Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection

Geng Yu[1] **Jianing Zhu**[2] **Jiangchao Yao**[1,3*] **Bo Han**[2,4]

[1]CMIC, Shanghai Jiao Tong University [2]TMLR Group, Hong Kong Baptist University
[3]Shanghai A I Laboratory [4]RIKEN Center for Advanced Intelligence Project

{warriors30, Sunarker}@sjtu.edu.cn
{csjnzhu, bhanml}@comp.hkbu.edu.hk

Abstract

Out-of-distribution (OOD) detection is crucial for deploying reliable machine learning models in open-world applications. Recent advances in CLIP-based OOD detection have shown promising results via regularizing prompt tuning with OOD features extracted from ID data. However, the irrelevant context mined from ID data can be spurious due to the inaccurate foreground-background decomposition, thus limiting the OOD detection performance. In this work, we propose a novel framework, namely, *Self-Calibrated Tuning (SCT)*, to mitigate this problem for effective OOD detection with only the given few-shot ID data. Specifically, SCT introduces modulating factors respectively on the two components of the original learning objective. It adaptively directs the optimization process between the two tasks during training on data with different prediction uncertainty to calibrate the influence of OOD regularization, which is compatible with many prompt tuning based OOD detection methods. Extensive experiments and analyses have been conducted to characterize and demonstrate the effectiveness of the proposed SCT. The code is publicly available at: https://github.com/tmlr-group/SCT.

1 Introduction

The deep neural networks (DNNs) are demonstrated to be overconfident on the OOD data out of the pre-defined label space [Hendrycks and Gimpel, 2017], which can induce severe problems in those safety-critical applications like autonomous driving or medical intelligence. Various explorations [Liang et al., 2018, Djurisic et al., 2022, Du et al., 2022, Zhu et al., 2023b] thus have been conducted in designing scoring functions or fine-tuning methods with auxiliary outliers to improve the OOD distinguishability. Specially, with the emergence of the powerful pretrained vision-language models (VLMs) [Radford et al., 2021], a series of prompt tuning based methods [Miyai et al., 2024b, Tao et al., 2023, Bai et al., 2023, Ming et al., 2022c] show impressive performance in current OOD detection benchmarks, with the regularization given only few-shot in-distribution (ID) data.

Generally, these regularizations [Wang et al., 2023, Miyai et al., 2024b] are built upon the ID-irrelevant local context as the surrogate OOD source, which is extracted by VLMs (refer to Figure 1) based on its alignment with ID-class text features. Although this saves the costly collection of auxiliary outliers from the open world, the quality of the ID-irrelevant local context also becomes the bottleneck, which can be greatly affected by the foreground-background decomposition with VLMs. Specifically, as revealed in previous studies [Oh et al., 2023, Tu et al., 2024, Wang et al., 2024], the prevalent VLMs struggle with poor calibration, which means that the decomposition performance on downstream data might not be well guaranteed. Thus, it naturally motivates the following question:

*Correspondence to Jiangchao Yao (Sunarker@sjtu.edu.cn).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

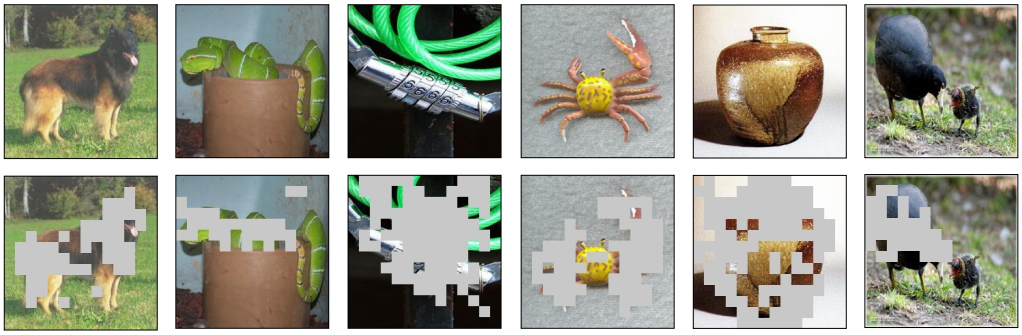


Figure 1: Imperfect foreground background decomposition. The top row shows the original images from ImageNet-1k and the bottom row shows the ID-irrelevant context extracted from the original images (shown as the colored patches of images on the second row), using CLIP fine-tuned with CoOp on 16-shot data. Due to the imperfect decomposition of fine-tuned vision-language models, large portions of the extracted local features from ID data belong to ID-related regions, thus harming the performance of OOD detection. More illustrations are presented in the Appendix A.3.3.

> *Can we flexibly leverage the imperfect OOD features extracted by the VLM itself, to facilitate the few-shot prompt tuning for effective OOD detection*?

As illustrated in Figure 1, a significant portion of extracted local context from ID data are not valid OOD features due to the inevitable imperfect decomposition. Consequently, OOD regularization based on such unreliable OOD features may potentially constrain the improvement of OOD detection. To investigate this problem, we conduct a proof-of-concept experiment on CLIP with ImageNet as the ID dataset and prompt-tune the model with different groups of data divided by their overall prediction uncertainty. Specifically, we find that the performance of prompt tuning based methods significantly deteriorates as the uncertainty of the given ID data rises, as presented in Figure 2, which motivates us to leverage such clues to overcome the current issue. Intuitively, as the model prediction on ID samples is less certain, the OOD features extracted from these data are less reliable. Performing OOD regularization on such unreliable surrogate OOD features can degrade the OOD detection performance of CLIP. Therefore, a potential idea is to adaptively adjust the importance of OOD features extracted from ID data according to their prediction uncertainty during model training.

Based on the previous observation, we propose a new learning framework, i.e., *Self-Calibrated Tuning* (SCT), to alleviate the problem induced by spurious OOD features. At the high level, we aim to dynamically adjust the weight of OOD regularization from different training samples based on their prediction uncertainty to calibrate their influence on model training. In detail, we introduce modulating factors based on the sample uncertainty estimation respectively on the two parts of the original learning objective of prompt tuning for OOD detection (refer to Eq (4)). Under this new learning framework, the model's attention is directed towards the classification task to better generalize to the downstream ID dataset when training with low-confidence data. OOD features extracted from high-confidence ID data are attached more importance to achieve more effective OOD regularization. The redirection effect of these two modulating factors facilitates VLMs learning from imperfect OOD features to ultimately improve the OOD detection of prompt tuning. Our main contributions can be summarized as follows,

- Conceptually, we investigate the problem of imperfect OOD features extracted in prompt tuning based OOD detection methods and observe that ID data with different prediction uncertainty exhibits a distinctive influence on the OOD regularization. (in Section 3.2)
- Technically, we propose a novel learning framework, namely *Self-Calibrated Tuning* (SCT), for facilitating prompt tuning for effective OOD detection given only few-shot ID samples, which conducts adaptive redirection of model optimization process between the two tasks to calibrate the influence of OOD features mined from different ID data. (in Sections 3.3)
- Empirically, extensive experiments from different perspectives are conducted to verify the effectiveness of SCT in improving OOD detection performance. To be specific, SCT improves the false positive rate (FPR95) by 3% compared to the previous best method on the

large-scale ImageNet-1k [Deng et al., 2009] benchmark. Furthermore, we perform various ablations and further discussions to provide a thorough understanding. (in Section 4)

2 Related works

Prompt tuning for VLMs. The concept of prompt tuning was originally applied in the field of natural language processing [Radford et al., 2018]. To eliminate the need for manual prompt crafting, prompt tuning exploits supervision signals from downstream tasks to automate the process of prompt generation. Autoprompt [Shin et al., 2020] searches for tokens that cause the greatest changes in gradients based on the label likelihood. Prefix-Tuning [Li and Liang, 2021] introduces a sequence of continuous vectors that can be end-to-end optimized in the token embedding space. Incorporating prompt tuning into computer vision. CoOp [Zhou et al., 2022a] adapts pretrained vision-language models by optimizing a set of learnable continuous prompt vectors. Various prompt tuning methods [Zhou et al., 2022b, Khattak et al., 2023, Sun et al., 2022] have been subsequently proposed to address different vision tasks. However, since these methods are not developed for OOD detection, they face challenges in identifying unknown OOD samples encountered at inference stages.

Out-of-distribution detection with VLMs. Large pretrained vision-language models have enriched the landscape of OOD detection through their remarkable generalization capability in both visual and textual domains. MCM [Ming et al., 2022a] employs the concept of maximum softmax probability [Hendrycks and Gimpel, 2017] into the inference process of CLIP for OOD detection, while CLIPN [Wang et al., 2023] trains an additional text encoder using and set of prompts with a large external dataset to improve its negative semantic understanding. Compared with zero-shot methods, prompt tuning based approaches achieve better OOD detection with access to few-shot ID training data. LoCoOp [Miyai et al., 2024b] adopts prompt tuning and extracts ID-irrelevant background from CLIP's local features as surrogate OOD data to regularize the learned prompts. [Bai et al., 2023] discovers ID-like outliers from ID samples via random cropping to learn a set of negative prompts. However, these two representative prompt learning-based methods suffer from spurious OOD features extracted from ID data due to the imperfect foreground-background decomposition of VLMs. In addition, LSN [Nie et al., 2023] introduce negative prompts to empower VLMs to learn negative semantics from ID samples while NegPrompt [Li et al., 2024] leverages negative prompts to investigate the novel setting of open-vocabulary OOD detection.

3 Method

In this section, we introduce our new framework, i.e., *Self-Calibrated Tuning* (SCT), which conducts adaptive redirection of model learning far away from OOD data region during prompt tuning with only the few-shot ID data. Firstly, we provide preliminaries and notations about prompt tuning based OOD detection (Section 3.1). Secondly, we present and discuss the critical motivation that inspires our method (Section 3.2). Thirdly, we introduce its newly derived learning objective with the explanation and analysis of the underlying intuition and present its algorithmic realization (Section 3.3).

3.1 Preliminaries

VLM-based OOD detection aims to identify test samples that do not belong to any ID class designated by the downstream tasks [Miyai et al., 2024a]. Therefore, the ID distribution is defined by the ID classes from the downstream tasks, which are different from those of the upstream pretraining. Formally, we consider multi-class classification as the original training task [Nguyen et al., 2015], where $\mathcal{X} \subset \mathbb{R}^d$ denotes the input space and $\mathcal{Y} = \{1, \dots, M\}$ denotes the label space. A reliable classifier should be able to detect the OOD input, which can be considered a binary classification problem. We consider $\mathcal{D}_{\text{in}}$ as the distribution of ID data over pairs of examples $\boldsymbol{x} \in \mathbb{R}^d$ and corresponding labels $y \in \mathcal{Y}$. At test time, the environment can present a distribution $\mathcal{D}_{\text{out}}$ over $\mathcal{X}$ of OOD data. In general, the OOD distribution $\mathcal{D}_{\text{out}}$ is defined as an irrelevant distribution of which the label set has no intersection with $\mathcal{Y}$ [Zhu et al., 2023a] and thus should not be predicted by the model parameterized by θ. A decision model $\Gamma(\cdot)$ can be made with the threshold μ:

$$\Gamma_{\mu}(\boldsymbol{x}; \theta) = \begin{cases} \text{ID} & S(\boldsymbol{x}; \theta) \geq \mu \\ \text{OOD} & S(\boldsymbol{x}; \theta) < \mu \end{cases}, \tag{1}$$

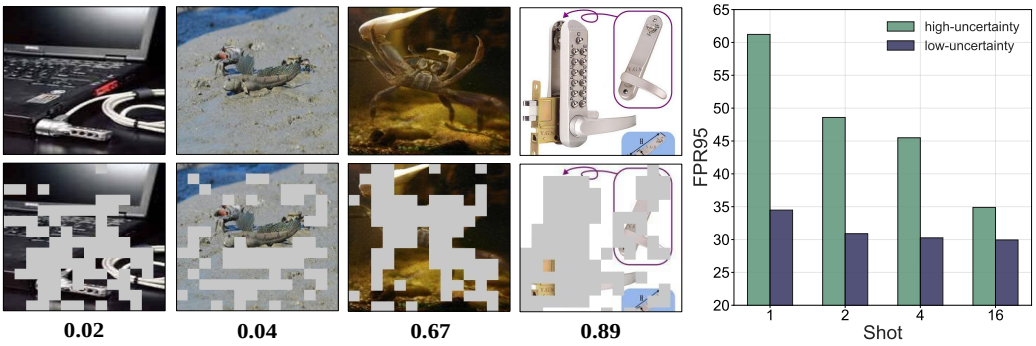


Figure 2: Empirical demonstration about invalid OOD features extracted from ID data in LoCoOp and the influence of sample uncertainty on OOD detection performance. In the left and middle panels, we illustrate the extracted OOD features at different levels of uncertainty and find that they become unreliable as the uncertainty increases. The numbers at the bottom denote the prediction probability for the ground-truth labels from fine-tuned CLIP. In the right panel, we collect ID samples of different uncertainty levels based on prompt-tuned CLIP and divide them into 2 groups. The result demonstrates that the OOD detection performance of LoCoOp significantly degrades as the uncertainty level of ID data rises. We leave the experimental details in Appendix A.3.1 for reference.

Vanilla prompt tuning. Given an ID image $\boldsymbol{x}$ and its corresponding label y, a global visual feature $\boldsymbol{f} = f(\boldsymbol{x})$ is obtained by the visual encoder f of CLIP. Then, the textual prompt vectors can be formulated as $\boldsymbol{t}_m = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, ..., \boldsymbol{\omega}_N, \boldsymbol{c}_m\}$, where $\boldsymbol{c}_m$ denotes the word embedding of the ID class name and $\boldsymbol{\omega} = \{\boldsymbol{\omega}_n|_{n=1}^{N}\}$ are N learnable context vectors, each of which has the same dimension as the word embedding. The text encoder g takes prompt $\boldsymbol{t}_m$ as the input and outputs the textual feature as $\boldsymbol{g}_m = g(\boldsymbol{t}_m)$. The final prediction probability of the CLIP model is computed as follows:

$$p(y = m \mid \boldsymbol{x}; \boldsymbol{\omega}) = \frac{\exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M} \exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_{m'}\right)/\tau\right)}, \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and τ represents the temperature of Softmax. We use $\boldsymbol{p}(\boldsymbol{x};\boldsymbol{\omega})$ to represent the probability vector $[p(y=1|\boldsymbol{x};\boldsymbol{\omega}), p(y=2|\boldsymbol{x};\boldsymbol{\omega}), \ldots, p(y=M|\boldsymbol{x};\boldsymbol{\omega})]$, denoting the prediction probability of ID image $\boldsymbol{x}$ for every ID class.

Prompt tuning for OOD detection. Compared with vanilla prompt tuning methods like CoOp [Zhou et al., 2022a], advanced prompt tuning based OOD detection methods extract surrogate OOD features from ID data via various methods to perform OOD regularization. LoCoOp [Miyai et al., 2024b] can further improve the detection performance by regularizing the learnable prompts $\boldsymbol{\omega}$ on OOD features $\tilde{\boldsymbol{X}}$ extracted from ID local features using a ranking-based method, and its corresponding learning objective is defined as follows,

$$\mathcal{L}_{\text{LoCoOp}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}}\left[\ell_{\text{CE}}(p(y|\boldsymbol{x};\boldsymbol{\omega}), y) + \lambda \ell_{\text{OOD}}(\boldsymbol{p}(\tilde{\boldsymbol{X}};\boldsymbol{\omega}))\right], \tag{3}$$

where λ is the balancing parameter, $\ell_{\text{CE}}(\cdot)$ is the Cross-Entropy (CE) loss, $\boldsymbol{p}(\tilde{\boldsymbol{X}})$ is the prediction probability vector for $\tilde{\boldsymbol{X}}$ and $\ell_{\text{OOD}}(\cdot)$ is the negative entropy of the given probability vector. IDLike [Bai et al., 2023] conducts multiple random cropping on ID data and chooses cropped regions as OOD features based on their feature similarity with textual features of ID classes.

3.2 Motivation

Given the only ID data, previous studies propose to extract the ID-irrelevant local context as the surrogate OOD source, which depends on the foreground-background decomposition using CLIP. However, since the model itself has uncertainty on the prediction results, the correctness of the decomposition cannot always be guaranteed. Thus, as illustrated in Figure 1, the extracted local regions based on the prediction of CLIP may result in spurious OOD features, which then limits the OOD detection performance. Thus, it naturally motivates the following critical research question:

> *How could we better utilize the surrogate OOD features extracted by imperfect foreground-background decomposition of CLIP for effective OOD regularization?*

In this work, we conduct a proof-of-concept experiment to investigate the relationship between sample uncertainty and OOD detection performance of prompt tuning based methods. We use ViT-B/16 CLIP as the model and large-scale ImageNet-1k OOD benchmarks. First, we observe that the quality of OOD features extracted by CLIP is highly correlated with the uncertainty estimation of ID data. The extracted OOD features become more spurious as the uncertainty escalates, as illustrated in the left and middle panel of Figure 2. Furthermore, as empirically shown in the right panel of Figure 2, the performance of prompt tuning based methods is heavily affected by the overall uncertainty level of the given ID data. Intuitively, as the predictions of the model on ID samples are less confident, the OOD features extracted from these data are less reliable. Regularizing on such invalid OOD features can undermine the calibration ability and OOD detection performance of CLIP, which can be reflected by the higher FPR95 score (indicating a higher error on OOD detection). Therefore, a new mechanism is required to take the sample uncertainty into consideration to assist the model in learning from these imperfect OOD features for more effective OOD detection.

3.3 Self-calibrated tuning

As aforementioned, the LoCoOp-based OOD detection paradigm relies on the extracted ID-background local features for OOD regularization. Given the imperfect foreground-background decomposition, the model is expected to effectively learn from the inaccurate OOD features for better OOD detection. Inspired by the previous observation as shown in Section 3.2, one conceptual idea to mitigate this problem is to adaptively adjust the importance of OOD regularization generated from different ID samples based on uncertainty estimation to alleviate the wrong guidance of invalid OOD features. Under this learning paradigm, the model can be regularized by more valid OOD features and simultaneously prevent itself from overconfidence to improve OOD detection. To this intuition, we consider reformulating the learning objective under the framework of prompt tuning as follows,

$$\mathcal{L}_{\text{SCT}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}}\left[\ell_{\text{CE}}(p(y|\boldsymbol{x};\boldsymbol{\omega}),y) * \phi(p(y|\boldsymbol{x};\boldsymbol{\omega})) + \lambda\ell_{\text{OOD}}(\boldsymbol{p}(\tilde{X});\boldsymbol{\omega}) * \psi(p(y|\boldsymbol{x};\boldsymbol{\omega}))\right], \quad (4)$$

where $\phi : \mathbb{R}^M \to \mathbb{R}$ and $\psi : \mathbb{R}^M \to \mathbb{R}$ indicate the newly-introduced modulating functions that calculate adaptive factors for the two components of the original loss function (i.e., Eq (3)) of LoCoOp based on the uncertainty estimation. In this loss function, the left part is for the ID classification task, and the right part is for the OOD regularization. Specifically, ϕ should be monotonically decreasing and ψ should be monotonically increasing with respect to $p(y|\boldsymbol{x};\boldsymbol{\omega})$ so that the modulating factors shift the focus of prompt learning between the two tasks during the training process.

In detail, when the model outputs low-confidence prediction for the ground-truth label, the importance of the ID classification task is highlighted in order to better generalize to the downstream task and simultaneously reduce the effect of regularization from invalid OOD features extracted from ID data. When the model can accurately and confidently classify the ID samples, its attention is redirected towards OOD regularization to strengthen the positive effect of useful ID-irrelevant features for better OOD detection. In the meantime, the loss contribution of the classification task is reduced to avoid the model overfitting to the downstream dataset, which benefits the calibration of the model [Mukhoti et al., 2020] and further enhances the validity of extracted OOD features.

Under this learning framework, the goal of confidence calibration and OOD detection can benefit from each other. This method calibrates the influence of OOD features mined from different ID data based on model prediction confidence during the training process to facilitate capturing more reliable OOD features from ID data. Among a wide range of functions that satisfy the simple requirements as discussed above, we choose the linear function due to its simple design. Concretely, we formulate the loss function of SCT as follows,

$$\mathcal{L}_{\text{SCT}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}}\left[\ell_{\text{CE}}(p(y|\boldsymbol{x};\boldsymbol{\omega}),y) * (1 - p(y|\boldsymbol{x};\boldsymbol{\omega})) + \lambda\ell_{\text{OOD}}(\boldsymbol{p}(\tilde{X};\boldsymbol{\omega})) * p(y|\boldsymbol{x};\boldsymbol{\omega})\right], \quad (5)$$

This implementation of $\mathcal{L}_{\text{SCT}}$ introduces no extra hyperparameters and we empirically demonstrate its effectiveness in the following experiment section 4. In the Appendix A.3.2, we consider other instantiations of the modulating function and demonstrate that these can be comparably effective.

Extraction of OOD local features. We adopt the ranking-based method as suggested by LoCoOp [Miyai et al., 2024b] to extract OOD local features from ID samples for OOD regularization.

Algorithm 1 Self-Calibrated Tuning(SCT)

Input: learnable prompt: $\boldsymbol{\omega}$, fine-tuning epochs: T, learning rate: η, a training set of ID data, extraction rank parameter: K, regularization term weight: λ;
Output: fine-tuned prompt $\boldsymbol{\omega}$;

for epoch $= 1, \ldots, T$ **do**
  for mini-batch $= 1, \ldots, M$ **do**
    Sample a mini-batch $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ from the training set
    Sample surrogate OOD features $\tilde{X}$ from ID local feature map by Eq. (6)
    $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \eta \nabla_{\boldsymbol{\omega}} \left\{ \frac{1}{n} \sum \ell_{\text{CE}}(p(y|\boldsymbol{x};\boldsymbol{\omega}), y) * \phi(p(y|\boldsymbol{x};\boldsymbol{\omega})) + \lambda \ell_{\text{OOD}}(\boldsymbol{p}(\tilde{X});\boldsymbol{\omega}) * p(y|\boldsymbol{x};\boldsymbol{\omega}) \right\}$
  end for
end for

To be specific, we select the region indices of ID-irrelevant regions from a set of all region indices $I = \{0, 1, 2, ..., H \times W - 1\}$, where H and W denote the height and width of the feature map. We calculate the classification prediction probabilities for each region i during training by computing the similarity between the image features $\boldsymbol{f}^{(i)}$ of each region i and the text features of the ID classes. Then we choose regions that do not include their ground truth class in the top-K predicted classes as ID-irrelevant regions J. The corresponding formulation is presented as follows:

$$J = \{i \in I : \text{rank}(p^{(i)}(y|\boldsymbol{x};\boldsymbol{\omega})) > K\}, \tag{6}$$

where $p^{(i)}(y|\boldsymbol{x};\boldsymbol{\omega})$ denotes the prediction probability of region i for the ground-truth label and $\text{rank}(p^{(i)}(y|\boldsymbol{x};\boldsymbol{\omega}))$ denotes the rank of the ground-truth label among all ID classes. We summarize the whole procedure of the proposed SCT in Algorithm 1.

Test-time OOD detection. At the testing stage, we use the GL-MCM score proposed by [Miyai et al., 2023] since it has been empirically proved to outperform the conventional MCM score. It combines the maximum softmax probability scores for both global and local image features. The detailed formulation is presented as follows:

$$S_{\text{GL-MCM}} = \max_{m} \frac{\exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M} \exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_{m'}\right)/\tau\right)} + \max_{m,i} \frac{\exp\left(\text{sim}\left(\boldsymbol{f}^{(i)}, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M} \exp\left(\text{sim}\left(\boldsymbol{f}^{(i)}, \boldsymbol{g}_{m'}\right)/\tau\right)}. \tag{7}$$

where $\boldsymbol{f}^{(i)}$ denotes the local image feature of ID samples $\boldsymbol{x}$ for region i and we set $\tau = 1$.

Comparison and compatibility. Compared with the previous prompt tuning based OOD detection algorithm [Miyai et al., 2024b], the critical idea behind SCT is trying to adaptively adjust the contribution of OOD features extracted from ID data with different uncertainty during training. It provides a general framework (under the guidance of the learning objective in Eq. (4)) to assist VLMs in learning from spurious OOD features for effective OOD detection. The discriminative feature regularized by our SCT can be utilized by those advanced post-hoc scoring functions [Sun et al., 2021, Huang et al., 2021, Sun and Li, 2022]. For prompt tuning based methods, the adaptive modulation introduced in SCT is orthogonal to current tuning objectives [Liu et al., 2020] and also compatible with different augmentation [Lu et al., 2023] or mining strategies [Bai et al., 2023, Zhu et al., 2023b].

4 Experiment

In this section, we present the comprehensive verification of the proposed SCT in the CLIP-based OOD detection scenario. First, we provide the experimental setups in detail (in Section 4.1). Secondly, we provide the performance comparison of our approach with a series of CLIP-based post-hoc methods and prompt tuning based methods (in Section 4.2). Thirdly, we conduct various ablation studies and further discussions to understand our method (in Section 4.3).

4.1 Experimental setup

Datasets. Following the common benchmarks used in previous works, we adopt the ImageNet-1K dataset [Deng et al., 2009] as the ID data. For OOD datasets, we adopt the same ones as in [Huang

and Li, 2021], including subsets of iNaturalist [Van Horn et al., 2018], SUN [Xiao et al., 2010], Places [Zhou et al., 2017], and TEXTURE [Cimpoi et al., 2014]. For the few-shot training, we use 1, 2, 4, and 16 shots ID data for training, respectively, and evaluate models in the full test set. We also present the comparison results on conventional CIFAR benchmarks [Hendrycks et al., 2019, Liu et al., 2020], which adopt CIFAR-10 and CIFAR-100 as ID datasets [Krizhevsky, 2009], in Appendix A.3.2.

Evaluation metrics. We employ the following two common metrics to evaluate the performance of OOD detection: (i) Area Under the Receiver Operating Characteristic curve (AUROC) [Davis and Goadrich, 2006] can be interpreted as the probability for a positive sample to have a higher discriminating score than a negative sample [Fawcett, 2006]; (ii) False Positive Rate (FPR) at 95% True Positive Rate (TPR) [Liang et al., 2018] indicates the probability for a negative sample to be misclassified as positive when the true positive rate is at 95%. We also adopt in-distribution testing accuracy (ID-ACC) to measure the preservation level of the performance for the original classification task on ID data and use Expected Calibration Error (ECE) [Naeini et al., 2015] to measure the performance of SCT on confidence calibration, which are both presented in Appendix A.3.2.

Implementation details. Following previous works [Tao et al., 2023, Ming et al., 2022a, Miyai et al., 2023], we use ViT-B/16 [Dosovitskiy et al., 2021] as the backbone model for the main experiment. For the hyperparameter K in the surrogate OOD features extraction, we use 200 in all experiments as recommended by [Miyai et al., 2024b]. For SCT, we adopt $\lambda = 0.4$ under the 1-shot setting and $\lambda = 0.2$ under the 16-shot setting. We train the CLIP for 25 epochs with a learning rate of 0.002 and other hyperparameters (*e.g.* batch size=32, SGD optimizer and token lengths N=16) are the same as those of CoOp [Zhou et al., 2022a]. We use two Nvidia 3090 GPUs for all experiments.

OOD detection baselines. We compare SCT with several competitive CLIP-based OOD detection methods in the two directions, including post-hoc methods and prompt tuning based methods. For post-hoc methods, we compare with MCM [Ming et al., 2022a] and GL-MCM [Miyai et al., 2024b] as zero-shot baselines. We also adopt Maximum Softmax Probability (MSP) [Hendrycks and Gimpel, 2017], ODIN [Liang et al., 2018], ReAct [Sun et al., 2021], MaxLogit [Hendrycks et al., 2022], and Energy score [Liu et al., 2020] as conventional scoring function baselines. In addition, we provide more discussions about post-hoc methods and our methods in Appendix A.3.2. For prompt tuning based methods, we adopt CoOp [Zhou et al., 2022a], LoCoOp [Miyai et al., 2024b], IDLike [Bai et al., 2023], NegPrompt [Li et al., 2024] and LSN [Nie et al., 2023] as baselines. For all prompt tuning based methods, we constrain all major experiments to few-shot learning scenarios, which is more practical in real cases. Note that LSN is a general learning framework that can be combined with various prompt tuning methods and we report the result of LSN incorporated with LoCoOp in Table 1. We leave more definitions and implementation details in the Appendix A.1.

4.2 Main results

In this part, we present the major performance comparison with some representative baseline methods for OOD detection to demonstrate the effectiveness of the proposed SCT. Specifically, we consider several zero-shot methods as the performance reference based on the pretrained CLIP and some prompt tuning based methods for specific comparison on fine-tuning with few-shot ID data. Note that we leave the experiment results on 2-shot and 4-shot settings in Appendix A.3.2.

Comparisons on conventional OOD detection In Table 1, we present the overall results of the comparison between different baseline methods and SCT for OOD detection. Since the prompt tuning based methods engage the ID data during training, the model will generally gain better empirical performance on OOD detection, reflected by evaluation metrics like FPR95 and AUROC. IDLike, NegPrompt, and LSN all introduce a set of negative prompts for each ID class to learn negative semantics of ID objects using different strategies, which obtain different levels of detection performance gains. Without sacrificing much classification performance (i.e., ID classification accuracy) on ID data, as shown in Table 6, our SCT can consistently achieve better OOD detection performance on the large-scale ImageNet-1k benchmark, which verifies the effectiveness of our methods with the newly proposed modulation factors.

Compatibility with other baselines. In Table 2, we report the results of compatibility experiments, in which we compare those prompt tuning based methods with their variants, incorporating our SCT to dynamically adjust the importance of OOD regularization from ID samples with different uncertainty levels. We can find that our SCT can consistently help them gain better or comparable

Table 1: **Comparison results on ImageNet-1k OOD benchmarks.** All methods are trained on the same backbone CLIP-ViT-B/16. Bold numbers are superior results. ↑ indicates larger values are better, and ↓ indicates smaller values are better. Results marked with † are taken from [Wang et al., 2023] and [Miyai et al., 2024b]. The prompt tuning based methods are run under multiple trials with reporting the mean and standard deviation of the performance.

Method	OOD dataset									
	iNaturalist		SUN		Places365		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-shot methods										
MCM	31.86	94.17	37.28	92.55	42.94	90.09	58.37	85.83	42.61	90.66
GL-MCM	15.16	96.71	29.16	93.41	37.07	90.37	58.85	83.11	35.06	90.90
CLIP-based post-hoc methods										
MSP†	74.57	77.74	76.95	73.97	79.72	72.18	73.66	74.84	74.98	76.22
ODIN †	98.93	57.73	88.72	78.42	87.80	76.88	85.47	71.49	90.23	71.13
Energy†	64.98	87.18	46.42	91.17	57.40	87.33	50.39	88.22	54.80	88.48
ReAct†	65.57	86.87	46.17	91.04	56.85	87.42	49.88	88.13	54.62	88.37
MaxLogit†	60.88	88.03	44.83	91.16	55.54	87.45	48.72	88.63	52.49	88.82
Prompt tuning based methods										
1-shot										
CoOp	$43.80^{\pm 4.33}$	$91.40^{\pm 0.70}$	$35.42^{\pm 1.56}$	$92.65^{\pm 0.53}$	$40.70^{\pm 2.63}$	$90.49^{\pm 0.68}$	$49.61^{\pm 3.03}$	$87.95^{\pm 0.79}$	$42.38^{\pm 1.92}$	$90.63^{\pm 0.38}$
LoCoOp	$28.81^{\pm 2.78}$	$94.05^{\pm 0.72}$	$25.76^{\pm 0.53}$	$94.51^{\pm 0.19}$	$33.68^{\pm 0.62}$	$91.59^{\pm 0.23}$	$51.53^{\pm 0.42}$	$86.85^{\pm 0.24}$	$34.94^{\pm 0.65}$	$91.75^{\pm 0.23}$
IDLike	$12.07^{\pm 0.88}$	$97.65^{\pm 0.10}$	$40.55^{\pm 5.84}$	$91.07^{\pm 1.80}$	$47.94^{\pm 5.24}$	$88.31^{\pm 2.05}$	$38.34^{\pm 13.39}$	$89.67^{\pm 4.03}$	$34.72^{\pm 0.80}$	$91.67^{\pm 0.07}$
NegPrompt	$65.03^{\pm 8.69}$	$84.56^{\pm 2.52}$	$44.39^{\pm 1.66}$	$89.63^{\pm 0.66}$	$51.31^{\pm 6.21}$	$86.55^{\pm 2.19}$	$87.60^{\pm 1.61}$	$63.76^{\pm 3.02}$	$62.08^{\pm 3.71}$	$81.13^{\pm 1.78}$
LSN	$59.28^{\pm 7.02}$	$87.20^{\pm 3.15}$	$40.15^{\pm 0.82}$	$91.47^{\pm 0.14}$	$46.11^{\pm 1.86}$	$88.74^{\pm 0.57}$	$60.34^{\pm 0.14}$	$83.92^{\pm 0.42}$	$51.47^{\pm 1.53}$	$87.84^{\pm 0.58}$
SCT	$19.16^{\pm 1.15}$	$95.70^{\pm 0.28}$	$23.52^{\pm 1.91}$	$94.58^{\pm 0.56}$	$32.81^{\pm 1.14}$	$91.23^{\pm 0.32}$	$48.87^{\pm 1.38}$	$86.66^{\pm 0.33}$	$\mathbf{31.09}^{\pm 0.48}$	$\mathbf{92.04}^{\pm 0.25}$
16-shot										
CoOp	$28.25^{\pm 4.68}$	$93.92^{\pm 1.17}$	$31.15^{\pm 0.89}$	$93.13^{\pm 0.38}$	$39.12^{\pm 1.05}$	$90.50^{\pm 0.37}$	$41.86^{\pm 1.88}$	$90.40^{\pm 0.54}$	$35.09^{\pm 1.60}$	$91.99^{\pm 0.42}$
LoCoOp	$17.58^{\pm 2.22}$	$96.30^{\pm 0.57}$	$22.82^{\pm 0.34}$	$95.20^{\pm 0.06}$	$32.21^{\pm 0.53}$	$92.03^{\pm 0.20}$	$45.27^{\pm 0.95}$	$88.86^{\pm 0.26}$	$29.47^{\pm 0.29}$	$93.10^{\pm 0.03}$
IDLike	$9.71^{\pm 0.60}$	$98.05^{\pm 0.07}$	$38.93^{\pm 0.10}$	$90.54^{\pm 0.68}$	$47.06^{\pm 1.44}$	$88.06^{\pm 0.90}$	$32.82^{\pm 5.12}$	$91.89^{\pm 1.49}$	$32.12^{\pm 1.09}$	$92.14^{\pm 0.01}$
NegPrompt	$37.79^{\pm 0.11}$	$90.49^{\pm 0.01}$	$32.11^{\pm 3.77}$	$92.25^{\pm 1.00}$	$35.52^{\pm 0.41}$	$91.16^{\pm 0.03}$	$43.93^{\pm 9.09}$	$88.38^{\pm 3.31}$	$37.34^{\pm 1.41}$	$90.57^{\pm 0.59}$
LSN	$36.17^{\pm 4.81}$	$92.66^{\pm 1.16}$	$34.27^{\pm 0.44}$	$93.53^{\pm 0.20}$	$41.47^{\pm 0.85}$	$90.52^{\pm 0.37}$	$46.43^{\pm 0.60}$	$89.38^{\pm 0.24}$	$39.58^{\pm 0.73}$	$91.53^{\pm 0.09}$
SCT	$13.94^{\pm 0.68}$	$95.86^{\pm 0.28}$	$20.55^{\pm 1.07}$	$95.33^{\pm 0.12}$	$29.86^{\pm 0.67}$	$92.24^{\pm 0.05}$	$41.51^{\pm 0.48}$	$89.06^{\pm 0.09}$	$\mathbf{26.47}^{\pm 0.39}$	$\mathbf{93.37}^{\pm 0.07}$

Table 2: OOD detection performance on compatibility experiments. All methods are trained on the same backbone. ↑ indicates larger values are better, and ↓ indicates smaller values are better. Methods are run under multiple trials reporting the mean and standard deviation of the performance.

Method	OOD Dataset									
	iNaturalist		SUN		Places365		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1-shot										
LoCoOp	$28.81^{\pm 2.78}$	$94.05^{\pm 0.72}$	$25.76^{\pm 0.53}$	$94.51^{\pm 0.19}$	$33.68^{\pm 0.62}$	$91.59^{\pm 0.23}$	$51.53^{\pm 0.42}$	$86.85^{\pm 0.24}$	$34.94^{\pm 0.65}$	$91.75^{\pm 0.23}$
SCT	$19.16^{\pm 1.15}$	$95.70^{\pm 0.28}$	$23.52^{\pm 1.91}$	$94.58^{\pm 0.56}$	$32.81^{\pm 1.14}$	$91.23^{\pm 0.32}$	$48.87^{\pm 1.38}$	$86.66^{\pm 0.33}$	$\mathbf{31.09}^{\pm 0.48}$	$\mathbf{92.04}^{\pm 0.25}$
IDLike	$12.07^{\pm 0.88}$	$97.65^{\pm 0.10}$	$40.55^{\pm 5.84}$	$91.07^{\pm 1.80}$	$47.94^{\pm 5.24}$	$88.31^{\pm 2.05}$	$38.34^{\pm 13.39}$	$89.67^{\pm 4.03}$	$34.72^{\pm 0.80}$	$\mathbf{91.67}^{\pm 0.07}$
IDLike+SCT	$12.24^{\pm 2.09}$	$97.58^{\pm 0.54}$	$31.98^{\pm 1.07}$	$92.27^{\pm 0.26}$	$44.79^{\pm 2.82}$	$87.62^{\pm 0.69}$	$43.57^{\pm 1.99}$	$86.71^{\pm 1.3}$	$\mathbf{33.14}^{\pm 0.58}$	$90.97^{\pm 0.36}$
LSN	$59.28^{\pm 7.02}$	$87.20^{\pm 3.15}$	$40.15^{\pm 0.82}$	$91.47^{\pm 0.14}$	$46.11^{\pm 1.86}$	$88.74^{\pm 0.57}$	$60.34^{\pm 0.14}$	$83.92^{\pm 0.42}$	$51.47^{\pm 1.53}$	$87.84^{\pm 0.58}$
LSN+SCT	$51.38^{\pm 1.24}$	$89.54^{\pm 0.14}$	$35.60^{\pm 0.53}$	$92.48^{\pm 0.09}$	$40.84^{\pm 0.59}$	$90.29^{\pm 0.10}$	$55.24^{\pm 0.01}$	$86.00^{\pm 0.03}$	$\mathbf{45.76}^{\pm 0.59}$	$\mathbf{89.58}^{\pm 0.09}$
16-shot										
LoCoOp	$17.58^{\pm 2.22}$	$96.30^{\pm 0.57}$	$22.82^{\pm 0.34}$	$95.20^{\pm 0.06}$	$32.21^{\pm 0.53}$	$92.03^{\pm 0.20}$	$45.27^{\pm 0.95}$	$88.86^{\pm 0.26}$	$29.47^{\pm 0.29}$	$93.10^{\pm 0.03}$
SCT	$13.94^{\pm 0.68}$	$95.86^{\pm 0.28}$	$20.55^{\pm 1.07}$	$95.33^{\pm 0.12}$	$29.86^{\pm 0.67}$	$92.24^{\pm 0.05}$	$41.51^{\pm 0.48}$	$89.06^{\pm 0.09}$	$\mathbf{26.47}^{\pm 0.39}$	$\mathbf{93.37}^{\pm 0.07}$
IDLike	$9.71^{\pm 0.60}$	$98.05^{\pm 0.07}$	$38.93^{\pm 0.10}$	$90.54^{\pm 0.68}$	$47.06^{\pm 1.44}$	$88.06^{\pm 0.90}$	$32.82^{\pm 5.12}$	$91.89^{\pm 1.49}$	$32.12^{\pm 1.09}$	$\mathbf{92.14}^{\pm 0.01}$
IDLike+SCT	$3.41^{\pm 1.88}$	$98.97^{\pm 0.38}$	$33.98^{\pm 2.30}$	$89.24^{\pm 2.18}$	$37.81^{\pm 0.44}$	$87.49^{\pm 1.33}$	$29.49^{\pm 2.58}$	$90.07^{\pm 0.35}$	$\mathbf{26.17}^{\pm 0.65}$	$91.44^{\pm 0.69}$
LSN	$36.17^{\pm 4.81}$	$92.66^{\pm 1.16}$	$34.27^{\pm 0.44}$	$93.53^{\pm 0.20}$	$41.47^{\pm 0.85}$	$90.52^{\pm 0.37}$	$46.43^{\pm 0.60}$	$89.38^{\pm 0.24}$	$39.58^{\pm 0.73}$	$91.53^{\pm 0.09}$
LSN+SCT	$35.14^{\pm 0.89}$	$92.64^{\pm 0.06}$	$29.80^{\pm 1.65}$	$94.06^{\pm 0.30}$	$37.00^{\pm 0.90}$	$91.25^{\pm 0.17}$	$45.45^{\pm 1.34}$	$89.46^{\pm 0.46}$	$\mathbf{36.85}^{\pm 0.08}$	$\mathbf{91.85}^{\pm 0.01}$

OOD detection performance across two evaluation metrics while keeping the classification accuracy comparable with the vanilla prompt-tuned model, as shown in Appendix A.3.2.

Comparisons on hard OOD detection. Following the setup in MCM [Ming et al., 2022a], we also explore the performance of SCT on hard OOD detection tasks, as shown in Table 3. SCT significantly outperforms LoCoOp in all four experimental settings, demonstrating that SCT has strong discriminative power for semantically hard OOD data.

4.3 Ablation study

In this part, we conduct various ablation experiments and further explorations to provide a thorough understanding of the characteristics of our proposed SCT. For the extra results and discussions (e.g., computational cost and social impact), we leave more details in Appendix A.3.2.

Table 3: OOD detection performance comparison with LoCoOp on hard OOD detection tasks. Bold numbers represents superior results.

ID Dataset	OOD Dataset	Method	FPR95↓	AUROC↑
ImageNet-10	ImageNet-20	LoCoOp	28.20	92.75
		SCT	**25.10**	**94.33**
ImageNet-20	ImageNet-10	LoCoOp	34.40	92.34
		SCT	**25.00**	**94.95**
ImageNet-10	ImageNet-100	LoCoOp	30.08	93.00
		SCT	**26.64**	**93.90**
ImageNet-100	ImageNet-10	LoCoOp	61.40	81.97
		SCT	**57.80**	**82.60**

Table 4: Ablation study on the modulation factors for the classification task and regularization term. Detailed results can be found in Appendix A.3.2.

Shot	ϕ	ψ	FPR95↓	AUROC↑	ID-ACC↑
1-shot	✗	✗	34.94	91.75	**69.03**
	✓	✗	31.14	**92.35**	68.60
	✗	✓	31.90	91.74	68.70
	✓	✓	**31.09**	92.04	68.80
16-shot	✗	✗	29.47	93.10	71.43
	✓	✗	29.30	92.66	71.50
	✗	✓	28.94	92.62	**71.90**
	✓	✓	**26.47**	**93.37**	71.77

Importance of the modulation factors in SCT. As an important aspect of the learning objective of SCT in Eq. (4), the newly introduced modulating factors in the learning objective of SCT conduct adaptive redirection of prompt tuning towards suitable tasks. Although each of modulating factors can seemingly achieve the redirection effect alone, we empirically find that both modulating factors play a significant role in enhancing OOD detection performance of VLMs, as shown in Table 4.

Influence of regularization weight in OOD regularization. The regularization weight λ controls the contribution of OOD regularization to the prompt learning process like the role of the two modulation factors. In Figure 3(a), we show the performance by varying the ratio $\lambda = 1$. It is worth noting that setting high regularization weights such as $\lambda = 1$ may even degrade the performance, indicating that the ID classification task should be attached more importance for better OOD detection.

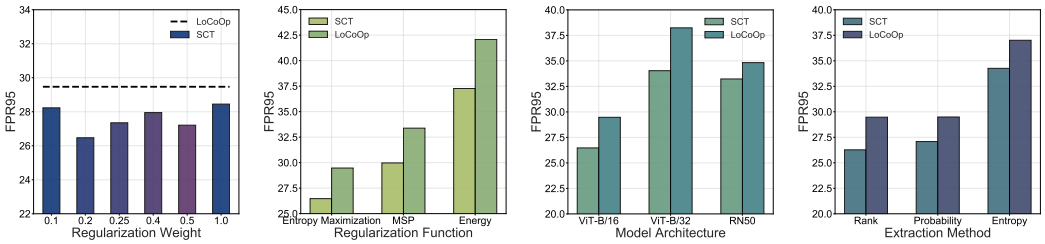
Generality of using different OOD regularization functions. Since SCT introduces a general learning framework of prompt tuning for OOD detection, the specific realization for the OOD regularization function can have multiple choices (e.g., MSP [Hendrycks and Gimpel, 2017] or Energy score function [Liu et al., 2020]). Here we report the performance using different OOD regularization functions in Figure 3(b), where they have different performance improvements compared with the original LoCoOp baseline. Specifically, the energy regularization needs tuning the two energy threshold hyperparameters m_{in} and m_{out}, limiting its advantages over other regularization functions.

Implementation with different CLIP architectures. We evaluate SCT with different VLM architectures and the results are shown in Figure 3(c). The result illustrates that the larger backbone boosts the performance of OOD detection and also shows SCT can outperform LoCoOp across various VLM architectures. It is important to note that we take the same hyperparameters across all architectures, demonstrating the robustness of SCT hyperparameters on different VLM architectures.

Comparison between different methods for extracting OOD features. In Figure 3(d), we perform the comparison of our method adopting different methods for extracting OOD features, including the probability-based and entropy-based methods. The results verify the superiority of SCT on OOD detection across all the OOD feature extraction methods. Specifically, the probability-based method and entropy-based method both have a threshold hyperparameter to discriminate between ID and OOD features. The sensitivity to hyperparameters of different methods might be the reason behind the different OOD detection performance. For the entropy-based method, the performance is significantly poor since it is challenging to determine the appropriate threshold [Miyai et al., 2024b].

5 Discussions and limitations

Comparisons with advanced post-hoc methods. Recently, advanced post-hoc approaches [Djurisic et al., 2022, Xu et al., 2024] exhibit comparable OOD detection performance to tuning based methods, despite the lack of training data. However, prompt tuning based methods can leverage the generalization ability of VLMs to better fit the domains of the downstream tasks given only few-shot ID training data. What's more, post-hoc methods and prompt tuning based methods are compatible with each other, further boosting the OOD detection performance. We provide more detailed discussions, including empirical experiments, on the compatibility of SCT with advanced post-hoc methods in Appendix A.3.2. Furthermore, future research efforts into post-hoc calibration methods for prompt tuning based OOD detection could also contribute to the community.



(a) Regularization Weight (b) Regularization Func. (c) Architecture (d) Extraction Method

Figure 3: Ablation study. (a) performance of using different regularization weights λ; (b) exploration of different regularization functions for OOD regularization; (c) using different CLIP architectures; (d) comparison of different methods for extracting OOD features.

Performance improvement on the AUROC metric. The experiment results in Table 1 and 2 indicate that the improvements of SCT over LoCoOp on AUROC are less notable than the FPR95 metric. However, as shown in the comparison of different baselines in Table 1, the improvement space for FPR95 is significantly larger than AUROC. Therefore, the progress on these two metrics should not be treated equally. Nevertheless, further investigation is necessary to enhance the OOD detection performance specifically targeting improvement on the AUROC metric.

The sensitivity to training data under the few-shot setting. VLMs are exposed to limited ID data samples under the few-shot scenarios, which means that the OOD detection performance of prompt tuning based methods, including SCT, can be susceptible to the quality of limited ID training data in practice. Exploration of selection strategies of suitable training data or overcoming the data sensitivity inherent in the few-shot setting could further facilitate the practical application of prompt tuning based OOD detection methods in real-world scenarios.

Theoretical analysis of the proposed method. Although we propose a novel framework to mitigate the problem of invalid extracted OOD features, we have not yet provided sufficient theoretical analysis to prove the effectiveness of our method. We choose the linear function as the modulation function for the sake of simplicity, instead of based on theoretical justification. Conducting theoretical analyses on the relationship between sample uncertainty and OOD detection performance of prompt tuning based methods under few-shot settings is also a potential direction for future work. .

6 Conclusion

In this paper, we propose a novel learning framework, i.e., *Self-Calibrated Tuning (SCT)*, that improves the OOD detection capability of VLMs with only the given ID training data. To mitigate the problem caused by invalid OOD features mined from ID data, SCT introduces two modulating factors to the original learning objective to conduct adaptive redirection of prompt tuning process between the tasks of ID classification and OOD regularization. Through the redirection effect, our method calibrates the impact of OOD features extracted from different ID samples based on the sample uncertainty estimation during the training process, which facilitates the model learning from imperfect surrogate OOD features for OOD regularization. We have conducted extensive experiments to demonstrate the effectiveness of SCT and its compatibility with a range of prompt tuning based methods, along with various ablation studies and further explorations to characterize the framework.

Acknowledgments and Disclosure of Funding

GY and JCY were supported by the National Key R&D Program of China (No. 2022ZD0160703), 111 plan (No. BP0719010) and National Natural Science Foundation of China (No. 62306178). JNZ and BH were supported by NSFC General Program No. 62376235, Guangdong Basic and Applied Basic Research Foundation Nos. 2022A1515011652 and 2024A1515012399, RIKEN Collaborative Research Fund, HKBU Faculty Niche Research Areas No. RC-FNRA-IG/22-23/SCI/04, and HKBU CSD Departmental Incentive Scheme.

References

Yichen Bai, Zongbo Han, Changqing Zhang, Bing Cao, Xiaoheng Jiang, and Qinghua Hu. Id-like prompt learning for few-shot out-of-distribution detection. *arXiv preprint arXiv:2311.15243*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

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. In *ICLR*, 2021.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 2006.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, 2022.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*, 2024.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. In *arXiv*, 2009.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. *arXiv preprint arXiv:2404.03248*, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

T Lin. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

Fan Lu, Kai Zhu, Kecheng Zheng, Wei Zhai, and Yang Cao. Likelihood-aware semantic alignment for full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2312.01732*, 2023.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022a.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *ICML*, 2022b.

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *AAAI*, 2022c.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. *arXiv preprint arXiv:2304.04521*, 2023.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, et al. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*, 2024a.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *NeurIPS*, 36, 2024b.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *ICLR*, 2023.

Changdae Oh, Mijoo Kim, Hyesu Lim, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *arXiv preprint arXiv:2311.01723*, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.

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

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. In *NeurIPS*, 2020.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *NeurIPS*, 2022.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICLR*, 2023.

Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *ICML*, 2024.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, 2023.

Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for vision-language models. In *ICML*, 2024.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *ICLR*, 2024.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. In *arXiv*, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022b.

Jianing Zhu, Hengzhuang Li, Jiangchao Yao, Tongliang Liu, Jianliang Xu, and Bo Han. Unleashing mask: explore the intrinsic out-of-distribution detection capability. In *ICML*, 2023a.

Jianing Zhu, Geng Yu, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. In *NeurIPS*, 2023b.

A Appendix / Supplemental Material

The whole Appendix is organized as follows. In Appendix A.1, we present the detailed definitions and implementation of zero-shot, post-hoc methods and several prompt tuning based methods that are considered in our experiments. In Appendix A.3, we provide our extra experimental details and more comprehensive results with further discussion on the underlying implications. Finally, in Appendix A.4, we discuss the potential broader impact and limitations of our work.

Reproducibility statement

To ensure reproducibility, we outline several key aspects about the experiments below:

- **Datasets.** The datasets we used are all publicly accessible, which are introduced in Section 4.1.
- **Open source.** The source code is publicly available at `https://github.com/tmlr-group/SCT`. We provide a backbone for our experiments as well as several auxiliary components, such as OOD detection performance evaluation.
- **Environment.** All experiments are conducted with multiple runs on NVIDIA GeForce RTX 3090 GPUs with Python 3.8 and PyTorch 1.12.

A.1 Details about considered baselines

In this section, we present the details about the baselines, including zero-shot, post-hoc methods, and several prompt tuning based methods, as well as related hyper-parameters that are considered in our work.

MSP. Hendrycks and Gimpel [2017] proposes to utilize maximum softmax probability(MSP) as the scoring function to differentiate between ID and OOD samples, of which the definition is as follows,

$$S_{\text{MSP}}(\boldsymbol{x}; f) = \max_m P(y = m|\boldsymbol{x}; f) = \max \ \texttt{softmax}(f(\boldsymbol{x})) \tag{8}$$

where f represents a given well-trained model and m is one of the classes $\mathcal{Y} = \{1, \ldots, M\}$. A higher MSP score signifies a greater probability that a sample belongs to the in-distribution distribution, indicating the model's confidence on the sample.

ODIN. Liang et al. [2018] designs the ODIN score function, utilizing the temperature scaling and minor perturbations to test samples to widen the gap between the distributions of ID and OOD data. The ODIN score is defined as follows,

$$S_{\text{ODIN}}(\boldsymbol{x}; f) = \max_m P(y = m|\tilde{\boldsymbol{x}}; f) = \max \ \texttt{softmax}(\frac{f(\tilde{\boldsymbol{x}})}{T}) \tag{9}$$

where $\tilde{x}$ denotes the perturbed samples (controled by ϵ) and T denotes the temperature.

Energy. Liu et al. [2020] proposes to use the Energy of the prediction logits to discriminate between the ID and OOD samples. The Energy score is formulated as follows,

$$S_{\text{Energy}}(\boldsymbol{x}; f) = -T \log \sum_{m=1}^{M} e^{f(\boldsymbol{x})_c / T} \tag{10}$$

where T denotes the temperature parameter. As theoretically proved in Liu et al. [2020], a lower Energy score represents a higher probability for a sample to belong to ID.

ReAct. Sun et al. [2021] designs a simple and effective approaches, named Rectified Activations (ReAct), to alleviate model overconfidence on out-of-distribution data. This work observes that OOD samples can induce unusually high unit activation in the deep layer of neural networks. ReAct improves OOD detection by simply rectifying the activations at an upper limit $c > 0$, which can be performed on a pretrained model without any modification to training.

MaxLogit. Hendrycks et al. [2022] demonstrates that the conventional MSP OOD detector does not scale well to challenging large-scale multiclass and multi-label OOD detection setting and proposes a surprisingly simple detector based on the maximum logit, called MaxLogit, that greatly outperforms previous post-hoc methods. The definition of MaxLogit is as follows,

$$S_{\text{MaxLogit}}(\boldsymbol{x}; f) = -\max_m f(\boldsymbol{x})_m \tag{11}$$

MCM. Ming et al. [2022a] explores the potential of large-scale vision-language models like CLIP for OOD detection and proposes a simple yet effective zero-shot OOD detection method called Maximum Concept Matching (MCM), which is based on aligning visual features with textual concepts. This method characterizes OOD uncertainty by appropriately scaling of the distance from the visual feature to the nearest ID prototype, of which the formulation is as follows,

$$S_{\text{MCM}} = \max_m \frac{\exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_m\right)/\tau\right)}{\sum_{m'=1}^{M} \exp\left(\text{sim}\left(\boldsymbol{f}, \boldsymbol{g}_{m'}\right)/\tau\right)}. \tag{12}$$

where τ denotes the temperature parameter, $\boldsymbol{f}$ and $\boldsymbol{g}_m$ represents image and text features respectively, and $sim()$ represents the cosine similarity between the image and text features.

GL-MCM. Miyai et al. [2023] designs a zero-shot OOD detection method called Global-Local Maximum Concept Matching(GL-MCM) that leverages the maximum softmax score of both global and local image features and designs. Utilizing the matching score with local features can compensate for the low global matching score and produce more accurate ID confidence. The detailed formulation of GL-MCM is shown in Eq (7).

CoOp. Drawing inspiration from the success in prompt learning within natural language processing, Zhou et al. [2022a] proposes Context Optimization (CoOp), a simple framework specifically to adapt vision-language models like CLIP for downstream image recognition. Concretely, CoOp represents the context words of a prompt as learnable vectors while keeping the entire pretrained encoders fixed. The specific framework of CoOp is presented in Section 3.1.

LoCoOp. Miyai et al. [2024b] introduces an approach named Local regularized Context Optimization (LoCoOp) to enhance the OOD detection performance of prompt tuning. It first extracts ID-irreverent contexts from CLIP's local features and utilizes it as OOD features to perform OOD regularization with entropy maximization. The details of this method are provided in Section 3.1

IDLike. Bai et al. [2023] designs a new framework of prompt tuning for OOD detection to focus on challenging OOD detection scenarios. It first constructs surrogate outliers from ID samples by conducting multiple random cropping on ID data and filtering them based on their cosine similarity with textual features of ID classes. Then it introduces a set of OOD prompts along with conventional ID prompts and optimizes them with a new loss function that consists of in-distribution loss, out-of-distribution loss and diversification regularization.

NegPrompt. Li et al. [2024] proposes a prompt tuning based OOD detection method, named NegPrompt, which learns a set of negative prompts for each ID class, with only ID data, to capture the negative semantics associated with these classes. The training process is divided into two stages. In the first stage, the model learns the positive prompts. In the second stage, the positive prompts are kept frozen and the negative prompts are optimized via three loss functions that enforce the separation between negative prompts and ID images, a proper degree of similarity between negative and positive prompts, and the diversity of the negative prompts.

LSN. Nie et al. [2023] reveals that CLIP cannot fully understand the negative semantics of textual information and proposes to learn a set of negative prompts for each class to alleviate this problem. The learned positive prompt (for all classes) and negative prompts (for each class) are utilized simultaneously to calculate similarity and dissimilarity in the feature space, thereby enhancing the accuracy of OOD sample detection.

A.2 In-depth comparison between SCT and hard example mining.

The part of ID classification in the learning framework of SCT bears a strong resemblance to hard example mining methods, such as focal loss [Lin, 2017]. In this section, we clarify the novelty and insights of our SCT by analyzing the difference between SCT and hard example mining as follows.

Conceptually, the motivation of SCT is to mitigate the problem of unreliable OOD features in prompt tuning based OOD detection methods. Generally, these methods rely on the ID-irrelevant local context extracted by VLMs as the surrogate OOD features to perform regularization, the quality of which is greatly affected by the inaccurate foreground-background decomposition of VLMs. As shown in Figure 1 and Figure 5, although VLMs can mask out some ID-related regions (shown as the grey patches of images), large portions of the extracted OOD features (shown as the colored patches of images) obviously belongs to ID features.

Empirically, we find that the quality of extracted OOD features significantly correlated with the uncertainty level of ID data. As illustrated in the left panel of Figure 2, the extracted OOD features become more inaccurate as the uncertainty increases. In the right panel of Figure 2, we train LoCoOp on multiple data groups with different uncertainty levels. The results demonstrate that the OOD detection performance of LoCoOp can be significantly impacted by the uncertainty level of ID data. Therefore, to mitigate the issue of unreliable OOD features, we propose SCT to calibrate the influence of OOD regularization from different ID samples based on their uncertainty level.

Technically, despite the simple design, SCT is significantly different from hard sample mining. The latter conducts reweighting directly on the samples based on the classification difficulty during training. The former adaptively adjusts the importance between the two components of the original learning objectives for every single sample. Data with high uncertainty are directly down-weighted in hard sample mining while they are utilized more for OOD regularization in SCT. As shown in Table 4, under 16-shot ID data, the OOD detection performance of simply assigning $1-p(y|x)$ to L_{ce} (denoted as ϕ ✓ and ψ ×) are significantly inferior to SCT (denoted as ϕ ✓ and ψ ✓), demonstrating the difference of SCT and hard sample mining.

A.3 Additional experimental results and further discussion

In this section, we provide more experiment results from various perspectives to characterize our proposed SCT. First, we introduce the additional experimental setups for the empirical verification in previous figures and our learning framework. Second, we offer more detailed results and analyses of our method in comparison to other advanced baselines. Finally, more demonstrations of the motivation of our method are provided.

A.3.1 Additional experimental setups

Figure 2. In the right panel of Figure 2, we conduct experiments to investigate the relationship between sample uncertainty and OOD detection performance of prompt tuning based methods. We first calculate the prediction probability for ground-truth labels of all the training samples in a 64-shot training set using a prompt-tuned CLIP model. This model is prompt-tuned with LoCoOp on a 4-shot training set which contains no overlapping samples with the 64-shot set. We use the prediction probability for ground-truth labels to represent uncertainty. High-uncertainty samples are assigned low prediction probability for their ground-truth labels and vice versa. We choose the data with the lowest and highest uncertainty for every ID class to generate two data groups of specific shots with different uncertainty levels respectively, and train the model with LoCoOp on these two data groups.

Figure 3. In Figure 3(b), we explore different regularization functions to perform OOD regularization, including entropy maximization [Miyai et al., 2024b], cross-entropy loss to the uniform distribution [Hendrycks et al., 2019, Ming et al., 2022b] and the energy-based function [Liu et al., 2020], under the 16-shot setting. In Figure 3(a), Figure 3(c) and Figure 3(d), we train the models of all the architectures with 16-shot training datasets. In Figure 3(d), we evaluate the performance of SCT with various methods for ID-irrelevant region extraction. To be specific, we consider three different methods, including the ranking-based method, probability-based method, and entropy-based method. Following the setups in [Miyai et al., 2024b], for the entropy-based method, we extract local regions where the entropy of $\boldsymbol{p}_i(\boldsymbol{x})$ is lower than $\frac{\log M}{2}$ since it is the half value of the maximum entropy of M-dimensional probabilities, as done in previous studies [Saito et al., 2020]. For the

Table 5: OOD detection performance of SCT under various few-shot settings. All methods are trained on the same backbone CLIP-ViT-B/16. Bold numbers are superior results. ↑ indicates larger values are better, and ↓ indicates smaller values are better.

Method	OOD dataset									
	iNaturalist		SUN		Places365		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
					2-shot					
LoCoOp	19.45	96.01	23.02	95.08	32.32	91.96	46.35	88.60	30.28	92.91
SCT	16.99	96.15	21.68	94.79	31.01	92.21	42.62	88.64	28.08	92.95
					4-shot					
LoCoOp	17.32	96.26	23.78	94.82	32.19	91.85	44.41	89.22	29.42	93.04
SCT	13.88	97.03	22.13	94.85	30.20	92.09	45.53	87.96	27.93	92.98

Table 6: ID classification performance of SCT and all the considered baselines. All methods are trained on the same backbone CLIP-ViT-B/16.

Method	ID Accuracy	
	1-shot	16-shot
Zero-shot methods		
MCM	66.7	
GL-MCM	66.7	
CLIP-based post-hoc methods		
MSP	66.7	
ODIN	66.7	
Energy	66.7	
ReAct	66.7	
MaxLogit	66.7	
Prompt tuning based methods		
CoOp	68.83	71.93
LoCoOp	69.03	71.43
IDLike	68.90	71.04
NegPrompt	68.83	71.93
LSN	68.83	71.93
SCT	68.80	71.77
IDLike+SCT	69.70	71.14
LSN+SCT	68.98	71.50

probability-based method, we extract regions where $p_i(y|\boldsymbol{x})$ is lower than $1/M$ simply because the ID dataset has M classes.

A.3.2 Additional experimental results and illustrations

Experiments on multiple few-shot settings. In order to further understand the effectiveness of our SCT on different few-shot settings, we report the evaluation results of our SCT with 2-shot and 4-shot training datasets, which are summarized in Table 5. The results demonstrate that SCT can gain significant improvement on OOD detection under all the few-shot settings.

ID classification performance of baselines and SCT in Table 1 and Table 2. To evaluate the performance of all the considered baselines and SCT on the original classification task, we report the classification accuracy on the ID test set in Table 6. Since NegPrompt and LSN learn positive prompts and negative prompts separately, they have the same ID classification performance as vanilla CoOp. The results show that SCT maintains comparable ID classification performance compared with vanilla prompt tuning CoOp and other advanced prompt tuning based OOD detection methods while achieving more effective OOD detection.

Fine-grained results of OOD test dataset for experiments in Table 4. To further evaluate the influence of both modulating factors introduced in Eq (4) on the performance on every OOD test

Table 7: Fine-grained results of OOD detection performance for experiments in Table 4. All methods are trained on the same backbone CLIP-ViT-B/16. Bold numbers are superior results. ↑ indicates larger values are better, and ↓ indicates smaller values are better.

ϕ	ψ	OOD dataset: iNaturalist FPR95↓	iNaturalist AUROC↑	SUN FPR95↓	SUN AUROC↑	Places365 FPR95↓	Places365 AUROC↑	Textures FPR95↓	Textures AUROC↑	Average FPR95↓	Average AUROC↑
						1-shot					
✗	✗	28.81	94.05	25.76	94.51	33.68	91.59	51.53	86.85	34.94	91.75
✓	✗	21.94	95.14	22.46	95.08	30.54	92.04	49.61	87.14	31.14	**92.35**
✗	✓	19.26	95.72	23.70	94.44	33.56	91.04	51.10	85.77	31.90	91.74
✓	✓	19.16	95.70	23.52	94.58	32.81	91.23	48.87	86.66	**31.09**	92.04
						16-shot					
✗	✗	17.58	96.30	22.82	95.20	32.21	92.03	45.27	88.86	29.47	93.10
✓	✗	18.14	94.50	22.42	94.04	31.90	91.18	44.72	90.09	29.30	92.66
✗	✓	15.08	96.92	21.42	95.16	30.60	92.07	48.64	86.35	28.94	92.62
✓	✓	13.94	95.86	20.55	95.33	29.86	92.24	41.51	89.06	**26.47**	**93.37**

Table 8: Different implementations of modulation functions.

Modulation function	ϕ	ψ
power	$(1-p(y\|\boldsymbol{x};\boldsymbol{\omega}))^{\alpha}$	$p(y\|\boldsymbol{x};\boldsymbol{\omega})^{\alpha}$
logarithmic	$1-\frac{\log\left(p(y\|\boldsymbol{x};\boldsymbol{\omega})+1\right)}{\log 2}$	$\frac{\log\left(p(y\|\boldsymbol{x};\boldsymbol{\omega})+1\right)}{\log 2}$
trigonometric	$\cos\left(\frac{\pi}{2}p(y\|\boldsymbol{x};\boldsymbol{\omega})\right)$	$\sin\left(\frac{\pi}{2}p(y\|\boldsymbol{x};\boldsymbol{\omega})\right)$

test, we present the fine-grained OOD detection results in Table 7. We observe from the results that although each of the modulation factors can individually contribute to redirect model training, their combination maximizes the improvement on OOD detection, which achieves the best results on most OOD test sets and shot numbers.

Experiments on other instantiations of modulation functions. To explore the generality of our proposed learning framework, we consider other instantiations of modulation functions. Specifically, we consider the power, logarithmic, and trigonometric functions, which are formulated in Table 8. For the experiment on the power function, we set $\lambda = 0.25$, $\alpha = 0.5$ for 1-shot and $\lambda = 0.25$, $\alpha = 4$ for 16-shot. The experiment results, as shown in Table 9, indicate that all the considered choices of modulation function can achieve significantly better OOD detection performance than LoCoOp, which further empirically proves the effectiveness of the learning framework of SCT.

Experiments on conventional CIFAR benchmarks. We conduct the experiments on conventional CIFAR benchmarks, following previous works [Liu et al., 2020]. We adopt CIFAR-10, CIFAR-100[Krizhevsky, 2009] as the ID datasets and use Textures [Cimpoi et al., 2014], Places365 [Zhou et al., 2016], iSUN [Xu et al., 2015], LSUN_Crop [Yu et al., 2015], and LSUN_Resize [Yu et al., 2015] as the OOD test datasets. We summarize the comparison results averaged across all OOD test datasets in Table 10, which confirm that SCT achieves superior performance to LoCoOp.

Experiments on Computational Cost SCT doesn't incur any extra computational cost to the LoCoOp due to its simple design. Technically, SCT introduces modulating factors respectively on the two components of the original learning objective. The modulating factors (instantiated as $1-p(y|x)$ for ϕ and $p(y|x)$ for ψ in Equation (4) in the submission) only involves the computation of the prediction probability of ground truth classes $p(y|x)$, which can be repeatedly used after the original forward pass of CLIP models. What's more, the operation of local features involved in LoCoOp and SCT are also relatively low-cost in terms of computation. The local are generated from the forward pass of the vision encoders of CLIP, which doesn't bring additional computational cost compared to regular training. Regarding the extraction of OOD features, we compute the similarity between every local feature and text feature of all the ID classes and we identify regions that do not include their

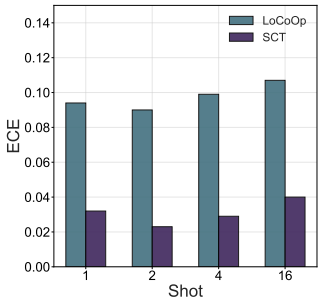


Figure 4: The comparison of calibration measured by ECE between SCT and LoCoOp trained with 1, 2, 4, 16 shots data. The evaluation is performed on the original validation set of ImageNet-1k.

Table 9: OOD detection performance of different instantiations of modulation functions. All methods are trained on the same backbone CLIP-ViT-B/16. Bold numbers are superior results. ↑ indicates larger values are better, and ↓ indicates smaller values are better. SCT-L denotes SCT with the linear function, SCT-Pow denotes SCT with the power function, SCT-Log denotes SCT with the logarithmic function and SCT-Tri denotes SCT with the trigonometric function as the modulation function.

Method	**OOD dataset**									
	iNaturalist		**SUN**		**Places365**		**Textures**		**Average**	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
1-shot										
LoCoOp	28.81	94.05	25.76	94.51	33.68	91.59	51.53	86.85	34.94	91.75
SCT-L	19.16	95.70	23.52	94.58	32.81	91.23	48.87	86.66	**31.09**	92.04
SCT-Pow	18.91	95.84	25.06	94.61	33.36	91.68	49.13	86.68	31.62	**92.20**
SCT-Log	18.32	95.98	26.00	94.62	33.42	91.39	51.17	85.97	32.23	91.99
SCT-Tri	18.46	95.84	24.32	94.98	33.23	91.69	54.10	86.15	32.53	92.17
16-shot										
LoCoOp	17.58	96.30	22.82	95.20	32.21	92.03	45.27	88.86	29.47	93.10
SCT-L	13.94	95.86	20.55	95.33	29.86	92.24	41.51	89.06	**26.47**	**93.37**
SCT-Pow	14.07	96.70	20.74	94.75	30.11	91.96	43.49	87.90	27.10	92.83
SCT-Log	13.11	97.01	20.56	95.50	29.03	92.66	45.55	87.61	27.06	93.20
SCT-Tri	14.88	96.81	20.30	95.44	29.33	92.50	44.84	87.90	27.34	93.16

Table 10: Experiments on CIFAR benchmark with 16-shot data.

Method	$\mathcal{D}_{\text{in}}$					
	CIFAR-10			**CIFAR-100**		
	FPR95↓	AUROC↑	ID-ACC↑	FPR95↓	AUROC↑	ID-ACC↑
MCM	16.36	95.68	90.10	74.92	77.62	68.40
LoCoOp	15.30	95.37	93.00	64.66	82.45	72.20
SCT	12.89	96.10	93.10	60.74	84.91	72.30

ground truth class in the top-K predicted classes as ID-irrelevant regions. Empirically, we evaluate the time and memory consumption of SCT compared with other baselines in Table 11 and the results show that SCT is relatively compute-efficient. The evaluation is conducted on a single GTX-3090 GPU with a batch size as 32.

Table 11: Time and memory cost of different methods on the ImageNet benchmark.

Method	Time for one iteration (s)	GPU Memory (MiB)	FPR95	AUROC	ID-ACC
CoOp	0.70	21140	35.09	91.99	**71.93**
LoCoOp	0.96	23036	29.47	93.10	71.43
SCT	0.96	23036	**26.47**	**93.37**	71.77

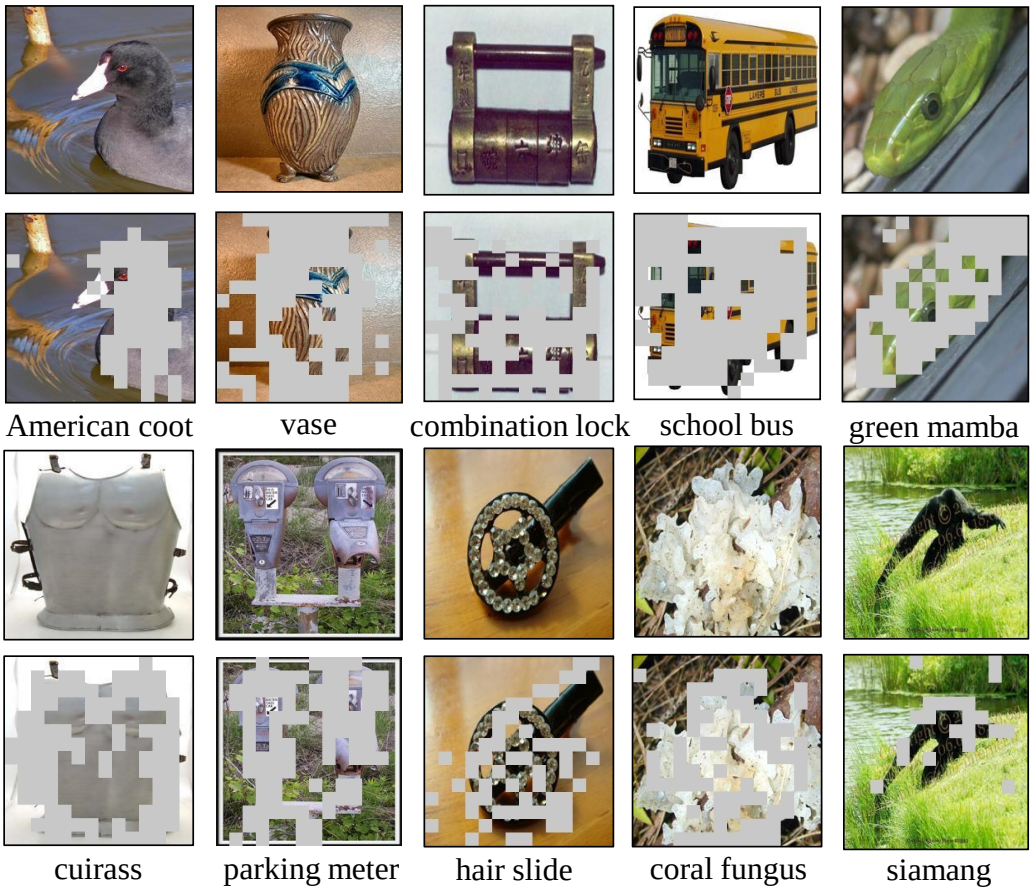


Figure 5: Examples of the invalid OOD features extracted by CLIP. The odd-numbered rows show the original images from ImageNet-1k and the even-numbered rows show the extracted ID-irrelevant context from the corresponding images. The ground-truth labels are annotated below every even-numbered row. Although CLIP can mask out some ID-related regions (shown as the gray patches of images), large portions of the extracted OOD features (shown as the colored patches of images) obviously belong to ID features.

Discussions about current advanced post-hoc methods and our method. prompt tuning based method can leverage the generalization ability of VLMs to better fit the domains of the downstream tasks with relatively low computational cost. Post-hoc methods need to be built on a well-trained model, the capacity of which greatly affects the OOD detection performance. What's more, post-hoc methods and prompt tuning based methods are compatible with each other, further boosting the OOD detection performance. We conduct experiments on the compatibility of an advanced post-doc method, NegLabel [Jiang et al., 2024], with SCT in the Table 12, and the results show that SCT can be combined with post-hoc methods for better OOD detection.

In addition, recent advanced post-hoc methods, such ASH [Djurisic et al., 2022] and ISH [Xu et al., 2024], improve OOD detection from the perspective of activation scale. However, VLMs like CLIP compute prediction probability based on the cosine similarity between image and text features. The computation of cosine similarity involves the normalization of features, which naturally eliminates the effect of the activation scale. Therefore, ASH and ISH can't apply to VLM models.

Improvement on confidence calibration. We use Expected Calibration Error (ECE) [Naeini et al., 2015] to measure the improvement of SCT on confidence calibration over LoCoOp, with lower values indicating better calibration. ECE is calculated by dividing predictions on samples into M equally-spaced bins by confidence scores, then computing the mean average of the difference between each bin's accuracy and confidence. It can be formulated as $ECE = \sum_{m=1}^{M} \frac{|Bin_m|}{n} |acc(Bin_m) -$

Table 12: Experiments on compatibility of Neg-Label and SCT on 16-shot data.

threshold	OOD dataset									
	iNaturalist		SUN		Places365		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
NegLabel	1.91	99.49	20.53	95.49	35.59	91.64	43.56	90.22	25.40	94.21
NegLabel+SCT	2.03	99.51	17.42	95.92	31.43	92.46	38.46	91.23	22.34	94.78

$conf(Bin_m)|$, where n denotes the number of samples. As shown in Figure 4, SCT can consistently enhance the calibration of VLMs across all few-shot settings.

A.3.3 More empirical demonstrations and illustrations of our research problem.

In this section, we present more empirical evidence and illustrations of the research problem that inspires our SCT in Fig 5. As illustrated, portions of the extracted local features from ID data are invalid OOD features, thus degrading the performance of OOD detection.

A.4 Broader impact

OOD detection is crucial for deploying reliable deep learning systems in real-world scenarios [Nguyen et al., 2015, Hendrycks et al., 2022]. This significance is particularly evident in safety-critical domains such as finance or medical intelligence, where a trustworthy model must accurately distinguish between samples belonging to distinct label spaces (e.g., animals) rather than simply providing predictions based on known classes (e.g., financial products or medical conditions). Our research focuses on a general and practical challenge concerning prompt tuning methods for improving OOD detection efficacy. We specifically focus on the issue of spurious OOD features extracted from ID samples. It is important for effective OOD detection to gain better empirical performance through regularizing prompt tuning.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and the introduction accurately reflect the paper's contributions and the scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations of the work in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setups are provided in Section 4.1 to ensure the result reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code is publicly available at: https://github.com/tmlr-group/SCT.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the detailed experimental setups in Section 4.1 and more discussion in Appendix A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We demonstrate the experimental statistical significance by reporting the mean and std value based on multiple trials of the experiments in Tables 1, 2, and other results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details of the used experiment compute resources in the reproducibility statement in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive and negative societal impacts of the work in Appendix A.4.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of the datasets and referred codes are properly mentioned in the references and their credits are properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.