query. In this black-box setting, we adopt the ZOO algorithm, SPSA, with our considerate modification to optimize our Coordinator without the oracle true gradient.

SPSA Spall et al. proposed Simultaneous Perturbation Stochastic Approximation (SPSA) [69, 72] that approximates the high-dimensional gradient efficiently. Given the positive decaying sequences of $a_i > 0$ and $c_i \in [0, 1]$, the gradient approximation, $\hat{g}$, and single-step parameter update of SPSA is described as follows:

$$\hat{g}_i(\boldsymbol{\phi}_i) = \frac{L(\boldsymbol{\phi}_i + c_i\Delta_i) - L(\boldsymbol{\phi}_i - c_i\Delta_i)}{2c_i}\Delta_i^{-1} \quad (1)$$

$$\boldsymbol{\phi}_{i+1} = \boldsymbol{\phi}_i - a_i\hat{g}_i(\boldsymbol{\phi}_i) \quad (2)$$

where L is an objective function, $\boldsymbol{\phi}_i \in \mathbb{R}^d$ is d-dimensional learnable parameters, and $\Delta_i \in \mathbb{R}^d$ is a i^{th}-step random perturbation vector, sampled from mean-zero distributions that satisfy finite inverse momentum condition [69, 74] such as Rademacher and Segmented Uniform distribution. With only two forward evaluations, i.e., querying twice to the API service model, SPSA parses the learning signal (estimate gradient) from the model's output difference, and we can optimize the parameters of Coordinator $\boldsymbol{\phi}$ to design the proper visual prompt for a given input.

Algorithm 1 BlackVIP algorithm

```
Require: Downstream dataset $\mathcal{D}$, pre-trained model $P_\theta$,
  Coordinator $h$ with encoder $f$ and prompt decoder $g$, is
  parameterized by $\boldsymbol{\phi}_i = \{\phi_{d,i}, \phi_{t,i}\}$, SPSA-GC decaying
  parameters $\{a_i, c_i\}$, and smoothing parameter $\beta$, prompt
  intensity $\epsilon$, and training iteration $R$.
  // Initialize $\boldsymbol{\phi}_1 = \{\phi_{d,1}, \phi_{t,1}\}$, $\{a_1, c_1\}$ and $m_1$
  for $i$ in 1 to $R$ do
      // Parse a batch $(x, y) \sim \mathcal{D}$ and design the prompt
      $h_{\boldsymbol{\phi}_i}(x) = g_{\phi_{d,i}}(f(x), \phi_{t,i})$
      $\tilde{x} = \text{clip}(x + \epsilon h_{\boldsymbol{\phi}_i}(x))$
      // Draw a sample $\Delta_i$, set $c_i$, and estimate the gradient
      $L(\boldsymbol{\phi}_i) := -\log P_{\theta;\boldsymbol{\phi}_i}(y|\tilde{x})$
      $\hat{g}_i(\boldsymbol{\phi}_i) = (L(\boldsymbol{\phi}_i + c_i\Delta_i) - L(\boldsymbol{\phi}_i - c_i\Delta_i))(2c_i\Delta_i)^{-1}$
      // Set $a_i$, and update parameters
      $m_{i+1} = \beta m_i - a_i\hat{g}_i(\boldsymbol{\phi}_i + \beta m_i)$
      $\boldsymbol{\phi}_{i+1} = \boldsymbol{\phi}_i + m_{i+1}$
  end for
```

SPSA with Gradient Correction Although the standard form of SPSA works well in myriad applications [6, 67, 75], like other ZOO algorithms, it may suffer slow convergence in practice [70, 71], and the problem gets even bigger on the high-dimensional problem setting such as neural networks' optimization. We speculate that the source of slow convergence is its noisy gradient estimation from the poor direction of random perturbations or intrinsic data noise. To mitigate this estimation noise, inspired by Nesterov's accelerated gradient (NAG) [54], we improve the parameter update rule in Eq. 2 as below:

$$\boldsymbol{\phi}_{i+1} = \boldsymbol{\phi}_i + m_{i+1} \quad (3)$$
$$m_{i+1} = \beta m_i - a_i\hat{g}_i(\boldsymbol{\phi}_i + \beta m_i)$$

where $\beta \in [0, 1]$ is smoothing parameter. As clearly noted in [78], when the poor update $\boldsymbol{\phi}_i + \beta m_i$ occurs, this NAG style update rule strongly pulls it back towards $\boldsymbol{\phi}_i$. Because of the SPSA's strongly stochastic nature, we conjecture that this *gradient correction* property is also highly effective for SPSA as well as first-order optimization algorithms. Algorithm 1 summarizes the BlackVIP algorithm.

5. Results

We first provide the experimental setup in Section 5.1. Next, Section 5.2 presents the comparison between SPSA-GC and the previous ZO method. Besides, we provide domain generalization and object location sensitivity experiments. Section 5.3 and 5.4 provide the results on 14 transfer learning benchmarks and ablation studies, respectively.

5.1. Experimental Setup

We extensively evaluate BlackVIP on 14 benchmarks (refer Supp A.1). These cover diverse visual domains and tasks, so they require understanding various visual semantics like scenes, actions, fine-grained categories, textures, satellite imagery, the number of objects, and the recognition of generic objects. Additionally, to investigate the importance of prompt design, we consider two synthetic datasets: Biased MNIST and Loc-MNIST (see Sec 5.2 and Fig. 4).

In this paper, we adopt CLIP ViT-B/16 [63] as a target PTM because it does not require a separate classifier for different tasks, and has a strong zero-shot generalization capability. For the frozen encoder of Coordinator, we use ImageNet pre-trained `vit-mae-base` checkpoint. As the baselines, we consider CLIP's zero-shot classifier (ZS), black-box adversarial reprogramming (BAR) [79], and VP with SPSA-GC that simply replace the backpropagation in VP [3] with SPSA-GC. Following the few-shot classification setting of [94], we use 16-shot training samples and the full testset by default. More details are provided in Supp A.

5.2. Synthetic Datasets

Comparison among optimization algorithms We validate our SPSA-GC on the well-known optimization benchmark, Rosenbrock function. We report the normalized loss ($\frac{|L(\theta^*)-L(\theta)|}{|L(\theta^*)-L(\theta_0)|}$) where $L(\theta^*)$ and $L(\theta_0)$ is the loss value on the optimal and initial point, respectively, and $L(\theta)$ is a loss value on the current parameter $\theta \in \mathbb{R}^{100}$. In Fig. 3 (left), SPSA-GC shows faster and more stable convergence than Random Gradient-Free (RGF) [49,79], and even achieves a comparable result to Nesterov's Accelerated Gradient (SGD-NAG) using true gradients. Besides, we simulate the noisy loss observation (emulating the mini-batch optimization) by adding Gaussian noise to learning loss, i.e., $L_{noisy}(\theta) = L(\theta) + \epsilon$, where $\epsilon \sim N(0, scale^2)$. In Fig. 3 (right), as the noise increases, RGF rapidly degenerates while SPSA is still relatively stable, and our gradient correction (SPSA-GC) gives further improvement.

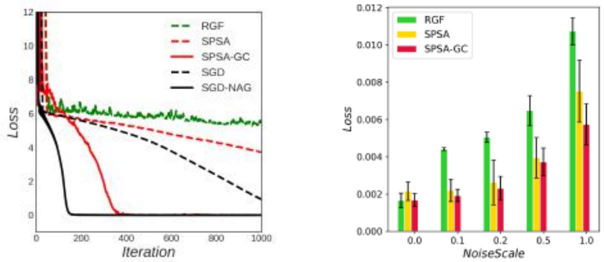


Figure 3. (Left) loss curve and (right) noise sensitivity analysis of 100-Dimensional Rosenbrock optimization.

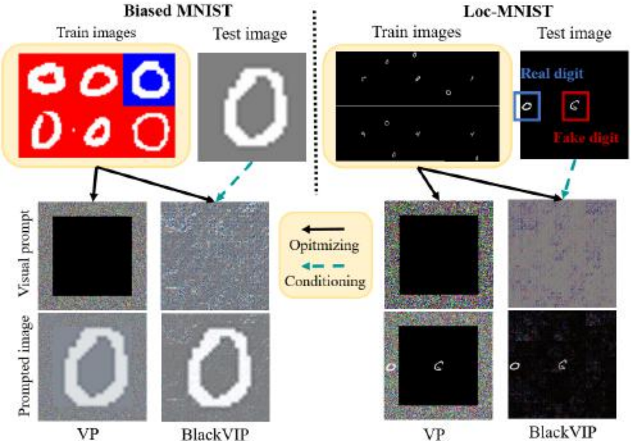


Figure 4. Prompt visualization on synthetic datasets. Unlike VP, our BlackVIP designs input-dependent conditional prompts contributing to the robustness under distribution/object-location shift.

Robustness on Distribution Shift Next, we evaluate our method on Biased MNIST [2] to investigate the robustness of BlackVIP's input-dependent automatic prompt design under distribution shift. Biased MNIST is a modified version of MNIST [43], constructed to validate a model's generalization ability under color bias shift. At train-time, each digit has a unique preassigned background color that strongly correlates with the label. The degree of correlation is determined by the value $\rho \in [0, 1]$, and the correlation ratio is reversed as 1-ρ at test-time. Results are summarized in Tab. 2 (left) and Fig. 5, respectively. In this setup, BlackVIP remarkably outperforms others (even white-box VP), and the performance gap goes larger under the stronger correlation. This means our input-dependent image-shaped prompts can be beneficial in *domain generalization* settings.

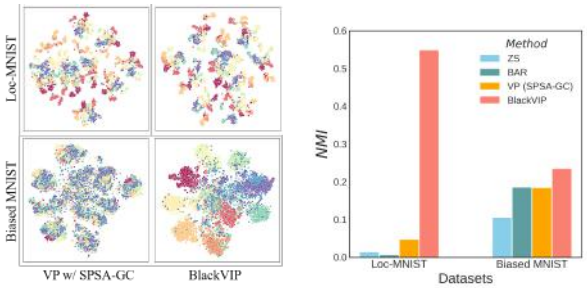


Figure 5. (Left) t-SNE [80] of prompted images' embedding on Biased MNIST and Loc-MNIST. (right) Normalized Mutual Information (NMI) [50] score of learned embedding.

Table 2. Results on synthetic datasets. BlackVIP shows robust performance under distribution/object-location shift.

Method	Biased MNIST				Loc-MNIST			
	16-Shot		32-Shot		16-Shot		32-Shot	
	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.9$	1:1	1:4	1:1	1:4
VP (white-box)	57.92	43.55	69.65	42.91	86.79	86.54	90 .18	92.09
ZS	37.56	37.25	37.56	37.25	29.70	22.70	29.70	22.70
BAR	53.25	53.07	53.93	53.30	33.98	26.05	34.73	27.72
VP w/ SPSA-GC	60.34	53.86	59.58	51.88	16.21	25.68	18.43	30.13
BlackVIP	**66.21**	**62.47**	**65.19**	**64.47**	**69.08**	**60.86**	**76.97**	**67.97**

Robustness on Object Location Shift We expect that BlackVIP adopts input-dependent image-shaped prompts, so does be still robust even if the object is not always located in the center of the image. To validate this, we create a variant of the MNIST, Loc-MNIST, by putting a real target digit on the four edges and an arbitrary fake digit in the center of the black blank image. The location of the target digit and the class of the fake digit are chosen randomly. We further consider a more challenging setup in that the fake digit is four times larger (1:4) than the real one. We summarize the results in Tab. 2 (right) and Fig. 5, respectively. Compared to input-independent frame-shaped prompting (BAR and VP), BlackVIP achieves significantly better performance which proves the superiority of the Coordinator's prompt design.

5.3. Few-shot Transfer Learning on Benchmarks

We consider the 14 few-shot benchmark datasets following [3, 93, 94]. As shown in Tab. 3, while BAR and VP undergo large performance variations across 14 datasets, BlackVIP boasts consistently high performance (i.e., improves the zero-shot performance on 13 over 14 datasets). Specifically, BAR shows promising results on the tasks that require understanding coarse semantics (DTD [15], EuroSAT [32], and RESISC [13]), but fails to show competitiveness on CLEVR [37] that requires visual reasoning (counting objects) by capturing the overall image semantics. Meanwhile, BlackVIP performs well across various tasks by extending or limiting attention of frozen PTM (Fig. 6), which denotes BlackVIP is a high-capability prompt learner that robustly adapts the PTM to diverse downstream tasks.

Practically, BlackVIP has three major advantages: 1) it only requires the 9K learnable parameters (see Tab. 4), while BAR and VP require 37K and 69K parameters. 2) It greatly reduces the peak memory allocation compared to white-box transfer learning methods. 3) BlackVIP shows outstanding query efficiency among the prompting methods (see Fig. 7). For instance, by sending just 10K queries with 12 USD (based on Clarifai Vision API), we can improve the performance of a zero-shot model about twice.

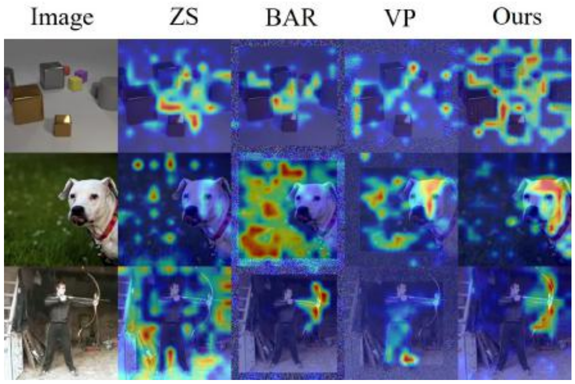


Figure 6. Grad-CAM analysis on CLVER, Pets, and UCF101.

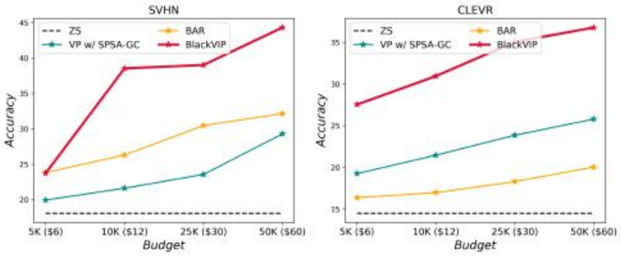


Figure 7. Query efficiency. (x-axis) A number of queries and cost for achieving (y-axis) corresponding performance.

5.4. Ablation Study

In this section, we provide two additional results for our BlackVIP: 1) we validate whether BlackVIP can achieve superior performance across four target backbones and two encoders of Coordinator. 2) we present the ablation study about pre-trained weights and optimization algorithms.

Architectures To study the versatility of our method, we vary the backbone architecture of the pre-trained target model and the encoder of Coordinator in Tab. 5. While BAR and the naive application of SPSA-GC on VP fail to improve the zero-shot performance of CNN-based target backbones that lack the global attention of Transformers

Table 3. Classification accuracy across 14 benchmarks that require natural, specialized, structured, and fine-grained visual recognition. BlackVIP shows outstanding results among input-space prompting methods. *Win* means the number of datasets that each method beats the zero-shot performance. Grays are the results of white-box learning. All experiments are done in 16 shots with three repeated runs.

Method	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	SVHN	EuroSAT	RESISC	CLEVR	UCF	IN	*Avg.*	*Win*
VP (white-box)	94.2	90.2	66.9	86.9	81.8	31.8	67.1	61.9	60.4	90.8	81.4	40.8	74.2	67.4	71.1	13
ZS	92.9	89.1	65.2	**71.3**	86.1	24.8	62.6	44.7	18.1	47.9	57.8	14.5	66.8	66.7	57.6	-
BAR	**93.8**	88.6	63.0	71.2	84.5	24.5	62.4	**47.0**	34.9	**77.2**	**65.3**	18.7	64.2	64.6	61.4	6
VP w/ SPSA-GC	89.4	87.1	56.6	67.0	80.4	23.8	61.2	44.5	29.3	70.9	61.3	25.8	64.6	62.3	58.8	4
BlackVIP	93.7	**89.7**	**65.6**	70.6	**86.6**	**25.0**	**64.7**	45.2	**44.3**	73.1	64.5	**36.8**	**69.1**	**67.1**	**64.0**	**13**

Table 4. Train-time peak memory allocation (Peak Memory) and the number of learnable parameters (Params) on ImageNet.

Method	Peak Memory (MB)		Params	
	ViT-B	ViT-L	ViT-B	ViT-L
FT (white-box)	21,655	76,635	86M	304M
LP (white-box)	**1,587**	3,294	513K	769K
VP (white-box)	11,937	44,560	69K	69K
BAR	1,649	3,352	37K	37K
VP w/ SPSA-GC	1,665	3,369	69K	69K
BlackVIP	2,428	**3,260**	**9K**	**9K**

Table 5. Ablation study for backbone architecture. Classification accuracy on EuroSAT across pre-trained target backbone architectures and BlackVIP's Coordinators (SSL encoder backbone).

Method	**Target Backbone**				
	RN50	RN101	ViT-B/32	ViT-B/16	*Avg.*
ZS	37.5	32.6	45.2	40.8	48.4
BAR	26.9	33.5	70.3	**77.2**	52.0
VP w SPSA-GC	34.7	31.2	**71.1**	70.9	52.0
Ours (RN50)	**51.3**	50.8	62.9	68.5	58.4
Ours (VIT-B/16)	48.4	**51.3**	67.9	73.1	**60.2**

[81], our BlackVIP consistently brings huge performance gains across all the architectures. It implies that BlackVIP is an *architecture-agnostic* approach, which pursues the general adaptation method for high-performing PTMs.

Table 6. Different Coordinator weights with SPSA variants. Mean classification accuracy of three repeated runs on EuroSAT

Encoder Type	Optim.	Acc.
Zero-Shot		47.9
scratch	SPSA	49.6
scratch	SPSA-GC	49.5
Sup. pre-trained	SPSA	59.4
Sup. pre-trained	SPSA-GC	65.2
SSL pre-tained	SPSA	69.4
BlackVIP (*SSL pre-trained* with SPSA-GC)		**73.1**

Coordinator weights and ZOO algorithms BlackVIP adopts the encoder-decoder structure to efficiently generate the input-dependent image-shaped prompts. We exploit an SSL pre-trained encoder while we plug the randomly initialized extremely lightweight decoder. From the design philosophy of BlackVIP, we expect that a pre-trained encoder extracts the rich semantic features of the given image, including the spatial features, and the decoder utilizes the features to produce a spatially and semantically structured prompt tailored to the input. We conjecture that an SSL pre-trained encoder is desirable to capture the demanding diverse semantics instead of a supervised one learned from pre-defined labels. Therefore, for Coordinator, BlackVIP adopts an SSL encoder (i.e., Masked Auto-Encoder [29]). Tab. 6 confirms that the SSL encoder outperforms the supervised pre-trained or randomly initialized encoder (scratch). Besides, SPSA-GC improves the 3.7% accuracy than SPSA, from 69.4 to 73.1. It denotes that approximated gradients by our SPSA-GC are more accurate than the original SPSA.

6. Conclusion

We pioneer *black-box visual prompting* for the realistic and robust adaptation of pre-trained models. We propose BlackVIP, which reparameterizes the input-space prompt as a conditional generative network Coordinator and equips our new ZOO algorithm, SPSA-GC, rather than backpropagation. BlackVIP does not require any accessibility on model architecture or parameters and efficiently adapts the pre-trained model to targeted downstream tasks. Extensive empirical results show that BlackVIP consistently improves the performance over baseline methods on few-shot adaptation, distribution shift, and object-location shift with minimal parameters, memory capacity, API queries, and cost.

Acknowledgement

This work was supported by a grant from National Research Foundation of Korea(2022R1A4A3033874, 2021R1F1A1060117, 2020R1C1C1008726, 2017R1D1A1 B05028565), and supported by Institute of Information & Communications Technology Planning & Evaluation grant funded by the Korea government (No.2021-0-02067)

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 6, 13

[3] 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. 1, 2, 3, 4, 6, 7, 14, 15

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 2

[5] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. Visual prompting via image inpainting. *arXiv preprint arXiv:2209.00647*, 2022. 1

[6] Andrei Boiarov, Oleg Granichin, and Olga Granichina. Simultaneous perturbation stochastic approximation for few-shot learning. In *2020 European Control Conference (ECC)*, pages 350–355, 2020. 5

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. 13

[8] 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. 1

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 5

[10] Pin-Yu Chen. Model reprogramming: Resource-efficient cross-domain machine learning. *arXiv preprint arXiv:2202.10629*, 2022. 4

[11] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022. 3

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[13] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 7, 13

[14] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 14

[15] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 7, 13

[16] 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, 2009. 3, 4, 13

[17] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022. 3

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[19] 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. In *International Conference on Learning Representations*, 2020. 1, 3

[20] Gamaleldin F Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*, 2018. 4, 14

[21] Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, and Xinggang Wang. Unleashing vanilla vision transformer with masked image modeling for object detection. *arXiv preprint arXiv:2204.02964*, 2022. 5

[22] Li Fei-Fei, R. Fergus, and P. 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, 2004. 13

[23] 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

[24] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 4

[25] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[26] Han Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. Text generation with efficient (soft) q-learning. *arXiv preprint arXiv:2106.07704*, 2021. 3

[27] Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18, 2003. 3

[28] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary computation*, 9(2):159–195, 2001. 3

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 5, 8, 15

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[31] 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. 2

[32] 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. 7, 13

[33] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[34] Gabriel Huang, Issam Laradji, David Vázquez, Simon Lacoste-Julien, and Pau Rodriguez. A survey of self-supervised and few-shot object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

[35] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2, 14

[36] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 1, 3

[37] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 7, 13

[38] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 1, 2

[39] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 3

[40] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022. 2

[41] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 13

[42] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2021. 1

[43] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6, 13

[44] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1, 3

[45] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 1, 3

[46] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 5

[47] Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*, 2022. 2, 5

[48] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020. 3

[49] Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018. 6

[50] Jochem Loedeman, Maarten C Stol, Tengda Han, and Yuki M Asano. Prompt generation networks for efficient adaptation of frozen vision transformers. *arXiv preprint arXiv:2210.06466*, 2022. 6

[51] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 13

[52] Igor Melnyk, Vijil Chenthamarakshan, Pin-Yu Chen, Payel Das, Amit Dhurandhar, Inkit Padhi, and Devleena Das. Reprogramming large pretrained language models for antibody sequence infilling. *arXiv preprint arXiv:2210.07144*, 2022. 4

[53] Paarth Neekhara, Shehzeen Hussain, Jinglong Du, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley. Cross-modal adversarial reprogramming. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2427–2435, 2022. 4, 5

[54] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983. 5

[55] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, apr 2017. 14

[56] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 13

[57] 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, 2008. 13

[58] Jiachun Pan, Pan Zhou, and Shuicheng Yan. Towards understanding why mask-reconstruction pretraining helps in downstream tasks. *arXiv preprint arXiv:2206.03826*, 2022. 5

[59] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387, 2016. 4

[60] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012. 13

[61] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[62] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 3

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 13, 14

[64] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 16

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[66] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[67] Qing Song, James C. Spall, Yeng Chai Soh, and Jie Ni. Robust neural network tracking controller using simultaneous perturbation stochastic approximation. *IEEE Transactions on Neural Networks*, 19(5):817–835, 2008. 5

[68] 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. 13

[69] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992. 2, 3, 5, 14, 15

[70] J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10):1839–1853, 2000. 5

[71] James C. Spall. Accelerated second-order stochastic optimization using only function measurements. *Proceedings of the 36th IEEE Conference on Decision and Control*, 2:1417–1424 vol.2, 1997. 5

[72] James C. Spall. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112, 1997. 5

[73] James C Spall. An overview of the simultaneous perturbation method for efficient optimization. *Johns Hopkins apl technical digest*, 19(4):482–492, 1998. 3

[74] James C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., USA, 1 edition, 2003. 5, 15

[75] Daniel Stein, Jochen Schwenninger, and Michael Stadtschnitzer. Simultaneous perturbation stochastic approximation for automatic speech recognition. In *Proc. Interspeech 2013*, pages 622–626, 2013. 5

[76] Tianxiang Sun, Zhengfu He, Hong Qian, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. Bbtv2: Towards a gradient-free future with large language models. In *Proceedings of EMNLP*, 2022. 3

[77] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of ICML*, 2022. 3

[78] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. 5

[79] Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources.

In *International Conference on Machine Learning*, pages 9614–9624. PMLR, 2020. 3, 4, 6, 14, 15

[80] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 8

[82] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 5

[83] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 5

[84] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 1

[85] Guangxuan Xiao, Ji Lin, and Song Han. Offsite-tuning: Transfer learning without full model. *arXiv preprint arXiv:2302.04870*, 2023. 2

[86] 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, 2010. 13

[87] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[88] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020. 4

[89] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[90] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 1, 2

[91] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2

[92] Tingting Zhao, Hirotaka Hachiya, Gang Niu, and Masashi Sugiyama. Analysis and improvement of policy gradient estimation. *Advances in Neural Information Processing Systems*, 24, 2011. 3

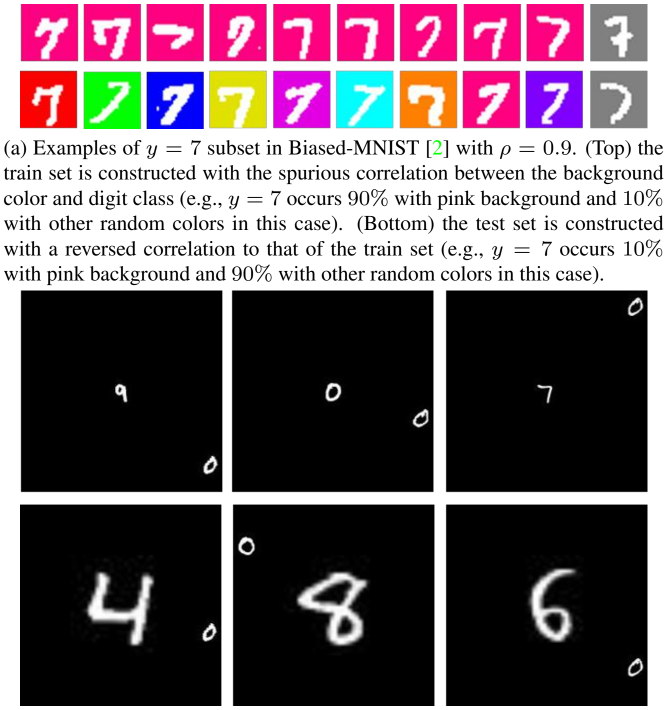
[93] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 3, 7, 13

[94] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 6, 7, 13, 14

A. Experimental Setting

A.1. Datasets

Synthetic Datasets Our BlackVIP generates the input-dependent image-size visual prompt which covers the whole image region, so we expect that this flexible prompt design can improve some kind of robustness as well as general recognition capability: (1) To evaluate the robustness on distribution shift (i.e., domain generalization), we consider Biased-MNIST [2] dataset. (2) To evaluate the robustness on adversarial noise and location-agnostic recognition capacity, we create a variant of the MNIST dataset called Loc-MNIST. Examples of these two datasets are provided in Figure 8.



(a) Examples of $y = 7$ subset in Biased-MNIST [2] with $\rho = 0.9$. (Top) the train set is constructed with the spurious correlation between the background color and digit class (e.g., $y = 7$ occurs 90% with pink background and 10% with other random colors in this case). (Bottom) the test set is constructed with a reversed correlation to that of the train set (e.g., $y = 7$ occurs 10% with pink background and 90% with other random colors in this case).

(b) Examples of Loc-MNIST dataset. The real digit from MNIST is located in the outer area, while the fake digit from another random MNIST image is placed in the center of the image. (Top) the case where the size ratio of the real digit to the fake digit is 1:1, and (Bottom) 1:4.

Figure 8. Examples of two synthetic datasets. (a) Biased MNIST and (b) Loc-MNIST.

Biased MNIST is a modified version of MNIST [43] where the biases reside in the background colors of the images of each digit. At train time, each digit has a unique preassigned background color that strongly correlates with the label. The degree of correlation is determined by the value $\rho \in [0, 1]$, such that $(100 \times \rho)\%$ of the images that belong to the same digit have the preassigned color of that digit as their background color, and the rest are uniformly assigned to have any of the other colors as their background color. At test time, we reverse the ratio so that $(100 \times (1-\rho))\%$ of the images now have the preassigned color as their background color and vice versa to evaluate the model's dependency on superficial features such as the color of the background that a digit is located on. We prepare the following two environments 1) easy: $\rho = 0.8$ and 2) hard: $\rho = 0.9$.

On the given black blank image with 224 × 224 resolution, i.e., zero's array, **Loc-MNIST** puts an original target digit image from MNIST that has 28 × 28 resolution on the edge-side (e.g., 0∼27 or 196∼223 for one of vertical or horizontal side and 0∼223 for another side) and puts a random fake digit (also from the MNIST dataset) on the center. The location of the target digit in the edge and the class of fake digit are chosen randomly with uniform probability. A synthetic image is created one by one for each original MNIST image. We prepare the following two environments 1) easy: the scale of the target and the fake digit is the same, i.e., 1:1, and 2) hard: the fake digit is four times larger than the original digit, i.e., 1:4.

For consistency, we perform the experiments on these two datasets with a few-shot evaluation protocol. To construct a train set, we randomly sample a subset (K-shot) of the created images for each class and use the whole test set.

Datasets To extensively evaluate the effectiveness of our proposed method and baseline approaches, we measure performance across the following 14 datasets that are widely used for transfer learning benchmark: Caltech101 [22], OxfordPets [60], StanfordCars [41], Flowers102 [57], Food101 [7], FGVCAircraft [51], SUN397 [86], DTD [15], SVHN [56], EuroSAT [32], Resisc45 [13], CLEVR [37], UCF101 [68], and ImageNet (IN) [16]. Note that these 14 datasets cover diverse visual domains, and they require understanding various visual semantics like scenes, actions, fine-grained categories, textures, satellite imagery, digits, the number of objects, and the recognition of generic objects.

Following the protocol in [93,94], we conduct a few-shot evaluation for all datasets: 16-shot for the train set, 4-shot for the validation set, and the whole test set. We use the few-shot split by [94] for each dataset those are also used in [94], while for Resisc45 and CLEVR, we randomly select the 16-shot and 4-shot samples for training and validation dataset, respectively.

A.2. Backbone Model

In this work, we aim at the robust adaptation of pre-trained models on diverse downstream tasks. For these pre-trained models, all experiments in this paper are done with the off-the-shelf vision-language model CLIP [63], and we adopt the ViT-B/16 for image encoder backbone architecture by default. During the adaptation (training) phase, the entire components of the pre-trained model are frozen without any architectural modification, and we only manage and optimize the learnable module Coordinator from the outside of the pre-trained model.

While input space visual prompting allows it to be applied to not only VLM, but also any other vision models like CNNs and ViTs, it requires the user to define the output space mapping, which maps the output prediction category set of a pre-trained task to a new downstream category set [3, 20, 79]. This is another non-trivial problem. Therefore, we limit our focus to only the VLM that can dynamically build the task-specific head from manual text template [35, 63] so that free from defining output space mapping.

A.3. Baseline Methods

CLIP Zero-Shot (ZS) CLIP [63] is one of the most popular vision-language zero-shot models that is widely exploited for classification, detection, segmentation, and other vision or vision-language tasks. Based on its well-aligned vision-language joint embedding space, the zero-shot classification can be performed with a manual text prompt (also called template) of each pre-defined class category. In this paper, we are mainly aiming to improve the CLIP's strong zero-shot performance in the few-shot adaptation setting.

BAR Black-Box Adversarial Reprogramming (BAR) [79] was proposed for efficient transfer learning of pre-trained model to the medical image domain. Different from the previous works on Adversarial Reprogramming (AR), BAR exploits the perturbation-vulnerability of neural networks for *adaptation* purpose rather than attack. By optimizing the frame-shaped learnable program, which embeds a downstream target image inside of that, BAR steers the ImageNet pre-trained model to classify the specialized medical images. Moreover, BAR adopts the zeroth-order optimizer (ZOO), Randomized Gradient-Free (RGF) [55] minimization algorithm for black-box transfer learning to broaden its applications.

When the resolution of the downstream input image is over that of the pre-training phase, Tsai et al. [79] set the embedded target image size for 64×64 resolution in the 299×299-size learnable program by default. However, we observe that such a heavy-pad thin-image design of prompt degrade the performance significantly, so we tune the resolution of the embedded image and set 194×194.

VP Similarly, Visual Prompting (VP) aims at adapting a pre-trained model to downstream tasks via learning input space visual prompts. Among some candidates for prompt designs, Bahng et al. [3] adopt the padding-style prompt so that realized prompts look like the frame-shape program of ARs. VP learns a universal visual prompt per each downstream task, and it just adds to all of the images in a task. Unlike the AR methods or our BlackVIP, the range of prompted images is unbounded. Following [3], we use the padding-style prompt, which is 30-pixel sized for each side by default.

While VP optimizes the parameters in the input space, it relies on a first-order optimization algorithm that uses the true gradient of entire model parameters, and we establish the performance of VP as an upper bound for other input space black-box optimization approaches, including BlackVIP. Additionally, by replacing the first-order algorithm with zeroth-order counterparts, we build two new baselines **VP w/ SPSA** and **VP w/ SPSA-GC** on our extensive experiments. These two methods confirm the effectiveness of our new components *Coordinator* and SPSA-GC.

Discussion Although BAR, VP, and BlackVIP share the generic goal: efficient transfer learning of pre-trained models via input-space optimization, there are several significant differences. (1) We propose a novel prompt design that is automatically formed in an input-dependent manner rather than the frame-shaped manual design of the input-independent prompt (or program) of VP (or BAR). (2) While VP relies on first-order algorithms and BAR adopts the RGF, we utilize the new variants of SPSA [69], SPSA-GC, which is enhanced with a proper modification in the parameter update rule. (3) Contrary to the medical imaging-only validation in BAR, based on the above two technical difference, BlackVIP successfully adapt the pre-trained model to diverse data domains (described in Section B.1.).

A.4. Implementation Details

Architecture For the fixed text prompt design of each dataset those are shared across all baseline methods and BlackVIP, we use the same templates provided by [3] for SVHN, CLEVR, and Resisc45, and [94] for remaining 11 datasets. For the frozen feature extractor (encoder) part of our *Coodinator*, we use the ImageNet pre-trained `vit-mae-base` checkpoint[†] from the HuggingFace. The output shape of the encoder is $N \times 768$, where N is the number of instances in the batch. We design the decoder based on *depth-wise separable convolution* (DSC) layer [14] for parameter efficiency. Specifically, we build a block of [`NORM-ACT-CONV`] and stack it five times. The `NORM` and `ACT` denote Batch Normalization and Gaussian Error Linear Unit, respectively. The `CONV` operation of the first four blocks is DSC, and the last one is a standard convolutional layer. Our implementation code is enclosed in `.zip` file.

To satisfy a fully convolutional design without loss of expressiveness, tensors that are fed into the decoder must be shaped in a 3D feature map. For this, we additionally govern a task-specific single continuous vector ϕ_t (called

[†] https://huggingface.co/docs/transformers/model_doc/vit_mae

prompt trigger vector), which is concatenated with the output feature vector of encoder leading the appropriate size of 1d vector for reshaping to 3d tensor. In this work, we set the dimension of the prompt trigger vector to 800, resulting in 1568 dimensions of concatenated vector that can be reshaped to $32 \times 7 \times 7$ shaped 3D tensor. The prompt trigger is shared across all instances for a given task.

Optimization and other configurations For a stable approximation of gradient in practice, ZOO algorithms repeat the gradient estimation step for several times and use the mean of those estimates as a final approximation of the gradient. Usually, the approximation quality is proportional to the number of these repeats. We set this repeat as five times for all baselines that use ZOO.

Besides the learning rate and learning rate schedule parameters, ZOO algorithms have some additional algorithm-specific hyperparameters needed to be tuned. For RGF, these are the standard deviation of a random gaussian vector and a smoothing parameter, and for SPSA, these are the perturbation magnitude and its decaying factor. We provide the search range of each hyperparameter in Table 7. The search range for algorithm-specific parameters is based on the proposal of authors of SPSA [74] and BAR [79]. Moreover, among the valid perturbation distributions of SPSA, we adopt the Segmented Uniform $[-1.0, -0.5] \cup [0.5, 1.0]$.

The learning objective is a cross-entropy loss for VP and BlackVIP and focal loss for BAR (following [79]). For all black-box approaches, the batch size is set to 128 across all datasets. Except for the SUN397 (1,000), StanfordCars (2,500), and ImageNet (500), we optimize all methods during 5,000 epochs for convergence. Note that the input space visual prompting with first-order algorithm already requires sufficiently large iterations, e.g., 1,000 epoch [3] with full dataset, and ZOO demands much more iterations due to the lack of gradient information.

A.5. Hyperparameter Sweep

In this section, we provide the hyperparameter search range of each algorithm, summarized in Table 7.

B. Detail Description of *Coodinator*

On the transfer learning of a pre-trained model which provides no accessibility about any architectural information or actual model parameters, BlackVIP treats this situation with two novel mechanisms: (1) parameter-efficient instance-aware prompt generation network, and (2) stable zeroth-order optimization algorithm that is based on SPSA [69]. In this section, we provide a detailed description of the first component, Coordinator.

Different from existing works on visual prompting, we reparameterize the input space visual prompt ϕ as a neural network, *Coordinator* $h_{\boldsymbol{\phi}}(\cdot)$ that generates an input-dependent visual prompt $h_{\boldsymbol{\phi}}(x)$. Coordinator is composed with encoder $f(\cdot)$, decoder $g_{\phi_d}(\cdot)$ and task-specific learnable vector ϕ_t. The encoder is used for extracting instance-specific latent feature vector $z_x = f(x)$ contributing to the construction of the optimal input space visual prompt for each instance. Because our goal in this work is the broad utilization of pre-trained models on diverse downstream tasks, we adopt a pre-trained encoder network optimized by a self-supervised learning objective, not by a supervised learning objective or scratch network. Specifically, we use the ViT-B/16 weights from the *Masked AutoEncoding* pre-training [29]. We present the grounds for using the self-supervised learning encoder in the main paper, refer to Sec. 3. During the training phase, this pre-trained encoder part is frozen (not updated) and just acts as a feature extractor. Then, the instance-specific feature vector from the encoder is conveyed to the decoder for a prompt generation.

Table 7. Hyperparameter sweep. Large LR (learning rate) of BAR and VP is based on [3] to directly optimize pixel values rather than the neural network's weights. PM denotes perturbation scale, c_i.

Hyperparameter	Algorithm	Search Range
initial LR	BAR, VP	{40.0, 20.0, 10.0, 5.0, 1.0}
initial LR (a_1)	BlackVIP	{1.0, 0.1, 0.01, 0.005}
min LR	BAR	{0.1, 0.01, 0.001}
decaying step	BAR	{0.9, 0.5, 0.1}
LR decaying factor	VP, BlackVIP	{0.6, 0.5, 0.4, 0.3}
initial PM (c_1)	BlackVIP	{0.01, 0.005, 0.001}
PM decaying factor	BlackVIP	{0.2, 0.1}
std. of perturbation	BAR	{1.0, 0.5}
smoothing	BAR	{0.1, 0.01, 0.001}
gradient smoothing	VP, BlackVIP	{0.9, 0.7, 0.5, 0.3}

Prompt decoder $g_{\phi_d}(\cdot)$ is a lightweight convolutional neural network, which has learnable parameters **less than 10K** by default. Note that the generated prompt has the same shape as the input image, so our prompt covers the entire region of the image, unlike previous visual prompting and reprogramming works applied to the partial region of the image by human-designed.

In addition to the feature vector from the fixed encoder, the decoder also incorporates an additional input which is shared for all instances across the current dataset. The so-called *prompt trigger vector* ϕ_t is a continuous vector that also contributes to the design of a visual prompt by collaborating with the instance-specific rich feature vector from the encoder. By introducing this prompt trigger vector, the decoder of the Coordinator can enjoy additional information to generate more proper prompts for a given task. Besides, it helps to build the 3D feature map for the decoder's input, which is necessary for designing a parameter-efficient fully convolutional decoder network.

C. Grad-CAM Analysis

To investigate whether visual prompts produced by each method adapt the pre-trained model, we visualize the Grad-CAM [64] on the original image and prompted image of the encoder's penultimate layer (Figure 9-16). We select eight datasets that represent the diverse image domains and experimental conditions: (*Natural*) OxfordPets and SVHN, (*Specialized*) EuroSAT, (*Structured*) CLEVR-count, (*Action Recognition*) UCF101, (*Fine-Grained*) StanfordCars, (*Synthetic*) Biased-MNIST and Loc-MNIST. Detail descriptions for each dataset are provided in Sec. A.1.

BlackVIP generates diverse prompts to properly adapt the pre-trained model to the targeted data domains. When the task is counting the number of objects (CLEVR) in the entire region of an image, BlackVIP extends the attention of the pre-trained model to the whole view of an image as shown in Figure 9. If the task requires a fine-grained understanding of objects or recognition of actions (UCF101), BlackVIP concentrates the model's attention on class-related regions, as shown in Figure 10.

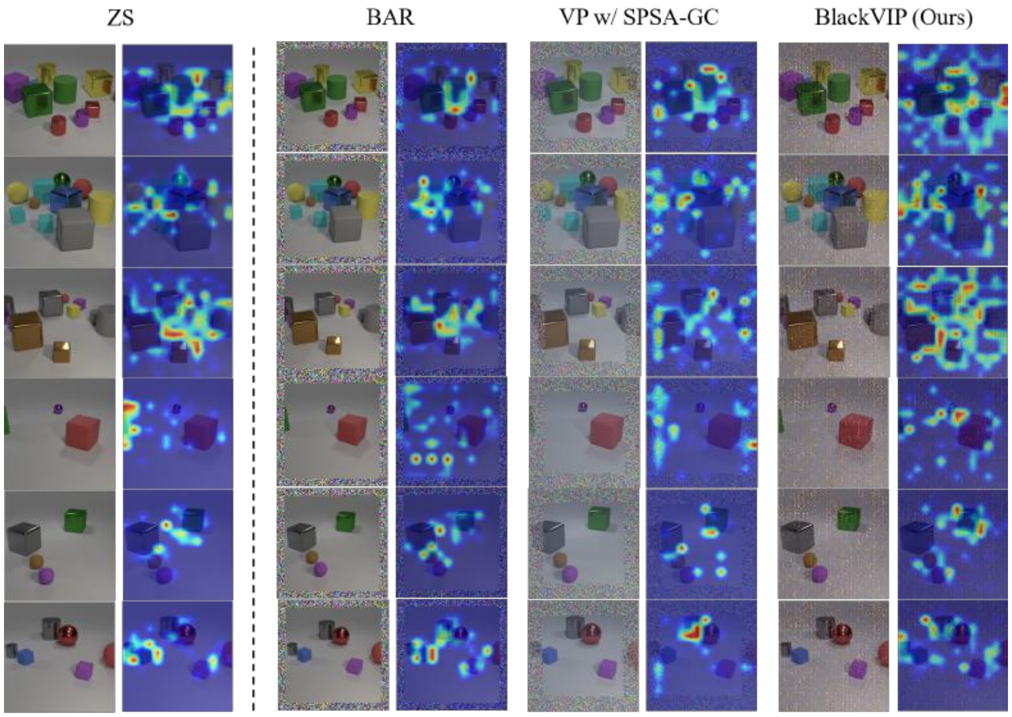


Figure 9. Grad-CAM on CLEVR. Compared to baseline methods, BlackVIP extends the attention of models to broad areas of the image for effective reasoning on the number of objects.

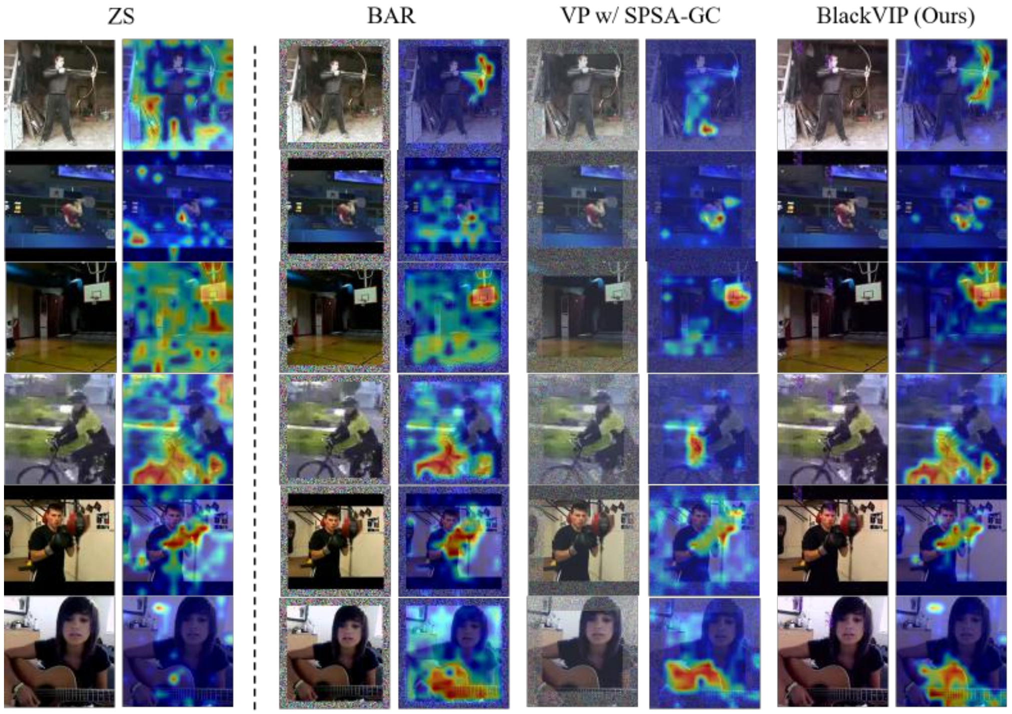


Figure 10. Grad-CAM on UCF101. Compared to baseline methods, BlackVIP concentrates the attention of models on local areas of the image for effective recognition of the specific actions.

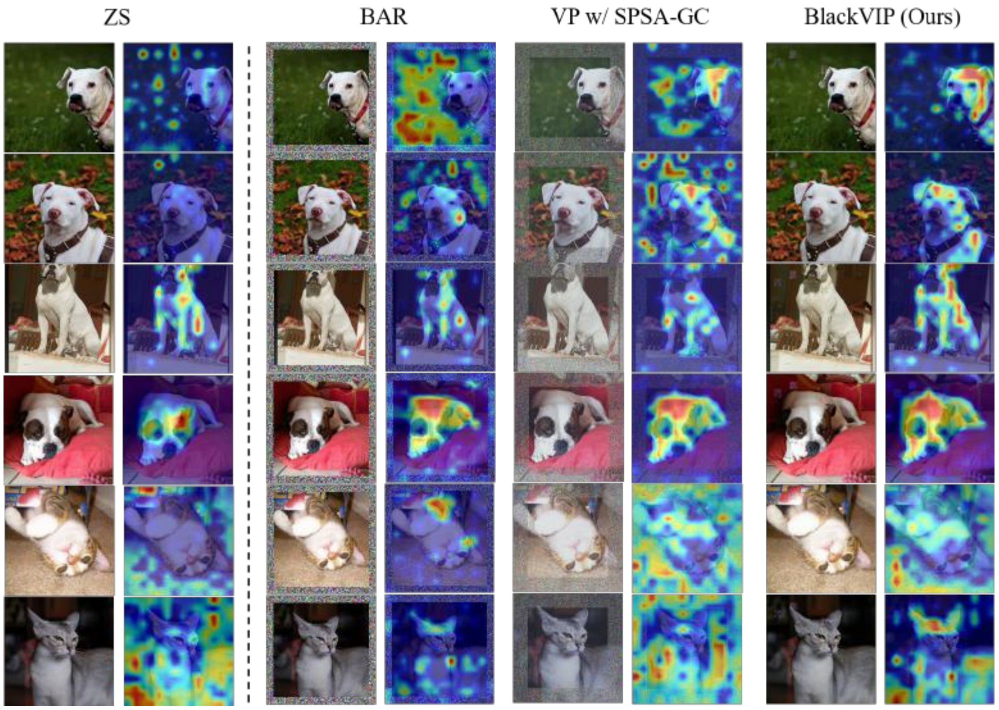


Figure 11. Grad-CAM on OxfordPets. Compared to baseline methods, BlackVIP effectively adapts the model to focus on the target object rather than spurious features such as the background.

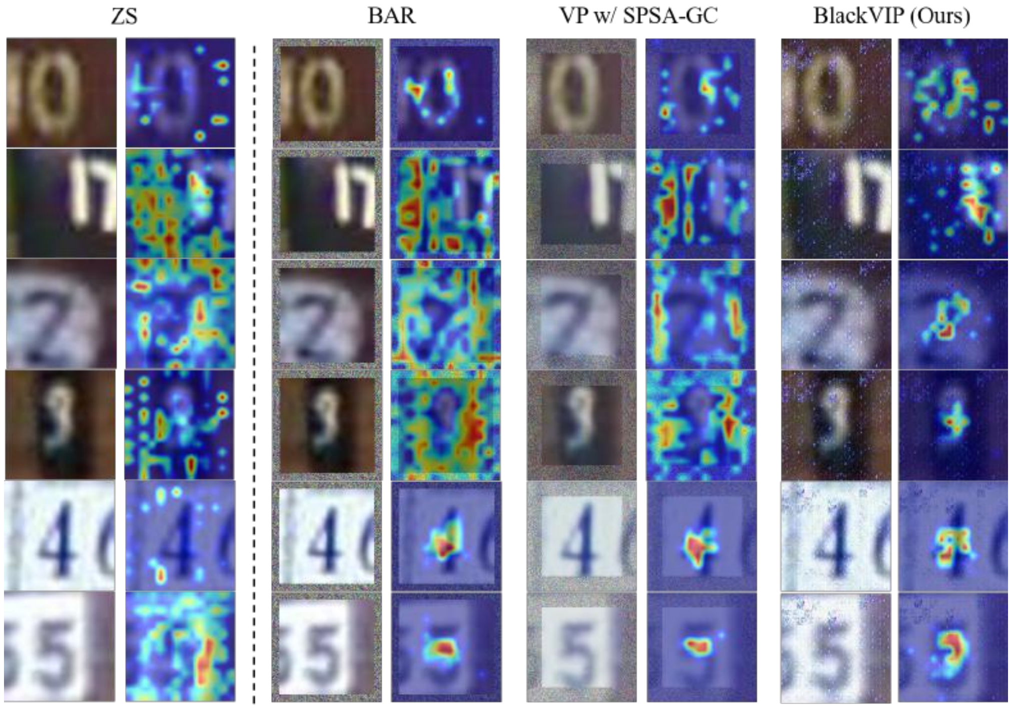


Figure 12. Grad-CAM on SVHN. Compared to baseline methods, BlackVIP effectively adapts the model to focus on the target digit rather than spurious features such as the background.

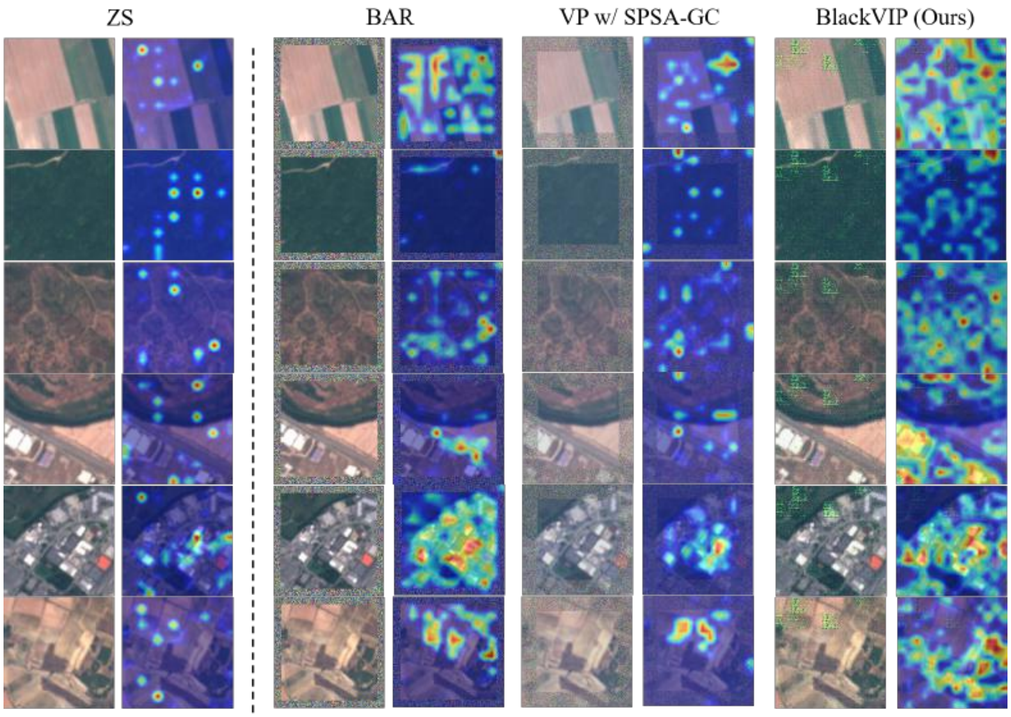


Figure 13. Grad-CAM on EuroSAT. Compared to baseline methods, BlackVIP extends the attention of models to broad areas of the image for effective classification of satellite imagery.

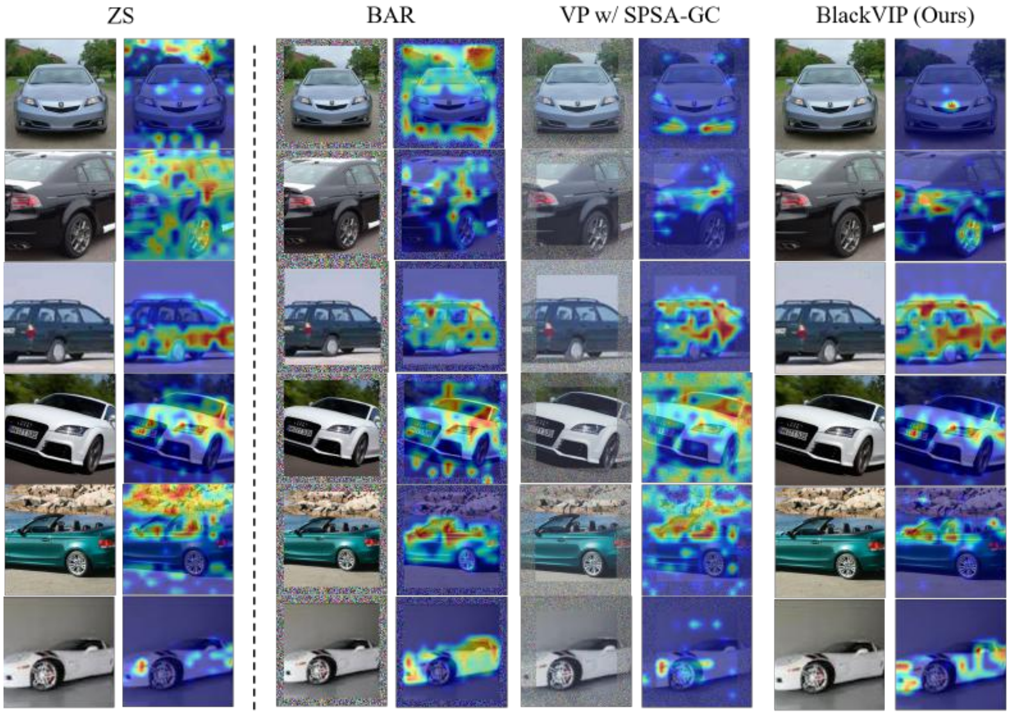


Figure 14. Grad-CAM on StanfordCars. Compared to baseline methods, BlackVIP concentrates the attention of models on an object or local areas of an image for effective fine-grained classification.

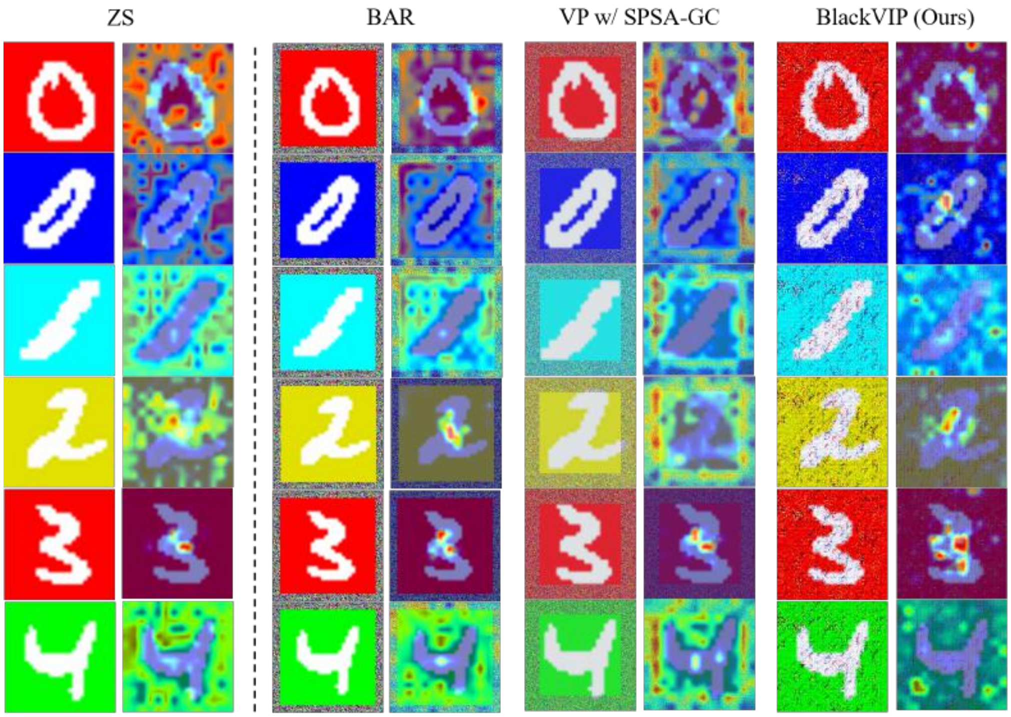


Figure 15. Grad-CAM on Biased-MNIST. While baseline methods attend to the background rather than digit shape, our BlackVIP can bypass this spurious feature through a widely scattered visual prompt and focus more of the attention on the shape of the digit.

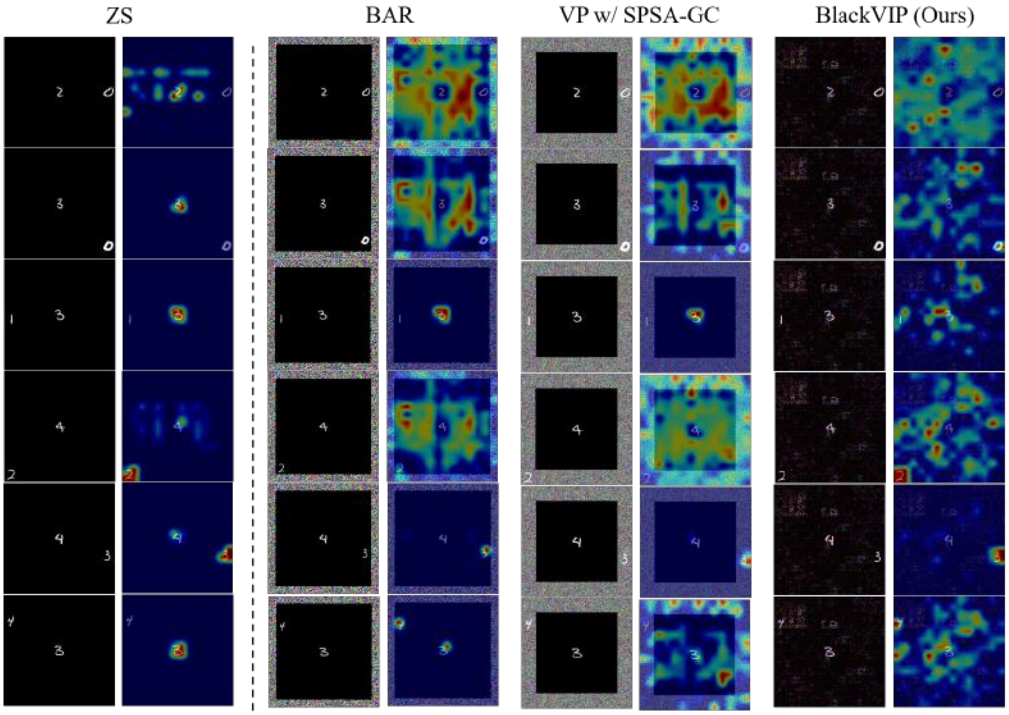


Figure 16. Grad-CAM on Loc-MNIST. Compared to baseline methods, BlackVIP effectively adapts the model to aim at edge-located true digit corresponding true label rather than the obstructive fake digit in the center of the image.