Black-box Prompt Tuning for Vision-Language Model as a Service

Lang Yu[1,2], **Qin Chen**[1,2*], **Jiaju Lin**[1] and **Liang He**[1,2]
[1]School of Computer Science and Technology, East China Normal University
[2]Shanghai Institute of AI for Education, East China Normal University
{lyu, jiaju_lin}@stu.ecnu.edu.cn, {qchen, lhe}@cs.ecnu.edu.cn

Abstract

In the scenario of Model-as-a-Service (MaaS), pre-trained models are usually released as inference APIs. Users are allowed to query those models with manually crafted prompts. Without accessing the network structure and gradient information, it's tricky to perform continuous prompt tuning on MaaS, especially for vision-language models (VLMs) considering cross-modal interaction. In this paper, we propose a black-box prompt tuning framework for VLMs to learn task-relevant prompts without back-propagation. In particular, the vision and language prompts are jointly optimized in the intrinsic parameter subspace with various evolution strategies. Different prompt variants are also explored to enhance the cross-model interaction. Experimental results show that our proposed black-box prompt tuning framework outperforms both hand-crafted prompt engineering and gradient-based prompt learning methods, which serves as evidence of its capability to train task-relevant prompts in a derivative-free manner.

1 Introduction

With the promise to learn universal cross-model representations, pre-trained vision-language models (VLMs) have achieved impressive performance in an extensive range of research fields [Du *et al.*, 2022]. Increased attention has focused on potential fine-tuning approaches to adapt these models to downstream tasks (e.g. linear probe [Radford *et al.*, 2021], adapter tuning [Houlsby *et al.*, 2019] and prompt tuning [Li and Liang, 2021]). However, making VLMs benefit everyone is still challenging. On the one hand, given a large number of tunable parameters, previous fine-tuning approaches can be computationally expensive; Moreover, institutes tend to keep the pre-trained model parameters closed-source due to commercial considerations such as GPT-3 [Brown *et al.*, 2020]. Since no gradient information is available, it's tricky for local users to perform fine-tuning when a model is deployed as a remote service (MaaS).

*Corresponding Author

Derivative-free Optimization, also referred as Black-box Optimization, is a promising way to optimize problems without the availability of gradient [Larson *et al.*, 2019]. Over the past decades, various methods have been explored, such as Bayesian Optimization [Shahriari *et al.*, 2015] and Evolution Strategies (ESs) [Loshchilov, 2014]. Particularly, ESs are widely used in automated machine learning. However, most ESs suffer from the "high-dimensional" problem and can only deal with thousands of parameters, which poses a challenge for black-box optimization over the large-scale pre-trained models, especially for the VLMs that involves multiple modalities with more parameters.

Recently, [Sun *et al.*, 2022b] provide a solution to perform black-box optimization on pre-trained language models (LMs). Inspired by the low intrinsic dimension of LMs [Aghajanyan *et al.*, 2021], the authors use projected prompts to bridge the gap between black-box optimization and fine-tuning. Whereas, this work is restricted to the single linguistic modality, and no cross-modal interaction is involved. Moreover, how to effectively find the optimal intrinsic vector in a multi-modal parameter subspace remains to be studied.

In this paper, we propose a Black-box Prompt Tuning framework for VLMs (BPT-VLM) to set the stage in the scenario of MaaS. Our framework regards multi-modal prompt tuning as black-box optimization based on empirically successful ESs (e.g. CMA-ES [Hansen and Ostermeier, 2001]). From the perspective of evolutionary learning, our BPT-VLM mainly consists of three parts (as shown in Figure 2):

(1) **Population**: According to the previous finding that pre-trained models have very low intrinsic dimensions [Aghajanyan *et al.*, 2021], optimization can actually perform on individuals (intrinsic vectors) from a small parameter subspace. Individuals forming a population can be further evaluated for distribution updates.

(2) **Objective Function**: As shown in Figure 1(a), model as a service is only allowed to perform forward pass, thus it's reasonable to define VLM as a black-box objective function to evaluate the fitness values of individuals;

(3) **Optimization Algorithm**: With the fitness values of a population of individuals, an evolution-based optimization algorithm updates its multivariate distribution to produce a higher-quality population in the next generation (Figure 1(b)).

Intrinsic vectors participate in VLM's forward passes as projected vision and language prompts. Depending on the length of propagation path, we explore two variants of prompt tuning for BPT-VLM, namely Shallow Prompt and Deep Prompt. After generations of black-box optimization, the multivariate distribution can produce solid intrinsic vectors with low loss value. In other words, the task-relevant vision and language prompts are learned in a derivative-free manner. Compared to [Sun *et al.*, 2022b], BPT-VLM considers both the visual and linguistic modalities in black-box tuning, and involves cross-modal interaction by optimizing vision-language prompts in the shared intrinsic subspace. Experimental results on 9 downstream tasks show that BPT-VLM not only surpasses the hand-crafted prompts, but also outperforms the prompts learned by gradient-based methods, namely Linear Probe [Radford *et al.*, 2021] and CoOp [Zhou *et al.*, 2022]. The main contributions of our work can be summarized as follows[1]:

- We propose a novel black-box prompt tuning framework for VLMs in the scenario of MaaS, which incorporates cross-modal interaction by sharing the intrinsic parameter subspace of both vision and language modalities and jointly optimizing the prompts with different modalities in a derivative-free manner.
- We extend traditional evolution strategies (CMA-ES, MM-ES, MA-ES) to a new scope of black-box prompt tuning on VLMs, and explore different prompt tuning variants (shallow and deep prompt) to further enhance the cross-modal interaction.
- Extensive experimental results show that prompts optimized in multi-modal intrinsic subspace can successfully adapt VLM to downstream tasks without accessing the gradient and model structure, which is more effective and efficient compared with the baselines.

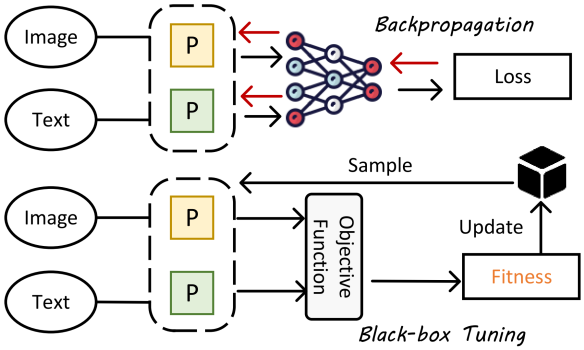
2 Related Work

2.1 Vision-Language Models

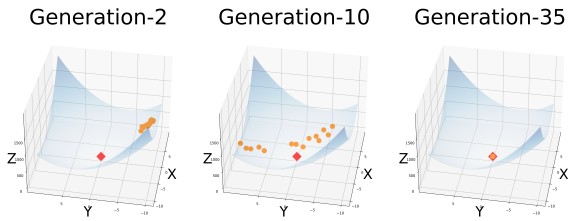
With the success of pre-trained models in the field of CV and NLP, many works attempted to pre-train large-scale models on both vision and language modalities, called Vision-Language Models (VLMs). According to the method of integrating multi-modal information, researchers categorize VLMs into fusion encoder-based models and dual encoder-based models. Fusion encoder-based VLMs feed the text embedding and image features into a unified Transformer or dual Transformers to perform further self-attention or cross-attention, such that information from two modalities could be integrated. Instead of relying on heavy transformer networks to model vision-language interaction, dual encoder-based VLMs adopt straightforward methods such as shallow attention layer or dot product to project the image embedding and text embedding to the same semantic space.

A representative of dual encoder-based VLMs is CLIP, which trains two single-model encoders using a contrastive loss to compute similarity scores for matching image-text pairs. After pre-training on 400 million data pairs collected from the Internet, CLIP demonstrates remarkable zero-shot image recognition capability on downstream tasks.

[1]Code is available at https://github.com/BruthYU/BPT-VLM



(a) Gradient-based and Black-box Prompt Tuning



(b) Covariance Matrix Adaptation Evolution Strategy

Figure 1: As shown in (a), black-box prompt tuning aims to optimize image and text prompts with derivative-free algorithms. The entire model is regarded as an objective function to evaluate the fitness of individuals. (b) gives an example of covariance matrix adaptation evolution, which searches for the minimum value of a binary function by adjusting the distribution of individuals through generations.

2.2 Prompt-based Learning

Prompt-based Learning originates from the NLP domain, which refers to prepend language instructions to the input text, and therefore reduces the gap between pre-trained LMs and downstream tasks. For instance, with the manually designed prompt "English: Hello, French: `[MASK]`", GPT-3 will output "Bonjour" as the result of a translation task. In addition to discrete prompts, recent works propose to treat prompt as continuous vectors and optimize them with gradient descent (see Figure 1(a)). Furthermore, prompt tuning can also be applied successfully to other fields: VPT [Jia *et al.*, 2022] extend prompt tuning to ViT for performance improvement on vision tasks; CoOp [Zhou *et al.*, 2022] introduces gradient-based prompt tuning to vision-language model.

In contrast to gradient-based learning, [Sun *et al.*, 2022b] empirically show that black-box tuning is feasible on large-scale LMs with prompts acting as a bridge, since the intrinsic dimension (the minimum number of parameters needed to be optimized) of pre-trained models can be compressed to several hundreds [Aghajanyan *et al.*, 2021].

2.3 Black-box Optimization Algorithms

Black-box optimization refers to optimizing the objective function without knowing the analytic expression and gra-

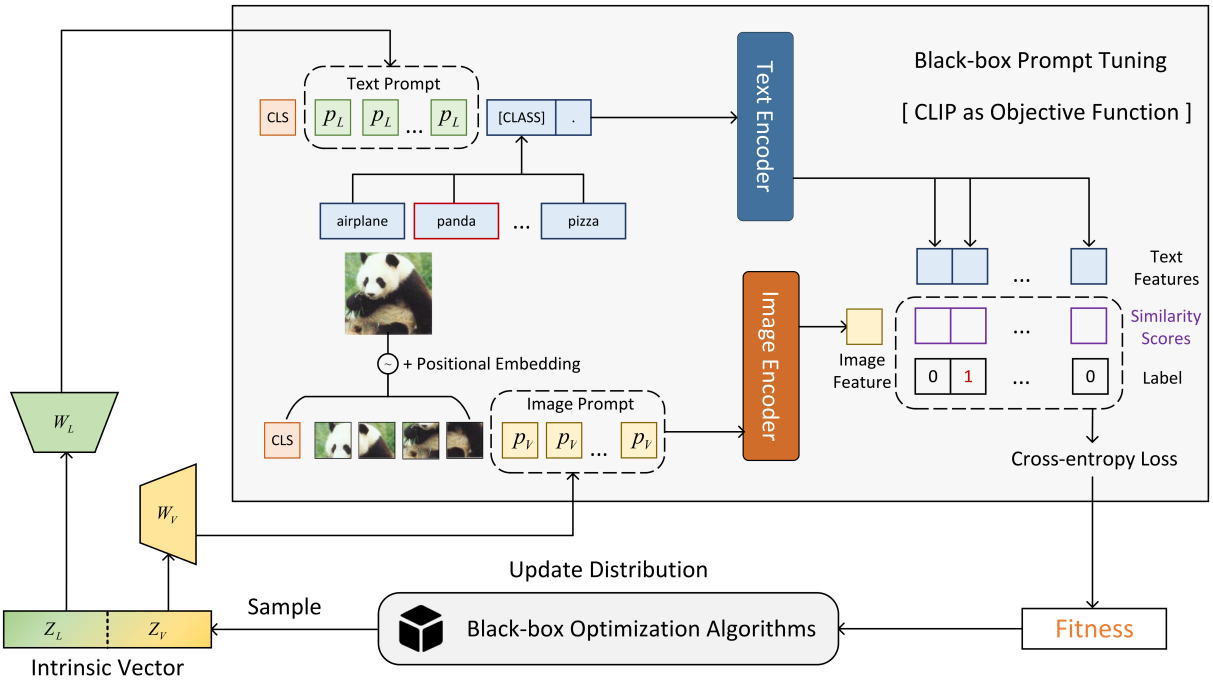


Figure 2: Overview of our black-box prompt tuning framework for VLMs. The recent advanced CLIP model is used as a VLM backbone. Given sampled joint intrinsic vectors (individuals), the matrices W_L and W_V project $[z_L \| z_V]$ into text and image prompts p_L and p_V. Through the forward pass of CLIP, these prompts are then evaluated on a downstream few-shot dataset. According to each sampled intrinsic vector's fitness value (loss value), the derivative-free algorithms adjust the distribution and reproduce a new generation of individuals.

dient information, which is similar to model fine-tuning in the scenario of MaaS. Here we mainly review the studies of Evolution Strategies (ESs) - a class of powerful algorithms for black-box optimization. CMA-ES [Hansen and Ostermeier, 2001] is one of the most successful implementations of evolution strategy, which adapts the covariance matrix to approximate the shape of function landscape. The adaptation of CMA-ES aims to increase the probability of reproducing individuals towards promising search directions.

However, the sampling procedure requires the decomposition of the covariance matrix, which leads to $\Theta(n^2)$ time complexity. Researchers have worked to reduce the computation the simplify the adaptation. MA-ES [Beyer and Sendhoff, 2017] removes the evolution path the simplify the update for the Cholesky factor; MM-ES [He *et al.*, 2020] applies fast mixture sampling to approximate the covariance matrix C in a non-recursive manner. Recent studies also explore limited-memory evolution strategy for memory efficient optimization (e.g. LM-MA-ES [Loshchilov *et al.*, 2018]).

3 Approach

This section introduces the details of our proposed BPT-VLM framework, where the recent advanced model CLIP is used as the VLM backbone. In particular, given sampled intrinsic vectors from the shared original parameter space, the projection matrices transform them into image and text prompts. Then the CLIP working as a black-box objective function evaluates the fitness of these intrinsic vectors for ES-based black-box optimization. Different prompt variants are also incorporated for enhancing cross-modal interaction. The details are introduced in the following sections.

3.1 Intrinsic Vector based Population

As discussed by [Li and Liang, 2021] and [Jia *et al.*, 2022], dozens of prompt tokens are required to be learned for single-modality Transformers like RoBERTa [2019] and ViT [2020]. Considering the embedding dimension of CLIP (512 for text encoder and 756 for image encoder), the total parameters of the continuous prompts $P \in \mathbb{R}^{D_1+D_2}$ need to be optimized can be tens of thousands, which is challenging for black-box optimization. While [Aghajanyan *et al.*, 2021] empirically demonstrates that large-scale LM actually has a very low intrinsic dimension, we extend this insight into VLMs to transfer prompt optimization from the original parameter space into a intrinsic subspace.

In particular, we define the parameter subspace of language as $z_L \in \mathbb{R}^{d_1}$ and the visual counterpart as $z_V \in \mathbb{R}^{d_2}$.

$$Z = [z_L \| z_V] \tag{1}$$

with a concatenation operation $\|$, the joint intrinsic parameter subspace is denoted as $Z \in \mathbb{R}^{d_1+d_2}$. Intrinsic vectors belonging to this subspace can be projected to vision-language prompt tokens p_L and p_V through randomly initialized matrices $W_L \in \mathbb{R}^{d_1 \times D_1}$ and $W_V \in \mathbb{R}^{d_2 \times D_2}$.

$$p_L = z_L W_L, p_V = z_V W_V \tag{2}$$

Note that the weights of W_L and W_V are fixed through generations, but directly initializing them with standard uniform

distribution $\mathcal{N}(0, 1/d)$ [Sun *et al.*, 2022a] may result in slow convergence and inferior performance. Thus, to approximate the output distribution of dual encoders' entry embedding and convolutional layers, we use normal distribution to initialize W_L and W_V with means and standard deviations as follows:

$$\begin{cases} \mu_L = \dfrac{\hat{\mu_L}}{d_1 - \hat{\sigma_L}^2}, \sigma_L = \dfrac{\hat{\sigma_L}}{\sqrt{d_1 - \hat{\sigma_L}^2}} \\ \mu_V = \dfrac{3k^2\hat{\mu_V}}{d_2}, \sigma_V = \hat{\sigma_V}\sqrt{\dfrac{3k^2}{d_2}} \end{cases} \tag{3}$$

where $\hat{\mu_L}$ and $\hat{\sigma_L}$ are observed mean and deviation of the word embedding layer in the text encoder, and $\hat{\mu_V}$ and $\hat{\sigma_V}$ are the counterparts of the entry convolutional layer in the image encoder, which uses 3-channel kernels with the size k.

In this way, we're able to perform prompt tuning in a much smaller parameter subspace ($d_1 + d_2 \ll D_1 + D_2$), and multiple intrinsic vectors as individuals can form a population for evolutionary learning.

3.2 Objective Function

In evolution strategies, an objective function is required to evaluate the fitness of each individual, and knowing its precise analytic form is unnecessary. In the scenario of MaaS, it's reasonable to recognize CLIP as a black-box objective function, since only forward pass is allowed and thus cannot compute gradients or the Hessian matrix.

CLIP consists of two Transformer-based encoders - one image encoder and the other text encoder (shown in Figure 2). Assume that CLIP-like models take a batch of texts x_L and images x_V as input, and we recognize the forward pass as a black-box function f, which outputs the similarity scores between each image-text pairs. With the output and the labels Y, we can calculate the cross entropy loss $\mathcal{L}$. Our framework aims to learn the optimal prompts in a derivative-free manner:

$$\begin{cases} f(p_L, p_V) = CLIP[(p_L, x_L); (p_V, x_V)] \\ P^\star = \arg\min_{p_L, p_V} \mathcal{L}[Y; f(p_L, p_V)] \end{cases} \tag{4}$$

where P is the unified formulation of (p_L, p_V). The black-box function f can be evaluated with forward pass, but is not available to the optimizer in a closed form.

As discussed in 3.1, optimization of vision and language prompts actually performs in an intrinsic parameter subspace. Thus the prompt tuning can also be denoted as

$$Z^\star = \arg\min_{z_L, z_V} \mathcal{L}[Y; f(z_L W_L, z_V W_V)] \tag{5}$$

3.3 Black-box Optimization with ES

Optimization in the intrinsic space only requires algorithms to deal with hundreds of parameters. Thus Evolution Strategy (ES) is a reasonable approach to tackle black-box prompt tuning. In the $t+1$-th iteration of CMA-ES, individuals are sampled from the following distribution:

$$Z^{t+1} \sim m^t + \sigma^t \mathcal{N}(0, C^t) \tag{6}$$

where m^t is the mean of top-λ out of all individuals evaluated by the objective function $\mathcal{L}$ (Equation 5), σ^t denotes the mutation step-size. Through generations, the distribution is adapted by adding a random Gaussian mutation defined by a covariance matrix C^t.

However, the original CMA-ES performs eigendecomposition $C = AA^T$ with $\Theta(n^2)$ time complexity to update the covariance matrix C, which precludes its application on large-scale optimization. While Cholesky-CMA-ES samples candidate solutions only with the iteratively updated Cholesky factor A_c:

$$Z^{t+1} \sim m^t + \sigma^t A_c^t \mathcal{N}(0, C^0) \tag{7}$$

which simplifies a lot the implementation of the algorithm. MA-ES further removes the evolution path and simplifies the update for the Cholesky factor; While MM-ES applies Fast Mixture Sampling to approximate the covariance matrix C in a non-recursive manner. The performance comparison of different black-box optimization is shown in Figure 4.

3.4 Prompt Design

Given the vision-language model CLIP, we introduce continuous prompts p_L and p_V to transfer pre-trained knowledge to downstream tasks. However, there are certain differences between its dual encoders (linguistic Transformer and ViT), and the prompt can be involved in multiple Transformer layers. This section explains different prompt designs used in our framework.

Preliminaries

For the text encoder, each word token in the input sentence S is first embedded to a d_L-dimensional subspace through an embedding layer as

$$e_L = Embed(S) \tag{8}$$

where $e_L \in \mathbb{R}^{m_L \times d_L}$ are the embedding features of the text input padded to a fixed length m_L.

On the other hand, for a plain ViT, an input image G will be first divided into m_V patches, each patch is then projected to a d_V-dimensional vector with a convolutional layer

$$e_V = Conv2D(G) \tag{9}$$

where $e_V \in \mathbb{R}^{m_V \times d_V}$ are the convolutional features of the image patches.

Prompt Location

As is widely used and evaluated in previous works [Li and Liang, 2021], prefix positioning is adopted for the text encoder. Text prompt p_L is placed before the sentence embedding e_L, and the number of tokens in p_L is denoted as n_L:

$$I_L = [c_L, p_L, e_L] \tag{10}$$

where c_L is the `[CLS]` token for text inputs, $I_L \in \mathbb{R}^{(1+n_L+m_L)\times d_L}$ represents the input features with prompt for the text encoder.

While for the image encoder (ViT), suffix positioning is applied to keep the information of pre-trained positional embedding, which has a length just matched with m_V image patches. A set of n_V visual tokens is referred as p_V:

$$I_V = [c_V, e_V, p_V] \tag{11}$$

where c_V is the `[CLS]` token for image inputs, and $I_V \in \mathbb{R}^{(1+m_V+n_V)\times d_V}$ denotes the visual features with prompt fed to the image encoder.

Prompt Depth

Since prompts are inserted to transformer-based image and text encoders, they have a strong ability to influence the output through sufficient cross-attention with input features. However, [Liu *et al.*, 2022] point out that long propagation path from prompt to the final signal may lead to information loss. Thus we explore two variants of prompt tuning in BPT-VLM, namely **Shallow Prompt** and **Deep Prompt**, depending on the depth of transformer layers inserted with projected prompts (Figure 3 takes the text encoder for example).

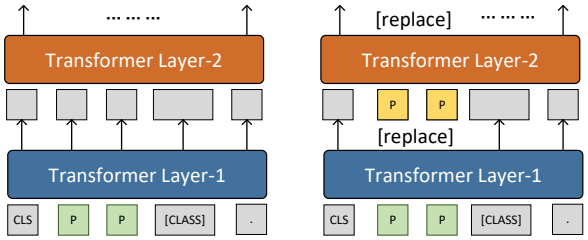


Figure 3: Shallow Prompt and Deep Prompt.

Shallow Prompt only inserts prompts for the input features of text encoder and image encoder (ViT), the propagation process can be formulated as:

$$\begin{cases} [c_L, h_L, e_L]_1 = L_1(I_L) \\ [c_L, h_L, e_L]_i = L_i([c_L, h_L, e_L]_{i-1}) \quad i \geq 2 \end{cases} \tag{12}$$

$$\begin{cases} [c_V, e_V, h_V]_1 = V_1(I_V) \\ [c_V, e_V, h_V]_i = V_i([c_V, e_V, h_V]_{i-1}) \quad i \geq 2 \end{cases} \tag{13}$$

where L_i and V_i indicate the i-th transformer layer of image and text encoder; h_L and h_V refers to the hidden representation of text prompt p_L and p_V. Prompts interacts with input features along layer-upon-layer propagation.

To avoid information loss caused by long propagation path, Deep Prompt involves prompts in each layer of transformer-based text and image encoders:

$$\begin{cases} [c_L, \#, e_L]_i = L_i([c_L, p_L, e_L]_{i-1}) \quad i \geq 1 \\ [c_V, e_V, \#]_i = V_i([c_V, e_V, p_V]_{i-1}) \quad i \geq 1 \end{cases} \tag{14}$$

where $\#$ is the hidden states to be replaced by another set of prompts. Since prompts inserted to different layers belong to different parameter spaces, multiple projection matrices are initialized. For N intrinsic parameter spaces, our framework recognizes the optimization as in-dependent sub-problems:

$$\arg\min\{Z_i^\star\}_{i=1}^N = (\arg\min Z_1^\star, ..., \arg\min Z_N^\star). \tag{15}$$

when one intrinsic space is under the process of black-box tuning, vectors belonging to other intrinsic spaces are kept fixed. Such iterative optimization enhances cross-modal interaction of VLM and generally outperforms shallow prompt.

4 Experiments

4.1 Experiment Setup

Datasets and Metrics

To evaluate the effectiveness of BPT-VLM, we conduct experiments on 9 visual image classification datasets: ImageNet [Deng *et al.*, 2009], Caltech101 [Fei-Fei *et al.*, 2004], OxfordPets [Parkhi *et al.*, 2012], Flowers102 [Nilsback and Zisserman, 2008], Food101 [Bossard *et al.*, 2014], UCF101 [Soomro *et al.*, 2012], SUN397 [Xiao *et al.*, 2010], EuroSAT [Helber *et al.*, 2019] and DTD [Cimpoi *et al.*, 2014]. These datasets covers a wide range of vision tasks.

The ImageNet and Caltech101 datasets are designed for generic image classification. While fine-grained object classification datasets including OxfordPets, Flowers102 and Food101 focus on differentiating between sub-classes belonging to the same meta-class. The SUN397 dataset is built for scene recognition. EuroSAT and DTD are specialized datasets catered for satellite classification and texture recognition. Following the few-shot setting adopted in [Zhou *et al.*, 2022], all methods use the same 16-shot split for prompt tuning and are evaluated on full test-sets for comparison.

Models for Comparison

As presented in Table 2, we compare BPT-VLM with two kinds of prompt tuning methods: gradient-based and derivative-free methods.

For *derivative-free method*, we consider **Manual Prompt** as our baseline: Following the prompt setting introduced in [Radford *et al.*, 2021], we use hand-crafted templates as task-relevant prompts to conduct zero-shot evaluation. For *gradient-based methods*, two baseline methods are considered: **(1) Linear Probe**: As suggested by [Tian *et al.*, 2020], training a linear layer as the classification head on top of CLIP can achieve competitive performance compared with other fine-tuning methods. We followed the same training method used by [Radford *et al.*, 2021] to train the linear probe model. **(2) CoOp**: Recently proposed CoOp [Zhou *et al.*, 2022] models text prompt's context words as learnable vectors while the other parameters of CLIP are kept fixed. We reproduced the results using on each tasks using 16 middle positional tokens and default 200 epoch training.

We devise four versions of BPT-VLM for comparison, namely **MM-ES-Shallow**, **MA-ES-Shallow**, **CMA-ES-Shallow** and **CMA-ES-Deep**, which adopt different evolution strategies [2001; 2020; 2018] and prompt designs (discussed in 3.4) for black-box optimization.

4.2 Implementation Details

Hyper-parameter	Default Setting
Intrinsic Dimension	1000
Vision Prompt Length	8
Language Prompt Length	5
Population Size	30
Loss Fucntion	Cross Entropy

Table 1: Default Setting of Hyper-parameters

Three derivative-free algorithms are introduced to our black-box prompt tuning frameword: CMA-ES, MM-ES and MA-ES, all of which are implemented based on open-source libraries PyCMA[2] and PyPop7[3]. Unless otherwise stated, all

[2]PyCMA: https://github.com/CMA-ES/pycma

[3]PyPop7: https://github.com/Evolutionary-Intelligence/pypop

		ImageNet	Caltech101	OxfordPets	StanfordCars	Food101	SUN397	DTD	EuroSAT	UCF101	**Average**
Gradient-Based	Linear Probe	55.87	90.63	76.42	70.08	70.17	67.15	63.97	82.76	73.72	72.31
	CoOp	62.95	91.83	87.01	73.36	74.67	69.26	63.58	83.53	75.71	75.77
Derivative-Free	Manual Prompt	58.18	86.29	85.77	55.61	77.31	58.52	42.32	37.56	61.46	62.56
	MM-ES-Shallow	–	93.67	90.49	62.49	**81.62**	–	48.40	86.25	70.76	–
	MA-ES-Shallow	–	93.59	90.57	65.03	81.54	–	59.63	86.93	76.34	–
	CMA-ES-Shallow	**65.08**	**94.16**	90.43	64.72	81.31	68.01	60.52	86.11	74.62	76.11
	CMA-ES-Deep	64.84	93.39	**90.62**	**67.84**	81.38	**69.83**	**64.13**	**89.37**	**76.66**	**77.56**
	$\Delta(\%)$	**+2.13**	**+2.33**	**+3.61**	**-5.52**	**+4.31**	**+0.57**	**+0.16**	**+5.84**	**+0.95**	**+1.79**

Table 2: Comparison results on various visual understanding tasks with the pre-trained vision-language model CLIP in 16-shot setting. MM-ES-Shallow, MA-ES-Shallow, CMA-ES-Shallow and CMA-ES-Deep are four versions of our BPT-VLM with different evolution strategies and prompt designs. The underlined values indicate the highest accuracy for gradient-based methods on each dataset, and the bold numbers are the counterparts for our BPT-VLM models. The last line Δ values indicate our maximum improvements over the best baselines.

models are built with the open-source CLIP with ViT-B/32 as the visual encoder's backbone. Table 1 demonstrates the default configuration of hyper-parameters used in our experiments. Note that only the intrinsic vector is required to be updated and no back-propagation is performed.

Since the optimization process of gradient descent algorithm and derivative-free algorithm are quite different, here we redefine the meaning of a training epoch. For the shallow variant of black-box prompt tuning, every 12 generations of evolution for the intrinsic vector are regarded as 1 training epoch. As for the deep prompt variant, 1 epoch of training indicates that each intrinsic vector (belonging to 12 transformer layers) iteratively performs 1 generation of evolution.

4.3 Main Results

Overall Performance. Table 2 shows the comparison results of baseline model and four versions of BPT-VLM framework (introduced in 4.1). We observe that the CMA-ES-Deep model using CMA-ES optimization and Deep Prompt outperforms both the derivative-free and gradient-based baselines on 8 out of 9 datasets, indicating the effectiveness of our black-box prompt tuning framework for large-scale VLMs. In particular, we achieve an average improvement of 1.79% over recent advanced baseline CoOp that uses gradient-based prompt tuning, which further demonstrates the advantage of our framework in the scenario of MaaS.

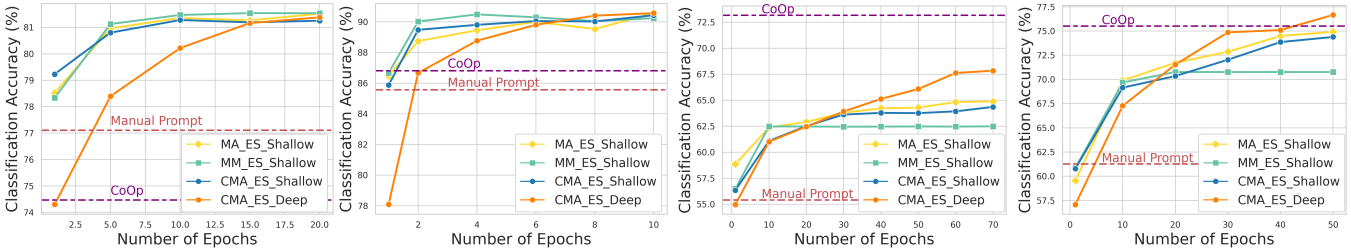
Performance on Various Datasets. It is worth noting that we achieve 2.13% improvement over the best gradient-based baseline on ImageNet, which is a challenging task that contains 1000 classes. The performance improvements are also significant on fine-grained classification datasets such as OxfordPets and Food101, as well as scene and action recognition datasets (i.e. SUN397 and UCF-101). The performance of our method on StanfordCars is not so appealing, thus we further analyze the failure cases. It is observed that our method under-performs the gradient-based methods when the text annotations only have subtle differences like "BMW X3 SUV" and "BMW X5 SUV", which increase the difficulty for the derivative-free methods without strong fitting mechanisms like gradient descent.

Performance of Various Optimization Algorithms and Prompt Designs. We plot the accuracy curves with different optimization algorithms and prompting methods during training in Figure 4. Since the results on each dataset is similar, the curves on four datasets are presented due to the limited space. We observe that the MM-ES algorithm performs slightly worse while converges faster, and there are no significant differences between other optimization algorithms. In addition, by comparing the two models as CMA-ES-Shallow and CMA-ES-Deep that use different prompting methods, we find that the shallow prompt achieves faster convergence speed, while slightly under-performs the deep variant on final accuracy. This coincides with our intuition since the deep prompt is iteratively optimized in different intrinsic subspaces involving each layer of the pre-trained VLMs, while the shallow prompt only involves the input layer that includes less parameters to be optimized as described in Section 3.4.

4.4 Further Analyses

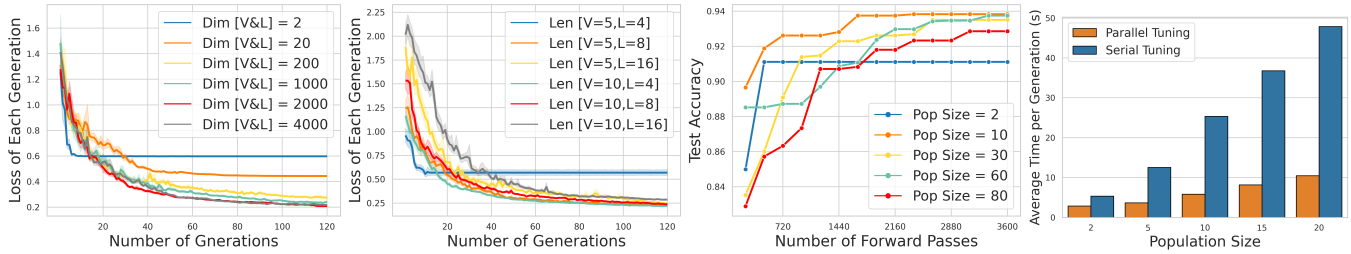
We conduct further analyses on various hyper-parameters to explore their effect. Each hyper-parameter is investigated while keeping the other hyper-parameters as default as listed in Table 1. For the limits of space, we show the experimental results on Caltech101 with the same 16-shot split in Figure 5. Similar results can be observed on other datasets.

Effect of Intrinsic Dimension. The parameter subspace of the intrinsic vectors is the space where optimization actually performs. As shown in Figure 5(a), the model with lower intrinsic dimension converges faster, but yields higher losses. When the intrinsic dimension increases over 1000, there are no significant differences in losses. Thus, the intrinsic dimension is recommended to be set to 1000 in our experiments.



(a) Food-101 (b) OxfordPets (c) StanfordCars (d) UCF-101

Figure 4: Accuracy curves on various datasets during training. Since one epoch of training in gradient-based and derivative-free methods represent different meanings as demonstrated in Section 4.2, the final performance of the recent advanced gradient-based method CoOp is presented with horizontal purple dash lines. The derivative-free baseline as manual prompt evaluated in zero-shot setting can be deemed as the initial points of our method, represented with horizontal red dash lines.



(a) Intrinsic Dimension (b) Prompt Length (c) Population Size (d) Parallel Evaluation

Figure 5: Results of ablation experiments on hyper-parameters. (a) and (b) illustrate the loss curves of different intrinsic dimension and prompt length settings; (c) shows the accuracy curves of black-box prompt tuning with various population sizes; (d) demonstrates comparison of time efficiency between serial and parallel evaluation.

Effect of Prompt Length. We utilize text prompt and image prompt for better inducing the knowledge contained in the pre-trained VLMs, which determine the original parameter space in text and image encoders. To investigate the effect of prompt length, we evaluate BPT-VLM with various lengths. As shown in Figure 5(b), the converged loss decreases when the prompt length grows first, but then it tends to be stable when the length becomes larger. Given that the model converges slowly with the increasing prompt length, we use a reliable setting as (V=5, L=8) for the text and prompt length for performance and efficiency balance.

Effect of Population Size. In one generation of the evolution strategy, individuals (intrinsic vectors) are sampled from a multivariate normal distribution. Then the algorithm evaluates their fitness to adjust the distribution for the next generation. Here we explore the effect of population size within 3600 times of individual evaluations. As suggested in Figure 5(c), a smaller population size as 2 can reach the convergence faster, but a noticeable performance degradation occurs. In addition, the larger population size does not always yield better results but converges more slowly. On the whole, the population size between 10 and 60 is recommended to be a reliable setting in our experiments.

Efficiency of Parallel Training. Open-source black-box optimization algorithms usually evaluate individuals in a time-consuming serial manner. To improve efficiency, our method can be easily adapted to support parallel prompt tuning, which can evaluate all individuals simultaneously in a single concatenated batch. Figure 5(d) shows the result of the efficiency analysis. It is worth noting that parallel tuning only takes about 1/4 time compared with traditional serial tuning when the population size varies in $\{5, 10, 15, 20\}$, and this advantage in training efficiency will be more prominent when the population size grows.

5 Conclusions and Future Work

In this paper, we propose a black-box prompt tuning framework for vision-language models in the scenario of MaaS. We extend derivative-free algorithms to the new scope of prompt tuning for pre-trained VLMs, and conduct expansive experiments to verify that large-scale VLMs also have very low intrinsic dimensions, which is as effective for fine-tuning as the full parameter spaces. Moreover, we incorporate cross-modal interaction in our framework by sharing the intrinsic parameter subspace of both vision and language modalities. Different prompt designs are also explored to enhance prompts' influence during propagation.

It could be interesting to consider black-box prompt tuning in different downstream tasks based on different pre-trained models, and the black-box algorithms could be exchanged, such as Particle Swarm and Bayesian Optimization. Our future work will also concern a deeper analysis of cross-modal interaction from the perspective of derivative-free optimization in shared intrinsic parameter spaces.

Acknowledgements

This research is funded by the National Key Research and Development Program of China (No. 2021ZD0114002), and the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901, No. 21511100402).

References

[Aghajanyan *et al.*, 2021] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, 2021.

[Beyer and Sendhoff, 2017] Hans-Georg Beyer and Bernhard Sendhoff. Simplify your covariance matrix adaptation evolution strategy. *IEEE Transactions on Evolutionary Computation*, 21(5):746–759, 2017.

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

[Brown *et al.*, 2020] 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.

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

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

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

[Du *et al.*, 2022] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

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

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

[He *et al.*, 2020] Xiaoyu He, Zibin Zheng, and Yuren Zhou. Mmes: Mixture model-based evolution strategy for large-scale optimization. *IEEE Transactions on Evolutionary Computation*, 25(2):320–333, 2020.

[Helber *et al.*, 2019] 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.

[Houlsby *et al.*, 2019] 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.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.

[Larson *et al.*, 2019] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[Liu *et al.*, 2022] Xiangyang Liu, Tianxiang Sun, Xuanjing Huang, and Xipeng Qiu. Late prompt tuning: A late prompt could be better than many prompts. *arXiv preprint arXiv:2210.11292*, 2022.

[Loshchilov *et al.*, 2018] Ilya Loshchilov, Tobias Glasmachers, and Hans-Georg Beyer. Large scale black-box optimization by limited-memory matrix adaptation. *IEEE Transactions on Evolutionary Computation*, 23(2):353–358, 2018.

[Loshchilov, 2014] Ilya Loshchilov. A computationally efficient limited memory cma-es for large scale optimization. In *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, pages 397–404, 2014.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over

a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

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

[Radford *et al.*, 2021] 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.

[Shahriari *et al.*, 2015] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[Soomro *et al.*, 2012] 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.

[Sun *et al.*, 2022a] 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.

[Sun *et al.*, 2022b] 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.

[Tian *et al.*, 2020] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020.

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

[Zhou *et al.*, 2022] 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.