# DECOOP: Robust Prompt Tuning with Out-of-Distribution Detection

**Zhi Zhou** [1] **Ming Yang** [1 2] **Jiang-Xin Shi** [1 2] **Lan-Zhe Guo** [1 3] **Yu-Feng Li** [1 2]

## Abstract

Vision-language models (VLMs), such as CLIP, have demonstrated impressive zero-shot capabilities for various downstream tasks. Their performance can be further enhanced through few-shot prompt tuning methods. However, current studies evaluate the performance of learned prompts separately on base and new classes. This evaluation lacks practicality for real-world applications since downstream tasks cannot determine whether the data belongs to base or new classes in advance. In this paper, we explore a problem setting called ***Open-world Prompt Tuning*** (OPT), which involves tuning prompts on base classes and evaluating on a combination of base and new classes. By introducing ***Decomposed Prompt Tuning*** framework (DEPT), we theoretically demonstrate that OPT can be solved by incorporating out-of-distribution detection into prompt tuning, thereby enhancing the base-to-new discriminability. Based on DEPT, we present a novel prompt tuning approach, namely, ***Decomposed Context Optimization*** (DECOOP), which introduces new-class detectors and sub-classifiers to further enhance the base-class and new-class discriminability. Experimental results on 11 benchmark datasets validate the effectiveness of DEPT and demonstrate that DECOOP outperforms state-of-the-art methods, providing a significant 2% average accuracy improvement.

## 1. Introduction

Vision-language models (VLMs), such as CLIP (Radford et al., 2021), have been developed to align images and language, demonstrating impressive zero-shot capabilities for a variety of downstream tasks (Deng et al., 2009; Maji et al., 2013; Krause et al., 2013), using only class names. The classification prediction is determined by calculating the cosine similarity between the image embedding, generated by the image encoder, and the text embedding, generated by the text encoder, using prompting techniques (Liu et al., 2023). For example, by inputting "a photo of class", the text encoder generates the corresponding text embedding, where "class" represents the class name.

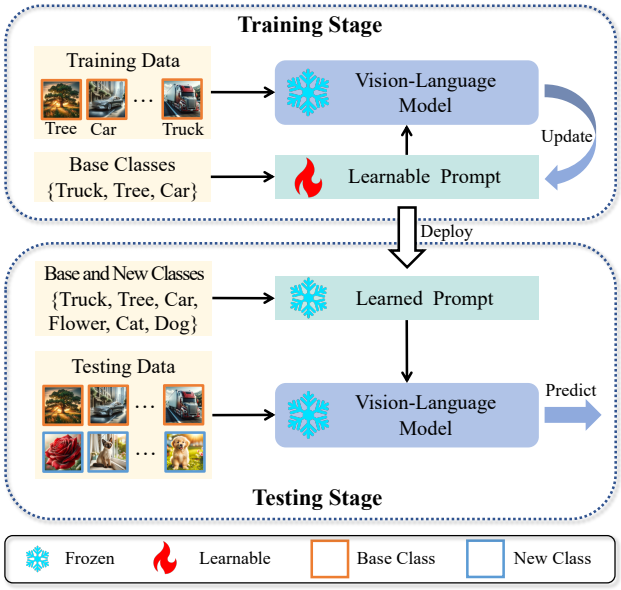
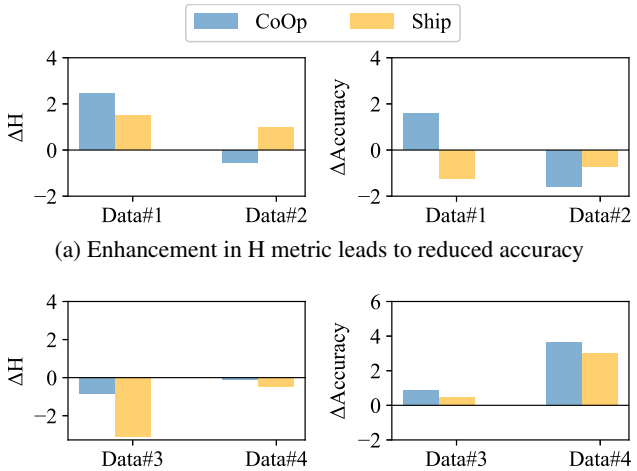



*Figure 1.* An illustration of the OPT evaluation paradigm. During the training, we finetune the model with data from base classes. During the testing, we evaluate the model on a mix of base and new classes.

In addition, it is possible to improve the performance of CLIP, particularly when dealing with downstream tasks that have limited labeled data. Few-shot prompt tuning methods (Lu et al., 2022; Zhou et al., 2022b; Shu et al., 2022b) utilize a small amount of labeled data from downstream datasets to fine-tune learnable prompts while keeping the other parameters unchanged. These approaches can yield substantial performance improvement compared to the zero-shot VLMs in downstream classification tasks. However, previous studies (Zhou et al., 2022a; Wang et al., 2023b)

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China [2]School of Artificial Intelligence, Nanjing University, China [3]School of Intelligence Science and Technology, Nanjing University, China. Correspondence to: Lan-Zhe Guo <guolz@nju.edu.cn>, Yu-Feng Li <liyf@nju.edu.cn>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).




(a) Enhancement in H metric leads to reduced accuracy

(b) Deterioration in H metric leads to improved accuracy

*Figure 2.* Delta performance of COOP and SHIP method compared to zero-shot baseline CLIP method. Subfigres (a) and (b) show that the changes in the H metric are not necessary indicators of performance improvements or degradation of accuracy, highlighting the significance of addressing the OPT problem.

have identified a limitation in which the learned prompts only operate effectively with labeled data from base classes. This limitation leads to a decrease in zero-shot performance for new classes which are unseen in the training set. To address this, the researchers propose an evaluation paradigm that assesses the performance of both base and new classes separately, as well as their harmonic mean, i.e., H metric.

Although this evaluation paradigm can comprehensively evaluate the performance of both base and new classes, it lacks practicality for real-world applications, which necessitate prior knowledge of whether the data belongs to base or new classes in the downstream task. For instance, in the context of biological underpinnings (Hayes et al., 2021; Kudithipudi et al., 2022) and visual classification (Lange et al., 2022; Mai et al., 2022), both base classes and new classes that arise during testing will be evaluated together. Therefore, we introduce a realistic problem setting, namely, ***O**pen-world **P**rompt **T**uning* (OPT), which evaluates the performance of the model on a mix of base and new classes while training model with base classes. An illustration of the OPT problem is shown in Figure 1. The results in Figure 2 show that the changes in the H metric are not necessary indicators of performance improvement or degradation when evaluating the combination of base and new classes, which highlights the significance of the OPT problem.

To address the OPT problem, we first analyze the original problem, which consists of three parts: base-to-new discriminability, base-class discriminability, and new-class discriminability. We observe that existing methods and settings fail to adequately consider the base-to-new discriminability. Motivated by this analysis, we propose the DEPT framework, which incorporates out-of-distribution (OOD) detection into prompt tuning to enhance the base-to-new discriminability and thereby prevents performance degradation on new classes. We theoretically prove that the DEPT framework can improve performance compared to the zero-shot baseline and prompt tuning methods. Building upon the DEPT framework, we introduce a novel prompt tuning approach called ***De**composed **Co**ntext **Op**timization* (DECOOP). This approach incorporates new-class detectors and sub-classifiers to further enhance the base-class and new-class discriminability, respectively. Empirical results validate the effectiveness of the DEPT framework and demonstrate that DECOOP approach outperforms current state-of-the-art (SOTA) methods by a significant margin.

The contributions of this paper are summarized as follows:

(1) We explore a practical OPT problem and break down the problem into two sub-problems: OOD detection and prompt tuning. Through decomposition, we uncover that base-to-new discriminability is crucial to address OPT, overlooked in existing methods and settings.

(2) We propose a novel DEPT framework, which introduces OOD detection into prompt tuning. Both our theoretical analysis and experimental results demonstrate the effectiveness of DEPT framework for OPT.

(3) Based on DEPT framework, we propose a novel prompt tuning approach DECOOP, which additionally enhances the base-class and new-class discriminability by introducing new-class detectors and sub-classifiers.

(4) We conduct comprehensive experiments on DECOOP using 11 benchmark datasets. The results show that our proposed scheme outperforms current SOTA comparison methods, delivering a significant 2% average improvement in accuracy.

## 2. Problem and Analysis

In this section, we first describe the notions and problem formulation for the OPT setting. Subsequently, we conduct an empirical analysis using a real-world dataset (Krause et al., 2013), wherein we identify two primary challenges to address: base-to-new discriminability and new-class discriminability. Finally, we decompose the original problem to demonstrate that the incorporation of the OOD detection technique can effectively resolve these two challenges.

### 2.1. Problem Formulation

We focus on the prompt tuning setting for multi-class classification problems that involve an input space $\mathcal{X}$, a class space $\mathcal{Y} = \mathcal{Y}_{\text{b}} \cup \mathcal{Y}_{\text{n}} = [C]$, and the text space $\mathcal{T}$, where $C$ represents the number of classes. Here, $\mathcal{Y}_{\text{b}}$ denotes the set of

base classes, and $\mathcal{Y}_{\text{n}}$ represents the set of new classes. The name of the $i$-th class is denoted as $\boldsymbol{t}_i \in \mathcal{T}$. Furthermore, $\boldsymbol{x} \in \mathcal{X}$ represents the data. $f(\boldsymbol{x}) \in \mathcal{Y}$ and $g(\boldsymbol{x}) \in \{\text{b}, \text{n}\}$ denote the label of $\boldsymbol{x}$ and the specific class space to which it belongs, where $f$ and $g$ are the mapping functions of the ground truth of the labels and the class space.

In OPT problem, we are given a pre-trained vision-language model $\mathcal{F} = \{\text{E}_V, \text{E}_T\}$, which consists of a visual encoder $\text{E}_V : \mathcal{X} \mapsto \mathbb{R}^d$ and a textual encoder $\text{E}_T : \mathcal{T} \mapsto \mathbb{R}^d$, where $d$ represents the dimension of model $\mathcal{F}$. During the training stage, we learn the prompt vector $\boldsymbol{p}$ on a few-shot dataset $\mathcal{D}$ containing data derived from $\mathcal{Y}_{\text{b}}$. To simplify the notation, we define $t_i(\boldsymbol{p})$ as the concatenation of the tokens of the class name $\boldsymbol{t}_i$ and the learned prompt $\boldsymbol{p}$. Consequently, weight vectors $\{\boldsymbol{w}_i(\boldsymbol{p})\}_{i=1}^{C}$ are generated for each class as textual embeddings, where $\boldsymbol{w}_i(\boldsymbol{p}) = \text{E}_T\left(t_i(\boldsymbol{p})\right) / \|\text{E}_T\left(t_i(\boldsymbol{p})\right)\|$. In the testing stage, given the test data $\boldsymbol{x}$ drawn from $\mathcal{Y}$, we initially obtain its visual embedding $\boldsymbol{z} = \text{E}_V(\boldsymbol{x}) / \|\text{E}_V(\boldsymbol{x})\|$. Subsequently, we calculate the prediction probabilities as follows:

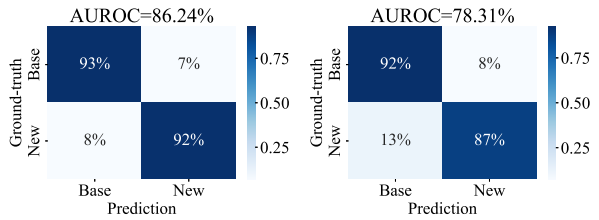
$$P(y|\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{z}^T \boldsymbol{w}_y / \tau\right)}{\sum_{i=1}^{C} \exp\left(\boldsymbol{z}^T \boldsymbol{w}_i / \tau\right)} \tag{1}$$

where $\tau$ represents the temperature determined by VLMs. For convenience, we will also use $P(\boldsymbol{x})$ to represent $P(y|\boldsymbol{x})$ in the subsequent paper. The prediction for $\boldsymbol{x}$ is given by $\arg\max_{y \in \mathcal{Y}} P(y|\boldsymbol{x})$. The objective of OPT is to train a model that can make robust predictions on $\mathcal{Y}$, which includes both base and new classes, without experiencing overall performance degradation due to the presence of new classes. In our following analyses and experiments, we perform a comparison between the zero-shot baseline method (referred to as ZS) and the prompt tuning method (referred to as PT) on OPT problem.

### 2.2. Problem Analysis

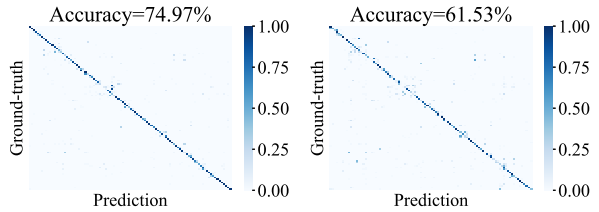
To tackle the OPT problem, we investigate a real-world dataset (Krause et al., 2013) to conduct detailed analyses of the challenges inherent in OPT. Our observation demonstrates that while prompt tuning methods can improve base-class discriminability, they compromise both base-to-new discriminability and new-class discriminability. To illustrate this observation, we present a comparison between the ZS methods and PT methods, where we employ CLIP as ZS method and COOP as PT method, in Figures 3 and 4.

Figure 3 indicates that the prompt tuning method results in a decreased base-to-new discriminability compared to the zero-shot baseline. Specifically, the AUROC for detecting new classes using the MSP technique (Hendrycks & Gimpel, 2016) decreases, and more false positive predictions are introduced for base classes. Moreover, Figure 4 illustrates




(a) Zero-shot Baseline ZS (b) Prompt Tuning Method PT

*Figure 3.* Performance of ZS and PT methods to distinguish data from base classes and new classes (base-to-new discriminability).




(a) Zero-shot Baseline ZS (b) Prompt Tuning Method PT

*Figure 4.* Performance of ZS and PT methods to distinguish data within new classes (new-class discriminability).

that the prompt tuning method also exhibits reduced new-class discriminability compared to the zero-shot baseline.

We emphasize that the existing H metric is incapable of measuring base-to-new discriminability, making it unsuitable for comprehensive practical applications. In OPT problem, the accuracy evaluated in the entire class space can effectively address this limitation.

### 2.3. Problem Decomposition

The above analysis reveals that the zero-shot baseline surpasses the prompt tuning method in terms of both new-class discriminability and base-to-new discriminability. This observation motivates us to incorporate OOD detection technique to combine ZS method and PT method. This approach aims to preserve the new-class discriminability using ZS while enhancing the base-class discriminability using PT. Therefore, we decompose the original classification problem into separate OOD detection and two classification problems:

$$\begin{aligned} P\left(y|\boldsymbol{x}\right) &= \sum_{i \in \{\text{b},\text{n}\}} P\left(y|y \in \mathcal{Y}_i, \boldsymbol{x}\right) \cdot P\left(y \in \mathcal{Y}_i|\boldsymbol{x}\right) \\ &= P\left(y|y \in \mathcal{Y}_k, \boldsymbol{x}\right) \cdot P\left(y \in \mathcal{Y}_k|\boldsymbol{x}\right) \end{aligned} \tag{2}$$

where $k$ always equals $g(\boldsymbol{x})$ for the sake of simplicity, representing the ground-truth label space of $\boldsymbol{x}$. The second term is an OOD detector to determine whether $\boldsymbol{x}$ belongs to the base or new class space. The first term is a classifier for the corresponding class space.

Equation 2 motivates us to propose a novel ***De**composed*

***Prompt Tuning*** framework (DEPT), which synergistically leverages the advantages of both the zero-shot baseline ZS and the prompt tuning method PT. The prediction probability $P_{\text{DEPT}}(y|\boldsymbol{x})$ of DEPT framework is:

$$\begin{cases} P_{\text{PT}}(y|\boldsymbol{x}), & P_{\text{OOD}}(y \in \mathcal{Y}_{\text{b}}|\boldsymbol{x}) \geq P_{\text{OOD}}(y \in \mathcal{Y}_{\text{n}}|\boldsymbol{x}), \\ P_{\text{ZS}}(y|\boldsymbol{x}), & P_{\text{OOD}}(y \in \mathcal{Y}_{\text{b}}|\boldsymbol{x}) < P_{\text{OOD}}(y \in \mathcal{Y}_{\text{n}}|\boldsymbol{x}). \end{cases} \tag{3}$$

where $P_{\text{OOD}}(y \in \mathcal{Y}_b|\boldsymbol{x})$ is the OOD detector to determine whether $\boldsymbol{x}$ belongs to the base or new class space. $P_{\text{ZS}}(y|\boldsymbol{x})$ and $P_{\text{PT}}(y|\boldsymbol{x})$ are classifiers of ZS and PT. In following theoretical analysis and empirical experiment, we adopt the ZS method using MSP method as the OOD detector, i.e., $P_{\text{OOD}}(y \in \mathcal{Y}_i|\boldsymbol{x}) = \max_{j \in \mathcal{Y}_i} P_{\text{ZS}}(y = j|\boldsymbol{x})$ for $i \in \{\text{b}, \text{n}\}$.

Then, we adopt the cross-entropy metric of two probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, i.e., $H(\boldsymbol{p}, \boldsymbol{q}) = -\sum_{i=1}^{C} p_i \log q_i$, to evaluate the performance of $P_{\text{ZS}}(y|\boldsymbol{x})$ and our DEPT framework $P_{\text{DEPT}}(y|\boldsymbol{x})$. We denote distributions $\tilde{\boldsymbol{k}} = \{\mathbb{I}[k = \text{b}], \mathbb{I}[k = \text{n}]\}$ and $\tilde{\boldsymbol{y}} = \{[f(\boldsymbol{x}) = i]\}_{i=1}^{C}$ for $\boldsymbol{x}$. Finally, we denote the following cross-entropy values for zero-shot baseline, prompt tuning method, and DEPT framework:

$$\begin{aligned} H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{k}}, \{P_{\text{ZS}}(y \in \mathcal{Y}_i|\boldsymbol{x})\}_{i=\{\text{b},\text{n}\}}\right), \\ H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{ZS}}(y = j|y \in \mathcal{Y}_k, \boldsymbol{x})\}_{j=1}^{C}\right), \\ H_{\text{PT}}^{\text{CLS}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{PT}}(y = j|y \in \mathcal{Y}_k, \boldsymbol{x})\}_{j=1}^{C}\right), \\ H_{\text{ZS}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{ZS}}(y = j|\boldsymbol{x})\}_{j=1}^{C}\right), \\ H_{\text{DEPT}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{DEPT}}(y = j|\boldsymbol{x})\}_{j=1}^{C}\right). \end{aligned} \tag{4}$$

**Theorem 2.1.** *If $\mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x})\right] \leq \delta$ for $\boldsymbol{x}$ belonging to both base and new classes, $\mathbb{E}_{\boldsymbol{x}}\left[H_{\text{PT}}^{\text{CLS}}(\boldsymbol{x})\right] \leq \delta - \Delta$ for $\boldsymbol{x}$ belonging to base classes, and $\mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x})\right] \leq \epsilon$, given a uniform mixing ratio ($\alpha : 1 - \alpha$) of base classes and new classes in the testing data, we can determine that:*

$$\begin{cases} \mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}(\boldsymbol{x})\right] & \leq \epsilon + \delta, \\ \mathbb{E}_{\boldsymbol{x}}\left[H_{\text{DEPT}}(\boldsymbol{x})\right] & \leq \epsilon + \delta - \alpha \cdot \Delta. \end{cases} \tag{5}$$

*Remark* 2.2. Theorem 2.1 demonstrates that decomposing the zero-shot baseline into an OOD detector and classifiers, and incorporating prompt tuning methods to aid in classifying base classes, can effectively decrease the upper bound of classification error. Moreover, enhancing the reliability of the OOD detector helps reduce the error term $\epsilon$ and ensures that the performance on new classes remains uncompromised compared to the baseline method. Consequently, this framework preserves base-to-new discriminability and new-class discriminability of ZS method. Additionally, refining the PT method increases $\Delta$, further enhancing base-class discriminability and reducing the upper bound of error.

The proof is presented in Appendix A. Theorem 2.1 motivates us to design a robust prompt tuning method based on Equation 3 using OOD detection techniques to solve OPT.

## 3. DECOOP Approach

We propose a novel prompt tuning framework, called DEPT, to address the OPT problem. The DEPT framework effectively maintains the discriminability between base classes and new classes, thus preventing degradation of discriminability when prompt tuning is applied. Our theoretical analysis, as presented in Theorem 2.1, demonstrates the superiority of DEPT when combining the zero-shot baseline and prompt tuning method. However, there are still two challenges that need to be addressed in order to further enhance the performance in complex real-world applications: (1) How can we train reliable OOD detectors to identify new-class data using limited labeled data from base classes? (2) With reliable OOD detectors, how to separately improve the base-class and new-class discriminability?

To tackle the challenges above, we present a novel prompt tuning approach named ***De**composed **Co**ntext **Op**timization* (DECOOP) based on our DEPT framework, containing $K$ new-classes detectors $\{\mathcal{M}_D^i\}_{i=1}^{K}$ and sub-classifiers $\{\mathcal{M}_C^i\}_{i=1}^{K}$. The introduction of new-class detectors aids in the improved detection of data from new classes in OPT problem, where the names of new classes are known and can be utilized. This differs from the traditional OOD detection problems and presents an opportunity for further performance enhancement. The sub-classifiers are designed to better classify the data from base classes and reduce the potential risks for new classes, which aims to enhance the base-class and new-class discriminability with a reliable base-to-new discriminability. The overall illustration of DECOOP approach is shown in Figure 5 and each component is described thoroughly in the following subsections.

### 3.1. New-class Detector $\mathcal{M}_D$

In the OPT problem, the model is trained with $\mathcal{Y}_{\text{b}}$ but has knowledge of the entire class space $\mathcal{Y}$ during testing. Therefore, the main challenge for new class detectors is to train the model to effectively utilize the knowledge of the new class $\mathcal{Y}_{\text{n}}$, which is only known during testing.

Specifically, Our proposed solution incorporates a leave-out strategy which divides the base class space $\mathcal{Y}_{\text{b}}$ into two distinct subsets during training stage: simulated base classes $\hat{\mathcal{Y}}_{\text{b}}$ and simulated new classes $\hat{\mathcal{Y}}_{\text{n}}$, where $\hat{\mathcal{Y}}_{\text{b}} \cup \hat{\mathcal{Y}}_{\text{n}} = \mathcal{Y}_{\text{b}}$. Respectively, we split the original training set $\mathcal{D}$ into $\mathcal{D}_{\text{b}} = \{(\boldsymbol{x}, y)|(\boldsymbol{x}, y) \sim \mathcal{D} \wedge y \in \hat{\mathcal{Y}}_{\text{b}}\}$ and $\mathcal{D}_{\text{n}} = \{(\boldsymbol{x}, y)|(\boldsymbol{x}, y) \sim \mathcal{D} \wedge y \in \hat{\mathcal{Y}}_{\text{n}}\}$. Then, our optimization objective function for the new-class detector is defined as:

$$\begin{aligned} \ell_{\text{OOD}} = &\frac{1}{|\mathcal{D}_{\text{b}}|} \sum_{(\boldsymbol{x}, y) \sim \mathcal{D}_{\text{b}}} \ell_{CE}(\boldsymbol{x}, y) \\ &+ \max\left\{0, \gamma + \ell_E^{\text{b}} - \ell_E^{\text{n}}\right\} \end{aligned} \tag{6}$$

where $\ell_{CE}(\boldsymbol{x}, y) = -\log P(\boldsymbol{x})_y$ represents the cross-

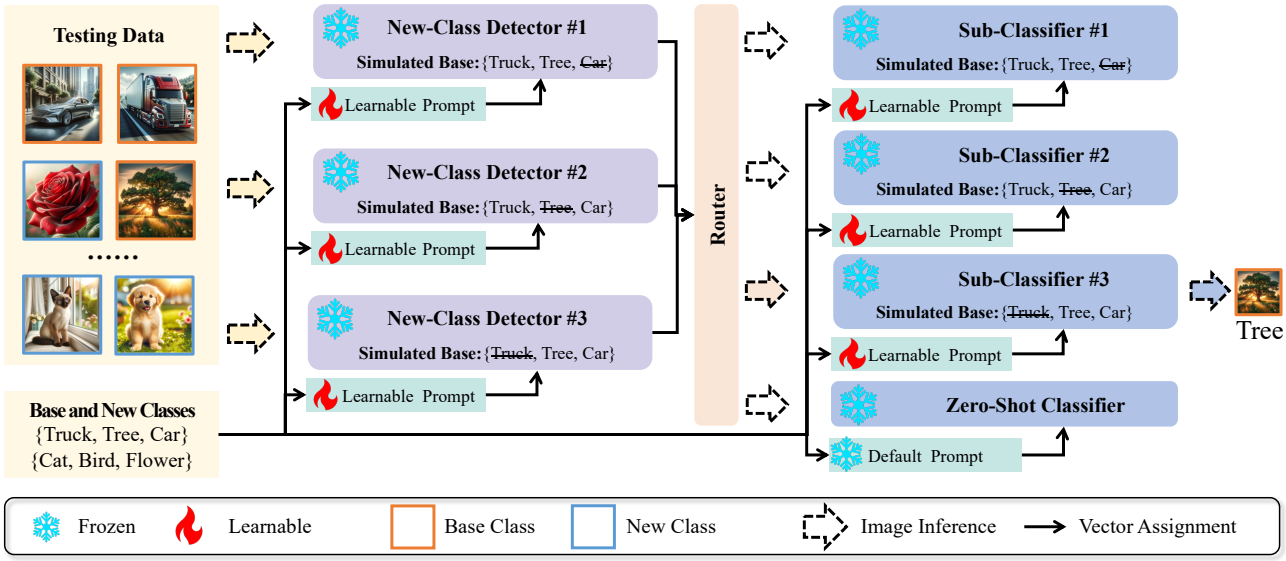



*Figure 5.* The overall illustration of DECOOP approach.

entropy loss, $\ell_E(\boldsymbol{x}) = -\sum_{i=1}^{C} P(\boldsymbol{x})_i \log P(\boldsymbol{x})_i$ represents the entropy loss, $\ell_E^{\text{b}} = \frac{1}{|\mathcal{D}_{\text{b}}|}\sum_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{b}}} \ell_E(\boldsymbol{x})$ represents the average entropy on the simulated base classes, and $\ell_E^{\text{n}} = \frac{1}{|\mathcal{D}_{\text{n}}|}\sum_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{n}}} \ell_E(\boldsymbol{x})$ represents the average entropy on the simulated new classes. Additionally, $\gamma$ is a hyper-parameter that controls the margin between $\ell_E^{\text{b}}$ and $\ell_E^{\text{n}}$ to ensure stable optimization. The objective function in Equation 6 encourages the model to make low-entropy predictions on simulated base classes and high-entropy predictions on simulated new classes, thereby enhancing base-to-new discriminability. However, partitioning the base class space causes the model's cognition to be limited to a subset of base classes, leading to the failure to distinguish between other base classes and new classes during testing. To address this issue, we propose the adoption of an ensemble of $K$ new-class detectors $\{\mathcal{M}_D^i\}_{i=1}^K$ that cover the entire base class space during training. Each new-class detector is trained with Equation 6 with a different class partition. Our class partitions of $K$ new-class detectors ensure each base class is considered as a simulated new class for at least one new-class detector. We denote $\mathcal{M}_D^i(\boldsymbol{x})$ as the new-class score computed for $\boldsymbol{x}$. Lower scores indicate a higher likelihood that $\boldsymbol{x}$ belongs to new classes.

In addition, a threshold remains crucial for the detection of new classes, even when well-trained new-class detectors are provided. Leveraging the benefits of our partition and ensemble strategy, we can directly estimate the threshold for each new-class detector during training using the Otsu algorithm (Otsu, 1979; Liu & Yu, 2009) and training data. This is possible due to the presence of naturally simulated base classes and new classes in the training data for each new-class detector. Subsequently, these estimated thresholds can be averaged to determine the threshold value, denoted as $\tau$, for testing.

### 3.2. Sub-Classifier $\mathcal{M}_C$

After training reliable new-class detectors, we proceed to train a sub-classifier for each detector, as each detector focuses on a specific subset of the base class space. Each of the $K$ sub-classifiers, denoted as $\{\mathcal{M}_C^i\}_{i=1}^K$, is designed to specialize in a particular base class space, thereby achieving better discriminability for the corresponding subset class space. For the $i$-th sub-classifier $\mathcal{M}_C^i$, we first utilize the trained new-class detector $\mathcal{M}_D^i$ partition the training data into two subsets: $\mathcal{D}_{\text{b}}^i$ and $\mathcal{D}_{\text{n}}^i$. Here, $\mathcal{D}_{\text{b}}^i = \{(\boldsymbol{x},y)|(\boldsymbol{x},y)\sim\mathcal{D} \wedge \mathcal{M}_D^i(\boldsymbol{x}) \geq \tau\}$ and $\mathcal{D}_{\text{n}}^i = \{(\boldsymbol{x},y)|(\boldsymbol{x},y)\sim\mathcal{D} \wedge \mathcal{M}_D^i(\boldsymbol{x}) < \tau\}$. Subsequently, we optimize the following objective function:

$$\ell_{\text{CLS}} = \sum_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{b}}^i} \ell_{CE}(\boldsymbol{x},y) + \sum_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{n}}^i} \ell_{KL}\left(P(\boldsymbol{x}), P_{\text{ZS}}(\boldsymbol{x})\right) \tag{7}$$

Here, $\ell_{KL}$ denotes KL-divergence loss, and $P(\boldsymbol{x})$ and $P_{\text{ZS}}(\boldsymbol{x})$ represent the prediction probabilities of DECOOP approach and zero-shot CLIP baseline. We denote $\mathcal{M}_C^i(\boldsymbol{x})$ as the prediciton probabilities computed for $\boldsymbol{x}$.

### 3.3. Inference

During testing, we evaluate an ensemble of $K$ new-class detectors $\{\mathcal{M}_D^i\}_{i=1}^K$ to determine whether each testing data should be predicted by one of the learned sub-classifiers $\mathcal{M}_C^i$ or the zero-shot CLIP baseline. Specifically, for a testing instance $\boldsymbol{x}$, we first compute the scores of the new-

*Table 1.* Comparison of average performance across 11 datasets was conducted among three approaches: ZS, PT, and our DEPT framework, utilizing ViT-B/16 and ViT-B/32 architectures. These results are consistent with our theoretical analysis.

| METHOD | VIT-B/16 | | VIT-B/32 | |
|---|---|---|---|---|
| | NEW ACC. | ACCURACY | NEW ACC. | ACCURACY |
| ZS | 65.49 | 63.92 | 63.95 | 60.36 |
| PT | 57.73 | 65.57 | 53.01 | 61.03 |
| DEPT | **68.15** | **68.03** | **65.45** | **62.92** |

*Table 2.* The average performance across 11 datasets using ViT-B/16 and ViT-B/32 architectures. The best performance is in bold.

| METHOD | VIT-B/16 | | VIT-B/32 | |
|---|---|---|---|---|
| | H | ACCURACY | H | ACCURACY |
| CLIP | 70.84 | 63.92 | 67.13 | 60.36 |
| PROMPT ENS. | 71.65 | 65.39 | 67.76 | 60.73 |
| COOP | 72.14 | 65.57 | 67.86 | 61.03 |
| COCOOP | 74.72 | 67.67 | 70.77 | 62.96 |
| SHIP | 72.26 | 64.51 | 69.25 | 59.91 |
| DECOOP(OURS) | **76.13** | **69.69** | **72.51** | **65.75** |

class detectors, $\{\mathcal{M}_D^i(\boldsymbol{x})\}_{i=1}^K$, and then make the prediction according to our DECOOP approach, defined as:

$$P_{\text{DECOOP}}(\boldsymbol{x}) = \begin{cases} P_{\text{ZS}}(\boldsymbol{x}), & \text{if } \max\limits_{i\in\{1,\cdots,K\}} \mathcal{M}_D^i(\boldsymbol{x}) < \tau, \\ \mathcal{M}_C^{i^\star}(\boldsymbol{x}), & \text{if } \max\limits_{i\in\{1,\cdots,K\}} \mathcal{M}_D^i(\boldsymbol{x}) \geq \tau, \end{cases} \tag{8}$$

where $i^\star = \arg\max_{i\in\{1,\cdots,K\}} \mathcal{M}_D^i(\boldsymbol{x})$. DECOOP approach selects single sub-classifier to predict each testing data instead of aggregating the results from all sub-classifiers. As a result, our approach requires $K$ times computation for the new-class detectors compared to the zero-shot CLIP baseline. In our experiments, we set $K$ to 3, which does not impose a heavy computational burden. We conduct experiments about evaluation time in Appendix B.7, demonstating that DECOOP is relatively efficient.

## 4. Experiments

In this section, we conduct experiments to answer the following three research questions:

**RQ1**: Can the empirical results of the DEPT framework on real-world datasets conform to our theoretical analysis?

**RQ2**: Can the DECOOP method surpass existing baseline and SOTA methods, thereby demonstrating its robustness?

**RQ3**: Does the DECOOP successfully improve the base-to-new discriminability, as designed?

### 4.1. Experimental Setup

**Evaluation Protocol.** We adopt the few-shot prompt tuning setting as previously explored in studies such as (Radford et al., 2021; Zhou et al., 2022a; Wang et al., 2023b). This setting involves partitioning the class space of each dataset equally, with 50% of the classes designated as base classes and the remaining 50% as new classes. Consequently, for each dataset, prompts are learned for downstream tasks using 16 labeled samples per base class, drawn from the training set. The efficacy of these learned prompts is subsequently evaluated on the entire testing set, encompassing both base and new classes. In DECOOP method, we report the Accuracy as well as previously reported H metric. As per the definition in CoCoOp (Zhou et al., 2022a), H metric separately evaluates the accuracy on base classes and new classes, denoted as $\text{Acc}_{\text{base}}$ and $\text{Acc}_{\text{new}}$. Then, H metric is computed using their harmonic mean, defined as $\text{H} = \frac{2\times\text{Acc}_{\text{base}}\times\text{Acc}_{\text{new}}}{\text{Acc}_{\text{base}}+\text{Acc}_{\text{new}}}$. The metric H evaluates the overall performance of classifying both base and new classes separately, which we refer to as base-class discriminability and new-class discriminability. We evaluate the accuracy of the entire class space, which includes a mix of base and new classes, denoted as Accuracy. This metric evaluates the overall performance of classifying both base and new classes, while additionally measuring base-to-new discriminability compared to the H metric.

**Datasets.** Following the CoOp framework (Zhou et al., 2022b), we conducted evaluations of our proposed DECOOP framework along with comparison methods on various image classification tasks. These tasks included general object recognition using ImageNet (Deng et al., 2009) and Caltech-101 (Fei-Fei et al., 2007) datasets, fine-grained object recognition involving datasets such as Oxford Pets (Krause et al., 2013), Food-101 (Bossard et al., 2014), Stanford Cars (Krause et al., 2013), Oxford Flowers 102 (Nilsback & Zisserman, 2008), and FGVC Aircraft (Maji et al., 2013). Additionally, we performed a remote sensing recognition task using the EuroSAT (Helber et al., 2019) dataset, a texture recognition task using the DTD (Cimpoi et al., 2014) dataset, an action recognition task using UCF101 (Soomro et al., 2012) dataset and a large-scale scene understanding task using SUN397 (Xiao et al., 2010) dataset. For each dataset, we developed a few-shot training set for prompt tuning and employed the full testing set to evaluate the effectiveness of the learned prompts.

**Compared Methods.** We compare our approach with five existing prompt-based methods. CLIP (Radford et al., 2021) uses a hand-crafted prompt to generate the target classifier on the downstream task. Furthermore, we compare the

*Table 3.* Performance comparison on 11 datasets using ViT-B/16 architecture. The best performance is in bold.

| | Average | | ImageNet | | Caltech101 | | OxfordPets | |
|---|---|---|---|---|---|---|---|---|
| | H | Acc. | H | Acc. | H | Acc. | H | Acc. |
| CLIP | 70.84 | 63.92 | 70.20 ± 0.00 | 66.73 ± 0.00 | 95.41 ± 0.00 | 92.90 ± 0.00 | 92.93 ± 0.00 | 88.03 ± 0.00 |
| PROMPT ENS. | 71.65 | 65.39 | 72.00 ± 0.00 | 68.48 ± 0.00 | 96.20 ± 0.00 | 94.08 ± 0.00 | 92.42 ± 0.00 | 86.37 ± 0.00 |
| COOP | 72.14 | 65.57 | 64.95 ± 1.11 | 61.79 ± 1.09 | 95.96 ± 0.39 | 93.24 ± 0.68 | 95.38 ± 0.33 | 89.61 ± 0.34 |
| COCOOP | 74.72 | 67.67 | 72.71 ± 0.33 | 69.41 ± 0.36 | 95.55 ± 0.24 | 93.43 ± 0.37 | **95.71 ± 0.76** | **90.24 ± 1.32** |
| SHIP | 72.26 | 64.51 | 67.29 ± 0.38 | 63.65 ± 0.32 | 95.83 ± 0.23 | 92.93 ± 0.37 | 94.44 ± 0.54 | 86.78 ± 1.32 |
| DECOOP(OURS) | **76.13** | **69.69** | **72.98 ± 0.04** | **69.62 ± 0.08** | **96.52 ± 0.09** | **94.50 ± 0.22** | 95.27 ± 0.08 | 88.87 ± 0.28 |

| | StandfordCars | | Flowers102 | | Food101 | | FGVCAircraft | |
|---|---|---|---|---|---|---|---|---|
| | H | Acc. | H | Acc. | H | Acc. | H | Acc. |
| CLIP | 68.75 ± 0.00 | 65.39 ± 0.00 | 72.74 ± 0.00 | 67.28 ± 0.00 | 90.18 ± 0.00 | 85.40 ± 0.00 | 30.25 ± 0.00 | 23.94 ± 0.00 |
| PROMPT ENS. | 69.36 ± 0.00 | 65.95 ± 0.00 | 72.14 ± 0.00 | 67.03 ± 0.00 | 90.32 ± 0.00 | 85.54 ± 0.00 | 29.42 ± 0.00 | 23.31 ± 0.00 |
| COOP | 68.22 ± 0.49 | 63.81 ± 0.44 | 78.33 ± 2.26 | 72.11 ± 2.36 | 86.65 ± 1.38 | 80.84 ± 1.50 | 29.38 ± 1.78 | 24.80 ± 1.23 |
| COCOOP | 71.49 ± 0.62 | 67.75 ± 0.68 | 80.04 ± 1.46 | 71.95 ± 1.24 | 90.41 ± 0.24 | 85.61 ± 0.43 | 27.87 ± 11.36 | 21.46 ± 7.42 |
| SHIP | 69.71 ± 0.43 | 64.67 ± 0.55 | 76.85 ± 2.18 | 70.40 ± 2.01 | 86.84 ± 1.49 | 77.39 ± 2.19 | 27.13 ± 1.10 | 24.44 ± 0.96 |
| DECOOP(OURS) | **73.24 ± 0.15** | **69.64 ± 0.19** | **84.16 ± 0.27** | **78.61 ± 0.59** | **90.68 ± 0.09** | **85.83 ± 0.07** | **31.44 ± 0.39** | **25.15 ± 0.31** |

| | SUN397 | | DTD | | EuroSAT | | UCF101 | |
|---|---|---|---|---|---|---|---|---|
| | H | Acc. | H | Acc. | H | Acc. | H | Acc. |
| CLIP | 72.26 ± 0.00 | 62.57 ± 0.00 | 57.32 ± 0.00 | 44.56 ± 0.00 | 58.16 ± 0.00 | 41.40 ± 0.00 | 71.00 ± 0.00 | 64.97 ± 0.00 |
| PROMPT ENS. | 75.04 ± 0.00 | 65.97 ± 0.00 | 59.63 ± 0.00 | 46.28 ± 0.00 | 58.45 ± 0.00 | 48.91 ± 0.00 | 73.17 ± 0.00 | 67.33 ± 0.00 |
| COOP | 71.37 ± 1.21 | 61.82 ± 1.11 | 57.22 ± 2.37 | 48.18 ± 1.78 | 74.33 ± 4.35 | 59.65 ± 5.07 | 71.68 ± 2.84 | 65.41 ± 2.18 |
| COCOOP | 77.17 ± 0.27 | 68.17 ± 0.33 | 60.59 ± 1.51 | 47.90 ± 1.43 | 73.77 ± 3.58 | 58.08 ± 1.49 | 76.59 ± 0.79 | 70.39 ± 1.25 |
| SHIP | 72.57 ± 0.38 | 60.42 ± 0.48 | 56.82 ± 2.18 | 47.58 ± 1.62 | 73.29 ± 2.67 | 54.11 ± 1.73 | 74.09 ± 2.09 | 67.24 ± 1.94 |
| DECOOP(OURS) | **78.11 ± 0.09** | **69.33 ± 0.05** | **62.72 ± 1.23** | **51.44 ± 1.04** | **74.61 ± 3.82** | **61.90 ± 3.72** | **77.67 ± 0.50** | **71.71 ± 0.79** |

PROMPT ENS. method, an ensemble technique that utilizes multiple classifiers to enhance the performance of CLIP, adhering to the guidelines set by CLIP. COOP (Zhou et al., 2022b) learns a soft prompt by minimizing the classification loss, and COCOOP (Zhou et al., 2022b) extends COOP by further learning a lightweight neural network to generate for each image an input-conditional token. SHIP (Wang et al., 2023b) follows variational autoencoders to introduce a generator that reconstructs the visual features by inputting the synthesized prompts and the corresponding class names to the textual encoder of CLIP.

**Implementation Details.** The number of tokens in each prompt is set to 16 for DECOOP approach and comparison methods. We train the prompts of new-class detectors for 50 epochs using the SGD optimizer and subsequently train the prompts for sub-classifiers for 100 epochs, also using the SGD optimizer. The learning rate $lr$ is set to 0.002, and it follows a cosine decay schedule. The margin $\gamma$ is set to $0.4$ for all datasets. We use the PROMPT ENS. method as our zero-shot baseline within the DECOOP approach. The batch size for images is 32 across all datasets. All experiments were conducted on Linux servers equipped with NVIDIA A800 GPUs. We report the average results over 5 runs with different random seed $\{1, 2, 3, 4, 5\}$.

### 4.2. Empirical Results

**RQ1**: Can the empirical results of the DEPT framework on real-world datasets conform to our theoretical analysis?

To verify Theorem 2.1, we conducted experiments on 11 datasets using ViT-B/16 and ViT-B/32 architectures. We employed CLIP as the zero-shot baseline ZS and COOP as the prompt tuning method PT. Subsequently, we constructed our DEPT framework by integrating these two methods, as presented in Equation 3. Here, the OOD detector used in our DEPT framework directly derives from CLIP using MSP method (Hendrycks & Gimpel, 2016). Each method is evaluated on the entire class space $\mathcal{Y}$, and the average performance across all datasets is reported. The results include New Acc. and Accuracy, indicating the average performance of new classes and all classes, respectively. The results presented in Table 1 consistently demonstrate that our DEPT framework outperforms both ZS and PT methods when evaluated using the New Acc. and Accuracy metrics. This observation suggests that the DEPT framework effectively mitigates performance degradation on new classes through the utilization of the OOD detector, which aligns well with our theoretical analysis.

**RQ2**: Can the DECOOP method surpass existing baseline and SOTA methods, thereby demonstrating its robustness?

To assess the effectiveness of the DECOOP approach, we conducted experiments on 11 datasets using ViT-B/16 and ViT-B/32 architectures. The average performance across all datasets, as well as the detailed performance on each dataset measured by two metrics, i.e., H and Accuracy, is reported. The results obtained using the ViT-B/16 architecture are presented in Table 3. Our DECOOP approach demonstrates superior average performance on both the H metric and

*Table 4.* The base-to-new discriminability of each method evaluated using MSP method (Hendrycks & Gimpel, 2016) and AUROC metrics. The best performance is in bold.

| DATASET | CLIP | COCOOP | SHIP | DECOOP(OURS) |
|---|---|---|---|---|
| IMAGENET | 88.34 | 88.05 | 84.71 | **97.48** |
| CALTECH101 | 97.03 | 95.71 | 96.94 | **99.58** |
| OXFORDPETS | 92.66 | 91.15 | 93.30 | **98.12** |
| STANFORDCARS | 86.24 | 83.00 | 87.23 | **97.63** |
| FLOWERS102 | 84.92 | 79.63 | 84.84 | **95.75** |
| FOOD101 | 89.88 | 88.19 | 89.92 | **97.59** |
| FGVCAIRCRAFT | 75.08 | 69.00 | 75.78 | **84.06** |
| SUN397 | 72.46 | 73.75 | 74.78 | **90.21** |
| DTD | 62.29 | 60.65 | 60.66 | **75.47** |
| EUROSAT | 56.40 | 57.74 | 59.32 | **77.78** |
| UCF101 | 82.03 | 79.03 | 80.35 | **93.56** |
| AVERAGE | 80.67 | 78.72 | 80.71 | **91.57** |

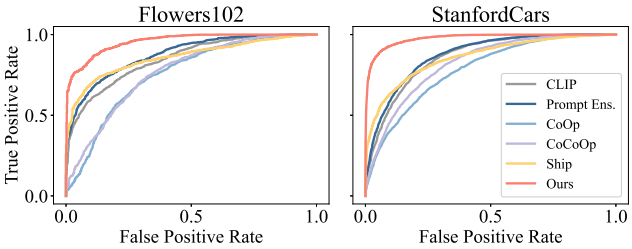



*Figure 6.* The ROC curve for detecting new classes of each method on Flowers102 and StandfordCars datasets.

Accuracy, showcasing its robustness. Regarding the detailed performance on each dataset, our approach outperforms the comparison methods on 10 out of 11 datasets, while achieving comparable performance on the remaining dataset. The detailed results using the ViT-B/32 architecture are provided in Appendix B.1, which yield similar conclusions.

Furthermore, these results reveal a positive correlation between the H metric and Accuracy in most cases. However, specific datasets such as FGVCAircraft (Maji et al., 2013) show that higher H metric values do not necessarily lead to improved Accuracy. This observation suggests that the H metric is inadequate for measuring base-to-new discriminability, emphasizing the significance of OPT problem.

**RQ3**: Does the DECOOP successfully improve the base-to-new discriminability, as designed?

The DECOOP approach introduces new-class detectors with the aim of improving base-to-new discriminability while simultaneously enhancing the discriminability of new classes. We evaluate the base-to-new discriminability of our approach and selected methods using the MSP (Hendrycks & Gimpel, 2016) method with the ViT-B/16 architecture. Specifically, for each method, we use the maximum probability on base classes as the scores and report the AUROC (Bradley, 1997) in Table 4. The results clearly indicate

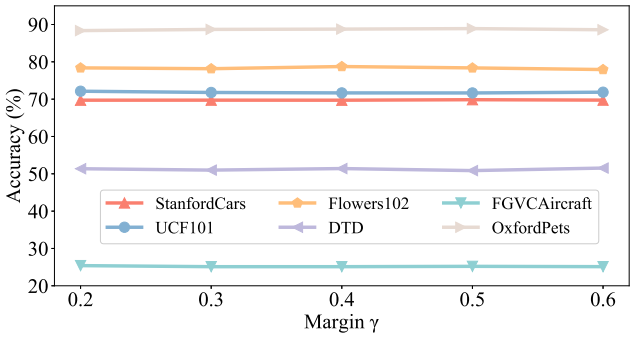



*Figure 7.* Performance using different values of margin $\gamma$.

that our DECOOP approach significantly improves base-to-new discriminability, which accounts for its SOTA performance. We have omitted some methods and standard deviations due to space limitations. Please refer to Appendix B.2 for full results. Additionally, we present the ROC curves for two representative datasets in Figure 6, which demonstrates similar findings. Due to space limitations, the ROC curves for all datasets are provided in Appendix B.3. Furthermore, we explore the correlation between the performance of new-class detectors and the model in Appendix B.4.

**Hyperparemeter.** The margin $\gamma$ serves as a hyperparameter for learning new-class detectors in our DECOOP approach. It controls the margin in the optimization process of the detectors, which may affect their performance. To answer the robustness question of $\gamma$, we conduct experiments on six datasets. Figure 7 demonstrates the robustness of the DECOOP approach to changes in $\gamma$.

**Comparison with Ensembling of COOP.** In Appendix B.6, we conduct an experiment to determine if directly combining multiple COOP prompts can lead to performance improvement. The results demonstrate that combining 2, 4, or 6 COOP prompts does not effectively enhance performance and, at times, even deteriorates the performance. This indicates that our performance gains cannot be attributed to simple prompt ensembling.

**Running Time.** In Appendix B.7, we conduct an experiment to compare the COOP, COCOOP, and DECOOP methods as shown in Table 9. On the EuroSAT dataset, the runtime of DECOOP increased only slightly compared to COOP (14.1s vs. 34.1s), but it is significantly more efficient than COCOOP (62.0s), demonstrating the efficiency of the DECOOP algorithm.

## 5. Related Work

**Few-shot Prompt Tuning.** Prompt learning aims to formalize various NLP tasks to mask language modeling problems, which is similar to the pre-training of language mod-

els (Devlin et al., 2018; Radford et al., 2019; 2021) by adopting different prompt templates. The previous works (Petroni et al., 2019; Radford et al., 2019; Brown et al., 2020) elaborately design human-crafted prompts, which is known as prompt engineering. Despite considerable progress in NLP, prompt learning remains underexplored in computer vision. Pretrained VLMs (Jia et al., 2021; Radford et al., 2021) introduce hand-crafted prompts to perform zero-shot inference on the downstream tasks. However, designing specific prompts for various downstream tasks is inefficient and costly and several studies (Shi et al., 2024b;c) peforms parameter-efficient fine-tuning to address this problem. CoOp (Zhou et al., 2022b) makes prompt learnable via minimizing the classification loss on the target task, adopting the prompt tuning approach of NLP. However, CoOp decreases the zero-shot capability of VLMs. To fix the problem, CoCoOp (Zhou et al., 2022a) introduces meta net to conditionally fine-tune prompts. LFA (Ouali et al., 2023) adopts a simple linear approach for vision-language alignment. VPT (Derakhshani et al., 2022) attempts to learn a collection of continuous prompts to capture the variational visual representation. SHIP (Wang et al., 2023b) follows the paradigm of variational autoencoders to generate visual features according to the prompts via the generative method. ProDA (Lu et al., 2022) proposes to learn the distribution of instance-specific prompts via variational inference. Ding et al. (2024) explores the integration of OOD detection methods for VLMs and present meaningful observations. However, these studies do not consider the OPT evaluation setting. Recent studies (Zhang et al., 2024; Shu et al., 2022a) also make the attempts to perform prompt tuning on changing datastreams in a test-time adaptation manner (Zhou et al., 2023; 2024; Zhao et al., 2024). These studies can be explored to address OPT problem in the furture studies.

**Out-of-distribution Detection.** Out-of-distribution detection refers to training the model on in-distribution (ID) dataset to classify OOD and ID samples. MSP (Hendrycks & Gimpel, 2016) takes the maximum softmax probability over ID categories as the score. RotPred (Hendrycks et al., 2019) includes an extra head to predict the rotation angle of rotated inputs in a self-supervised manner, and the rotation head together with the classification head is used for OOD detection. MCD (Yu & Aizawa, 2019) considers an ensemble of multiple classification heads and promotes the disagreement between each head's prediction on OOD samples. StyleAugment (Geirhos et al., 2018) applies style transfer to clean images to emphasize the shape bias over the texture bias. STEP (Zhou et al., 2021) focuses on exploring out-of-distribution detection within a semi-supervised setting (Tong et al., 2022; Lan-Zhe & Yu-Feng, 2024; Shi et al., 2024a; Jia et al., 2024). CIDER (Ming et al., 2022) regularizes the model's hyperspherical space by increasing inter-class separability and intra-class compactness. MixOE (Zhang et al., 2023) performs pixel-level mixing operations between ID and OOD samples and regularizes the model such that the prediction confidence smoothly decays as the input transitions from ID to OOD. RegMixup (Pinto et al., 2022) trains the model with both clean images and mixed images obtained from the convex combination. Recent studies (Ming & Li, 2024; Sun et al., 2024), such as Clipn (Wang et al., 2023a), LoCoOp (Miyai et al., 2023), attempt to explore the capability of zero-shot and few-shot ood detection via VLMs respectively. However, while these studies primarily focus on OOD detection tasks, our research utilizes OOD detection to enhance the generalization of VLMs.

## 6. Conclusion

In this paper, we explore the OPT problem in detail and uncover that the base-to-new discriminability is crucial but often overlooked by existing methods and settings. We first introduce the DEPT framework and demonstrate, through theoretical analysis, that incorporating OOD detection into prompt tuning can enhance the base-to-new discriminability and prevent degradation of new-class discriminability. Building upon DEPT, we propose a novel prompt tuning approach called DECOOP that introduces new-class detectors and sub-classifiers to further enhance the discriminability of both the base and new classes. Experimental results validate our analysis of DEPT and demonstrate the effectiveness of our DECOOP approach.

One limitation of our work is that we only take the initial step in combining OOD detection and prompt tuning. We believe there is potential for future researchers to explore. The other limitations are included in Appendix C.

## Code Availability Statement

The implementation code for this work is available at `https://github.com/WNJXYK/DeCoOp`.

## Acknowledgements

This research was supported by National Science and Technology Major Project (2022ZD0114803) and the National Science Foundation of China (62306133, 62176118).

## Impact Statement

This paper aims to advance prompt tuning for vision-language models. The work carried out in this study has various potential societal implications. We firmly believe that the majority of these impacts are positive and it is unnecessary to explicitly emphasize any specific ones in this paper. Additionally, we anticipate that the responsible utilization of technology will foster discourse concerning the best practices and regulations for implementing methods.

## References


Bossard, L., Guillaumin, M., and Gool, L. V. Food-101 - mining discriminative components with random forests. In *Proceedings of the 13th European Conference on Computer Vision*, pp. 446–461, 2014.

Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, pp. 1145–1159, 1997.

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

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceeings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Derakhshani, M. M., Sanchez, E., Bulat, A., da Costa, V. G. T., Snoek, C. G., Tzimiropoulos, G., and Martinez, B. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Ding, K., Zhang, H., Yu, Q., Wang, Y., Xiang, S., and Pan, C. Weak distribution detectors lead to stronger generalizability of vision-language prompt tuning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 1528–1536, 2024.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understandin*, pp. 59–70, 2007.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Hayes, T. L., Krishnan, G. P., Bazhenov, M., Siegelmann, H. T., Sejnowski, T. J., and Kanan, C. Replay in deep learning: Current approaches and missing biological elements. *Neural Computation*, pp. 2908–2950, 2021.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 2217–2226, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in neural information processing systems*, volume 32, 2019.

Ilharco, G., Wortsman, M., Gadre, S. Y., Song, S., Hajishirzi, H., Kornblith, S., Farhadi, A., and Schmidt, L. Patching open-vocabulary models by interpolating weights. In *Advances in Neural Information Processing Systems*, 2022.

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

Jia, L.-H., Guo, L.-Z., Zhou, Z., and Li, Y.-F. Lamda-ssl: a comprehensive semi-supervised learning toolkit. *Science China Information Sciences*, 67(117101), 2024.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J. C., Brna, A. P., Raja, S. C., Cheney, N., Clune, J., Daram, A. R., Fusi, S., Helfer, P., Kay, L., Ketz, N., Kira, Z., Kolouri, S., Krichmar, J. L., Kriegman, S., Levin, M., Madireddy, S., Manicka, S., Marjaninejad, A., McNaughton, B., Miikkulainen, R., Navratilova, Z., Pandit, T., Parker, A., Pilly, P. K., Risi, S., Sejnowski, T. J., Soltoggio, A., Soures, N., Tolias, A. S., Urbina-Meléndez, D., Cuevas, F. J. V., van de Ven, G. M., Vogelstein, J. T., Wang, F., Weiss, R., Yanguas-Gil, A., Zou, X., and Siegelmann, H. T. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, pp. 196–210, 2022.

Lan-Zhe, G. and Yu-Feng, L. Robust pseudo-label selection for holistic semi-supervised learning. *Science China Information Sciences*, 53(3):623–637, 2024.

Lange, M. D., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. G., and Tuytelaars, T. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 3366–3385, 2022.

Liu, D. and Yu, J. Otsu method and k-means. In *In Proceedings of the 9th International Conference on Hybrid Intelligent Systems*, pp. 344–349, 2009.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, pp. 1–35, 2023.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., and Sanner, S. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Ming, Y. and Li, Y. How does fine-tuning impact out-of-distribution detection for vision-language models? *International Journal of Computer Vision*, 132(2):596–609, 2024.

Ming, Y., Sun, Y., Dia, O., and Li, Y. How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450*, 2022.

Miyai, A., Yu, Q., Irie, G., and Aizawa, K. Locoop: Few-shot out-of-distribution detection via prompt learning. *arXiv preprint arXiv:2306.01293*, 2023.

Nilsback, M. and Zisserman, A. Automated flower classification over a large number of classes. In *Proceedings of the 6th Indian Conference on Computer Vision*, pp. 722–729, 2008.

Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

Ouali, Y., Bulat, A., Martínez, B., and Tzimiropoulos, G. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15488–15500, 2023.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Pinto, F., Yang, H., Lim, S. N., Torr, P., and Dokania, P. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *Advances in Neural Information Processing Systems*, pp. 14608–14622, 2022.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, pp. 9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Shi, J.-X., Wei, T., and Li, Y.-F. Residual diverse ensemble for long-tailed multi-label text classification. *Science CHINA Information Science*, 2024a.

Shi, J.-X., Wei, T., Zhou, Z., Shao, J.-J., Han, X.-Y., and Li, Y.-F. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Shi, J.-X., Zhang, C., Wei, T., and Li, Y.-F. Efficient and long-tailed generalization for pre-trained vision-language model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024c.

Shu, M., Nie, W., Huang, D., Yu, Z., Goldstein, T., Anandkumar, A., and Xiao, C. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, 2022a.

Shu, M., Nie, W., Huang, D.-A., Yu, Z., Goldstein, T., Anandkumar, A., and Xiao, C. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, pp. 14274–14289, 2022b.

Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

Sun, H., He, R., Han, Z., Lin, Z., Gong, Y., and Yin, Y. Clip-driven outliers synthesis for few-shot OOD detection. *CoRR*, abs/2404.00323, 2024.

Tong, W., Hai, W., Weiwei, T., and Yufeng, L. Robust model selection for positive and unlabeled learning with constraints. *Science China Information Sciences*, 65(212101), 2022.

Wang, H., Li, Y., Yao, H., and Li, X. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1802–1812, 2023a.

Wang, Z., Liang, J., He, R., Xu, N., Wang, Z., and Tan, T. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3032–3042, 2023b.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7949–7961, 2022.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Yang, X., Shao, J., Tu, W., Li, Y., Dai, W., and Zhou, Z. Safe abductive learning in the presence of inaccurate rules. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 16361–16369, 2024a.

Yang, X., Wei, W., Shao, J., Li, Y., and Zhou, Z. Analysis for abductive learning and neural-symbolic reasoning shortcuts. In *Proceedings of the 41st International Conference on Machine Learning*, 2024b.

Yu, Q. and Aizawa, K. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9518–9526, 2019.

Zhang, D., Zhou, Z., and Li, Y. Robust test-time adaptation for zero-shot prompt tuning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 16714–16722, 2024.

Zhang, J., Inkawhich, N., Linderman, R., Chen, Y., and Li, H. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5531–5540, 2023.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *Journal of Machine Learning Research*, 25(98):1 – 52, 2024.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16795–16804, 2022a.

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

Zhou, Z., Guo, L., Cheng, Z., Li, Y., and Pu, S. STEP: out-of-distribution detection in the presence of limited in-distribution labeled data. In *Advances in Neural Information Processing Systems*, pp. 29168–29180, 2021.

Zhou, Z., Guo, L., Jia, L., Zhang, D., and Li, Y. ODS: test-time adaptation in the presence of open-world data shift. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 42574–42588, 2023.

Zhou, Z., Zhang, D.-C., Li, Y.-F., and Zhang, M.-L. Towards robust test-time adaptation method for open-set recognition. *Journal of Software*, 35(4):1667–1681, 2024.

## A. Proof of Theorem 2.1

*Proof.* We first compute $H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x})$ and $H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x})$ for one specific instance $\boldsymbol{x}$. Recall that for an instance $\boldsymbol{x}$, we denote its ground-truth label space as $k$ (which always equals to $g(\boldsymbol{x})$) and its ground-truth label as $f(\boldsymbol{x})$. To facilitate the proof, we define additional label spaces:

$$\mathcal{Y}_{i,j} = \begin{cases} \{j\}, & j \in \mathcal{Y}_i, \\ \emptyset, & otherwise, \end{cases} \tag{9}$$

and additional class vectors for $\boldsymbol{x}$:

$$\tilde{y}_{i,j} = \begin{cases} 1, & f(\boldsymbol{x}) = j \wedge f(\boldsymbol{x}) \in \mathcal{Y}_i, \\ 0, & otherwise. \end{cases} \tag{10}$$

Our computational results are presented as follows:

$$\begin{aligned} H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{ZS}}(y=j|y \in \mathcal{Y}_k, \boldsymbol{x})\}_{j=1}^{C}\right) \\ &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{ZS}}(y \in \mathcal{Y}_{k,j}|y \in \mathcal{Y}_k, \boldsymbol{x})\}_{j=1}^{C}\right) \\ &= -\sum_{j=1}^{C} \tilde{y}_j \log P_{\text{ZS}}(y=j|y \in \mathcal{Y}_k, \boldsymbol{x}) \\ &= -\log P_{\text{ZS}}(y=f(\boldsymbol{x})|y \in \mathcal{Y}_k, \boldsymbol{x}), \end{aligned} \tag{11}$$

and

$$\begin{aligned} H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{k}}, \{P_{\text{ZS}}(y \in \mathcal{Y}_i|\boldsymbol{x})\}_{i=\{\text{b,n}\}}\right) \\ &= -\sum_{i \in \{\text{b,n}\}} \tilde{k}_i \log P_{\text{ZS}}(y \in \mathcal{Y}_i|\boldsymbol{x}) \\ &= -\log P_{\text{ZS}}(y \in \mathcal{Y}_k|\boldsymbol{x}). \end{aligned} \tag{12}$$

Then, we can bound $\mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}(\boldsymbol{x})\right]$ as follows:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}(\boldsymbol{x})\right] &= \mathbb{E}_{\boldsymbol{x}}\left[H\left(\tilde{\boldsymbol{y}}, \{P_{\text{ZS}}(y=j|\boldsymbol{x})\}_{j=1}^{C}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[H\left(\tilde{\boldsymbol{y}}, \{P_{\text{ZS}}(y \in \mathcal{Y}_{k,j}|\boldsymbol{x})\}_{j=1}^{C}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[-\log P_{\text{ZS}}(y \in \mathcal{Y}_{k,f(\boldsymbol{x})}|\boldsymbol{x})\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[-\log P_{\text{ZS}}(y \in \mathcal{Y}_{k,f(\boldsymbol{x})}|y \in \mathcal{Y}_k, \boldsymbol{x}) - \log P_{\text{ZS}}(y \in \mathcal{Y}_k|\boldsymbol{x})\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x}) + H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x})\right] \\ &= \mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}}\left[H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x})\right] \\ &\leq \delta + \epsilon. \end{aligned} \tag{13}$$

Further, we can similarily compute $H_{\text{PT}}^{\text{CLS}}(\boldsymbol{x})$ as follows:

$$\begin{aligned} H_{\text{PT}}^{\text{CLS}}(\boldsymbol{x}) &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{PT}}(y=j|y \in \mathcal{Y}_k, \boldsymbol{x})\}_{j=1}^{C}\right) \\ &= H\left(\tilde{\boldsymbol{y}}, \{P_{\text{PT}}(y \in \mathcal{Y}_{k,j}|y \in \mathcal{Y}_k, \boldsymbol{x})\}_{j=1}^{C}\right) \\ &= -\sum_{j=1}^{C} \tilde{y}_{k,j} \log P_{\text{PT}}(y \in \mathcal{Y}_{k,j}|y \in \mathcal{Y}_k, \boldsymbol{x}) \\ &= -\log P_{\text{PT}}(y \in \mathcal{Y}_{k,f(\boldsymbol{x})}|y \in \mathcal{Y}_k, \boldsymbol{x}). \end{aligned} \tag{14}$$

*Table 5.* Performance comparison between our proposed DECOOP method and comparison methods on 11 datasets using ViT-B/32 architecture. The best performance is in bold.

| | AVERAGE | | IMAGENET | | CALTECH101 | | OXFORDPETS | |
|---|---|---|---|---|---|---|---|---|
| | H | ACC. | H | ACC. | H | ACC. | H | ACC. |
| CLIP | 67.13 | 60.36 | 65.69 ± 0.00 | 62.05 ± 0.00 | 93.78 ± 0.00 | 91.08 ± 0.00 | 91.30 ± 0.00 | 85.01 ± 0.00 |
| PROMPT ENS. | 67.76 | 60.73 | 66.91 ± 0.00 | 63.22 ± 0.00 | 94.06 ± 0.00 | 91.20 ± 0.00 | 89.73 ± 0.00 | 83.18 ± 0.00 |
| COOP | 67.86 | 61.03 | 60.99 ± 0.09 | 57.61 ± 0.12 | 93.55 ± 0.76 | 91.09 ± 0.45 | 92.17 ± 0.77 | 85.21 ± 0.65 |
| COCOOP | 70.77 | 62.96 | 67.74 ± 1.23 | 64.06 ± 1.39 | 93.78 ± 0.92 | 91.01 ± 0.87 | **94.05 ± 0.56** | **87.84 ± 0.89** |
| SHIP | 69.25 | 59.91 | 61.72 ± 0.61 | 56.93 ± 1.26 | 93.35 ± 0.93 | 89.80 ± 0.83 | 92.19 ± 1.47 | 81.22 ± 1.03 |
| DECOOP(OURS) | **72.51** | **65.75** | **68.07 ± 0.06** | **64.49 ± 0.04** | **95.56 ± 0.22** | **93.36 ± 0.48** | 93.13 ± 0.50 | 86.25 ± 0.96 |

| | STANDFORDCARS | | FLOWERS102 | | FOOD101 | | FGVCAIRCRAFT | |
|---|---|---|---|---|---|---|---|---|
| | H | ACC. | H | ACC. | H | ACC. | H | ACC. |
| CLIP | 65.14 ± 0.00 | 60.39 ± 0.00 | 70.50 ± 0.00 | 64.27 ± 0.00 | 85.10 ± 0.00 | 79.16 ± 0.00 | 23.62 ± 0.00 | 18.30 ± 0.00 |
| PROMPT ENS. | 64.67 ± 0.00 | 59.82 ± 0.00 | 68.60 ± 0.00 | 63.30 ± 0.00 | 85.55 ± 0.00 | 79.59 ± 0.00 | 23.45 ± 0.00 | 18.30 ± 0.00 |
| COOP | 62.33 ± 1.21 | 56.95 ± 1.37 | 71.13 ± 1.95 | 65.25 ± 1.43 | 81.55 ± 0.91 | 74.32 ± 1.17 | 23.15 ± 1.71 | 18.88 ± 0.85 |
| COCOOP | 65.48 ± 0.66 | 60.27 ± 0.84 | 74.46 ± 1.10 | 65.86 ± 1.53 | **86.11 ± 0.29** | **80.09 ± 0.40** | 21.68 ± 5.89 | 15.28 ± 4.87 |
| SHIP | 64.38 ± 0.81 | 56.22 ± 1.00 | 70.41 ± 1.72 | 62.41 ± 1.88 | 81.76 ± 0.90 | 72.14 ± 1.43 | 19.34 ± 2.64 | 19.00 ± 0.98 |
| DECOOP(OURS) | **67.45 ± 0.15** | **62.55 ± 0.23** | **79.06 ± 0.43** | **72.84 ± 0.77** | 86.04 ± 0.10 | 79.98 ± 0.11 | **25.58 ± 0.33** | **20.03 ± 0.16** |

| | SUN397 | | DTD | | EUROSAT | | UCF101 | |
|---|---|---|---|---|---|---|---|---|
| | H | ACC. | H | ACC. | H | ACC. | H | ACC. |
| CLIP | 71.35 ± 0.00 | 61.99 ± 0.00 | 53.60 ± 0.00 | 42.85 ± 0.00 | 50.81 ± 0.00 | 38.17 ± 0.00 | 67.56 ± 0.00 | 60.67 ± 0.00 |
| PROMPT ENS. | 73.27 ± 0.00 | 63.74 ± 0.00 | 53.81 ± 0.00 | 43.44 ± 0.00 | 56.90 ± 0.00 | 40.75 ± 0.00 | 68.39 ± 0.00 | 61.49 ± 0.00 |
| COOP | 69.48 ± 1.01 | 59.89 ± 0.85 | 57.52 ± 1.82 | 48.90 ± 1.23 | 67.46 ± 7.70 | 51.07 ± 8.05 | 67.11 ± 3.56 | 62.12 ± 2.48 |
| COCOOP | 75.51 ± 0.37 | 65.96 ± 0.45 | 59.57 ± 2.21 | 47.08 ± 1.30 | 66.98 ± 8.67 | 49.19 ± 5.78 | 73.17 ± 1.24 | 65.98 ± 1.06 |
| SHIP | 70.33 ± 0.63 | 58.86 ± 0.71 | 57.22 ± 3.14 | 45.91 ± 1.07 | **77.74 ± 3.74** | 50.23 ± 1.92 | 73.27 ± 1.21 | 66.31 ± 0.72 |
| DECOOP(OURS) | **75.87 ± 0.14** | **66.59 ± 0.19** | **60.61 ± 0.48** | **50.39 ± 0.40** | 72.35 ± 2.42 | **58.93 ± 2.62** | **73.87 ± 0.36** | **67.83 ± 0.81** |

Finally, we can bound $\mathbb{E}_{\boldsymbol{x}}\left[H_{\text{DEPT}}(\boldsymbol{x})\right]$ as follows:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x}}\left[H_{\text{DEPT}}(\boldsymbol{x})\right] &= \mathbb{E}_{\boldsymbol{x}}\left[H\left(\tilde{\boldsymbol{y}}, \{P_{\text{DEPT}}(y=j|\boldsymbol{x})\}_{j=1}^{C}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[H\left(\tilde{\boldsymbol{y}}, \{P_{\text{DEPT}}(y\in\mathcal{Y}_{k,i}|\boldsymbol{x})\}_{i=1}^{C}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[-\log P_{\text{DEPT}}(y\in\mathcal{Y}_{k,i}|\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\wedge k=\text{b}}\left[-\log P_{\text{DEPT}}(y\in\mathcal{Y}_{k,i}|\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\wedge k=\text{n}}\left[-\log P_{\text{DEPT}}(y\in\mathcal{Y}_{k,i}|\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\wedge k=\text{b}}\left[-\log P_{\text{PT}}(y\in\mathcal{Y}_{k,f(\boldsymbol{x})}|y\in\mathcal{Y}_k,\boldsymbol{x}) - \log P_{\text{ZS}}(y\in\mathcal{Y}_k|\boldsymbol{x})\right] \\
&\quad + \mathbb{E}_{\boldsymbol{x}\wedge k=\text{n}}\left[-\log P_{\text{ZS}}(y\in\mathcal{Y}_{k,f(\boldsymbol{x})}|y\in\mathcal{Y}_k,\boldsymbol{x}) - \log P_{\text{ZS}}(y\in\mathcal{Y}_k|\boldsymbol{x})\right] \\
&= \mathbb{E}_{\boldsymbol{x}\wedge k=\text{b}}\left[H_{\text{PT}}^{\text{CLS}}(\boldsymbol{x}) + H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}\wedge k=\text{n}}\left[H_{\text{ZS}}^{\text{CLS}}(\boldsymbol{x}) + H_{\text{ZS}}^{\text{OOD}}(\boldsymbol{x})\right] \\
&\leq \alpha\cdot(\delta-\Delta+\epsilon) + (1-\alpha)\cdot(\delta+\epsilon) \\
&\leq \delta+\epsilon-\alpha\cdot\Delta.
\end{aligned}
\tag{15}
$$

□

## B. Additional Experimental Results

### B.1. Detailed Results on ViT-B/32 Architecture

To address the consistent performance of our proposal, we conduct experiments and compare our proposed DECOOP method, baseline methods, and SOTA prompting tuning methods on 11 datasets using ViT-B/32 architectures. Each dataset is trained with random seeds from 1 to 5. In terms of detailed performance on each dataset, our proposed method outperforms the comparison methods on 9 out of 11 datasets, while achieving comparable performance on the remaining 2 datasets, showcasing its robustness to different pre-trained architectures.

### B.2. Detailed AUROC Results

The full experimental results of Table 4 are presented in Table 6. Our DECOOP approach achieves the best base-to-new discriminability among all comparison methods.

*Table 6.* AUROC performance is compared with CLIP, Prompt Ensemble, COOP, COCOOP, SHIP and our proposed DECOOP. The results demonstrate that our proposal enhances base-to-new discriminability.

| DATASET | CLIP | PROMPT ENS. | COOP | COCOOP | SHIP | DECOOP(OURS) |
|---|---|---|---|---|---|---|
| IMAGENET | 88.34 ± 0.00 | 89.79 ± 0.00 | 77.14 ± 1.62 | 88.05 ± 1.22 | 84.71 ± 1.62 | **97.48 ± 0.03** |
| CALTECH101 | 97.03 ± 0.00 | 97.09 ± 0.00 | 94.53 ± 0.87 | 95.71 ± 0.50 | 96.94 ± 0.79 | **99.58 ± 0.03** |
| OXFORDPETS | 92.66 ± 0.00 | 92.21 ± 0.00 | 91.06 ± 1.00 | 91.15 ± 0.95 | 93.30 ± 1.23 | **98.12 ± 0.24** |
| STANFORDCARS | 86.24 ± 0.00 | 87.46 ± 0.00 | 78.25 ± 2.00 | 83.00 ± 2.24 | 87.23 ± 1.16 | **97.63 ± 0.02** |
| FLOWERS102 | 84.92 ± 0.00 | 87.78 ± 0.00 | 78.06 ± 1.82 | 79.63 ± 2.20 | 84.84 ± 1.41 | **95.75 ± 0.18** |
| FOOD101 | 89.88 ± 0.00 | 90.26 ± 0.00 | 87.53 ± 1.20 | 88.19 ± 1.07 | 89.92 ± 1.00 | **97.59 ± 0.04** |
| FGVCAIRCRAFT | 75.08 ± 0.00 | 75.86 ± 0.00 | 75.25 ± 1.36 | 69.00 ± 7.91 | 75.78 ± 1.65 | **84.06 ± 0.26** |
| SUN397 | 72.46 ± 0.00 | 75.29 ± 0.00 | 70.29 ± 1.47 | 73.75 ± 1.11 | 74.78 ± 1.14 | **90.21 ± 0.10** |
| DTD | 62.29 ± 0.00 | 61.10 ± 0.00 | 56.78 ± 1.93 | 60.65 ± 0.94 | 60.66 ± 1.22 | **75.47 ± 1.02** |
| EUROSAT | 56.40 ± 0.00 | 57.74 ± 0.00 | 52.26 ± 8.68 | 57.74 ± 2.49 | 59.32 ± 6.31 | **77.78 ± 3.85** |
| UCF101 | 82.03 ± 0.00 | 83.56 ± 0.00 | 72.72 ± 2.21 | 79.03 ± 1.52 | 80.35 ± 1.99 | **93.56 ± 0.62** |
| AVERAGE | 80.67 | 81.65 | 75.81 | 78.72 | 80.71 | **91.57** |

*Table 7.* Ablation study. We report average performance across 11 datasets was conducted among baselines, COOP, DEPT, DECOOP approaches, utilizing ViT-B/16 and ViT-B/32 architectures. The best performance is in bold. The second-best performance is underlined.

| METHOD | VIT-B/16 | | VIT-B/32 | |
|---|---|---|---|---|
| | H | ACCURACY | H | ACCURACY |
| CLIP | 70.84 | 63.92 | 67.13 | 60.36 |
| PROMPT ENS. | 71.65 | 65.39 | 67.76 | 60.73 |
| COOP | 72.14 | 65.57 | 67.86 | 61.03 |
| DEPT | <u>74.82</u> | <u>68.03</u> | <u>69.96</u> | <u>62.92</u> |
| DECOOP | **76.13** | **69.69** | **72.51** | **65.75** |

### B.3. Detailed ROC Curves

To evaluate whether our proposal can improve the performance for detecting, we conduct experiments on 11 datasets using ViT-B/16 architecture. Each curve is drawn using our experiment results with random seeds to 1. For each method, we adopt the maximum softmax probability over new classes as the detecting score for drawing the curve. The results in Figure 8 show that our proposal can achieve the best detection performance.

### B.4. Correlation between $\mathcal{M}_O$ and Performance

The objective of the DECOOP approach is to enhance the base-to-new discriminability through the $\mathcal{M}_O$, leading to improved performance. Hence, a key question arises: does a better $\mathcal{M}_O$ result in enhanced performance? To investigate this, we employ different new-class detectors with varying AUROC values for training and evaluate the performance as shown in Figure 9. This figure illustrates the correlation between the AUROC of the new-class detector and the performance metric. The results indicate a positive correlation between these two variables, validating our claim and aligning with our research objective.

### B.5. Ablation Study

We conduct ablation studies to validate the effectiveness of each component of our proposed DECOOP approach in Table 7. In this paper, we first propose a novel prompt tuning framework DEPT to introduce OOD detection into prompt tuning. Then, two advanced modules are integrated into DEPT framework to form our DECOOP approach. As the our two modules cannot be separated to perform classification, we compare baseline methods, DEPT framework, and our proposed DECOOP appraoch. The results show that DEPT framework enhances the base-to-new discriminability and prevents performance degradation of new classes, thereby outperforming CLIP, PROMPT ENS., and COOP methods. Further, our proposed DECOOP approach achieves the best performance among all methods, demonstrating it additionally enhances the base-class and new-class discriminability.

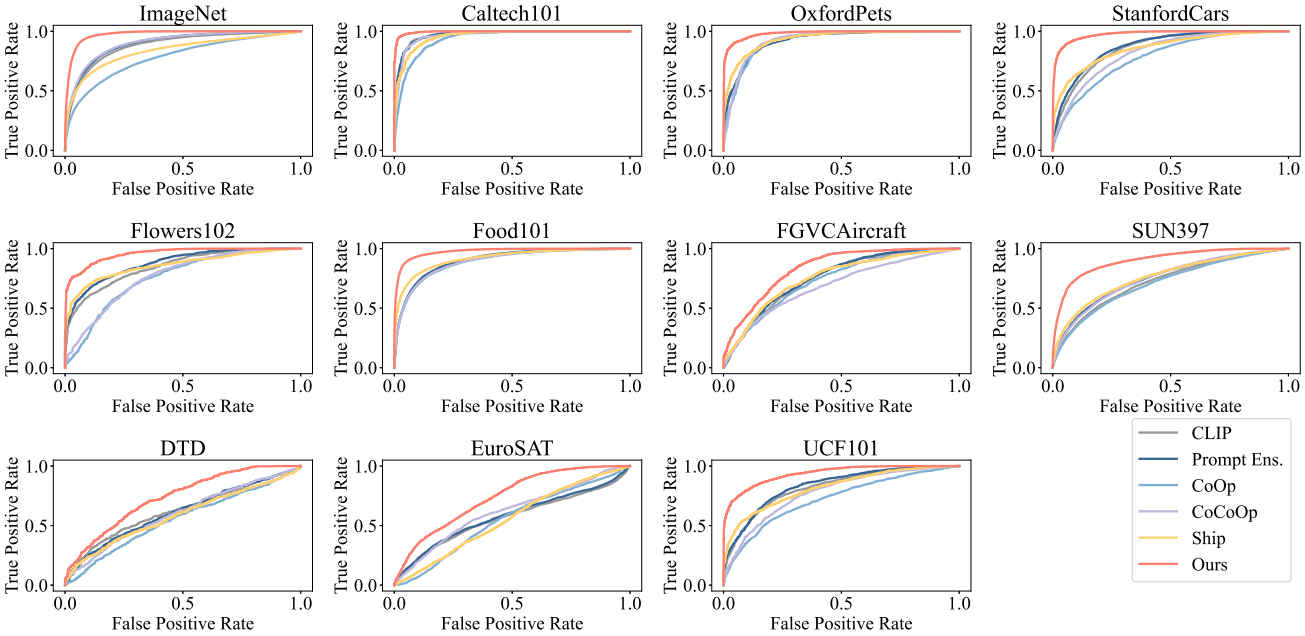



*Figure 8.* The roc curve for detecting new classes of each method on 11 datasets.

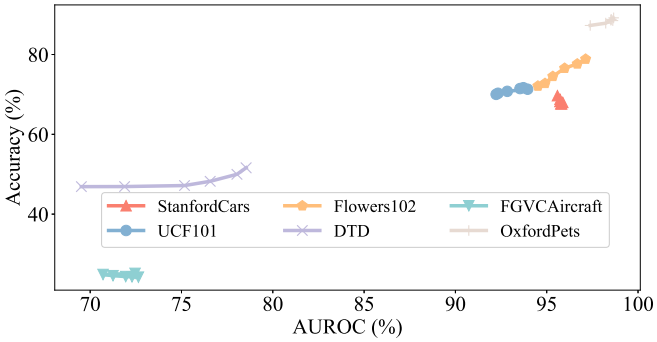



*Figure 9.* Correlation between performance of $\mathcal{M}_O$ and accuracy.

### B.6. Simple Ensembling of COOP Method

We also conduct experiments to evaluate whether directly ensemble multiple COOP learners can achieve similar performance. The results, shown in Table 8, indicate that the ensemble of multiple COOP prompts does not yield significantly better performance compared to the COOP method. These results prove that the performance gain does not derive from simple prompt ensembling.

### B.7. Evaluation Time

Our DECOOP approach adopts multiple prompts to detect OOD, so it may take more time. We compared the running time taken by COOP, COCOOP, and DECOOP methods when evaluating the testing set of two datasets in Table 9. The results show that the running time of the DECOOP is not significantly longer than the COOP method since the computation can be performed in parallel. However, our DECOOP approach runs in two stages (i.e., OOD detection and classification stages), therefore, the running time will be approximately double compared to the COOP method. However, the running time of COCOOP rises significantly as the number of categories increases, where demonstates our DECOOP is efficient.

*Table 8.* Performance comparison between our proposed method and the ensemble of multiple COOP prompts is conducted. The results demonstrate that directly combining multiple COOP learners does not yield significantly better performance compared to the COOP method. Moreover, our proposed algorithm outperforms other methods.

| METHOD | FLOWERS102 | DTD | CALTECH101 | STANDFORDCARS |
|---|---|---|---|---|
| COOP | 72.11 | 48.18 | 93.24 | 63.81 |
| COOP×2 | 71.62 | 50.08 | 93.31 | 64.20 |
| COOP×4 | 73.12 | 49.99 | 92.89 | 64.32 |
| COOP×6 | 71.89 | 49.69 | 92.87 | 65.03 |
| DECOOP | **78.61** | **51.44** | **94.50** | **69.64** |

*Table 9.* Evaluation running time of COOP, COCOOP, and DECOOP methods.

| DATASETS | #CLASSES | COOP | COCOOP | DECOOP |
|---|---|---|---|---|
| EUROSAT | 10 | 14.1S | 62.0S | 34.1S |
| FOOD101 | 101 | 50.5S | 711.7S | 131.5S |

*Table 10.* Comparison with weight interpolating methods.

| | NEW ACC. | ACCURACY |
|---|---|---|
| CLIP | 65.48 | 63.92 |
| COOP | 57.75 | 65.58 |
| RFT (WORTSMAN ET AL., 2022) | 65.34 | 69.26 |
| DECOOP | **66.54** | **69.69** |

### B.8. Comparison with Weight Interpolating Methods

Existing studies (Wortsman et al., 2022; Ilharco et al., 2022) observe that interpolating weights for tuned and original vision-language models can improve the generalization capacity. In the context of prompt tuning, we can interpolate weights of the tuned prompt and original prompt. We report the average results on all datasets using ViT-B/16 architecture in Table 10. The results show that interpolating weights can give better performance compared to both the original model and the tuned model, which aligns with the conclusion of existing studies. Our DECOOP outperforms other methods, demonstrating its effectiveness. Note that Weight Interpolating Methods and DECOOP have studied the different stages in fine-tuning, therefore, the combination of both to further enhance performance can be a direction for future research.

## C. Limitation and Future Work

Our paper proposes the integration of OOD detection into prompt tuning to prevent performance degradation on new classes. In addition to the content discussed at the end of Section 6, one limitation of our approach is the potentially increased time consumption due to the adoption of a two-stage classification process. Integrating knowledge (Yang et al., 2024b;a) into the prompt tuning to achieve the better generalization is also a future direction to explore. Experiments detailed in Appendix B.7 demonstrate that our method's running time is shorter than some existing methods (e.g., COCOOP), proving that the running time of our proposal is within an acceptable range. One possible solution is to integrate the two-stage classification into prompt training through the utilization of advanced training strategies, which can be explored as potential research directions in the future.