Prompt Optimization Meets Subspace Representation Learning for Few-shot Out-of-Distribution Detection

Faizul Rakib Sayem
faizulrakib.sayem@ucf.edu
University of Central Florida
Orlando, Florida, USA

Shahana Ibrahim
Shahana.ibrahim@ucf.edu
University of Central Florida
Orlando, Florida, USA

Abstract

The reliability of artificial intelligence (AI) systems in open-world depends heavily on their ability to flag out-of-distribution (OOD) inputs, which are unseen during the training phase. Recent advances in large-scale vision-language models (VLMs) have enabled promising few-shot OOD detection frameworks using only a handful of in-distribution (ID) samples. However, existing prompt learning-based OOD methods rely solely on softmax probabilities, overlooking the rich discriminative potential of the feature embeddings learned by VLMs trained on millions of samples. To address this limitation, we propose a novel context optimization (CoOp)-based framework that integrates subspace representation learning with prompt tuning. Our approach improves ID-OOD separability by projecting the ID features into a subspace spanned by prompt vectors, while projecting ID-irrelevant features into an orthogonal null space. To train such OOD detection framework, we design an easy-to-handle end-to-end learning criterion that ensures strong OOD detection performance as well as high ID classification accuracy. Experiments on real-world datasets showcase the effectiveness of our approach.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Computer vision**.

Keywords

Prompt learning, Out-of-distribution detection, Vision-language models, Subspace representation learning

ACM Reference Format:
Faizul Rakib Sayem and Shahana Ibrahim. 2025. Prompt Optimization Meets Subspace Representation Learning for Few-shot Out-of-Distribution Detection. In *Proceedings of ACM SIGKDD Knowledge Discovery and Data Mining Workshop on Prompt Optimization (KDD 2025).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Deep learning models often exhibit overconfidence when exposed to inputs from unseen, out-of-distribution (OOD) categories [6]. This overconfidence can lead to critical failures in open-world and safety-sensitive applications such as autonomous driving [5] and medical diagnostics [21]. Traditional OOD detection approaches [9–11, 13] typically rely on designing scoring functions or incorporating auxiliary outlier datasets during training. While such methods have demonstrated promise in controlled settings, they often fail to generalize in dynamic, real-world environments where the nature of the OOD data is unpredictable.

Recently, large-scale vision-language models (VLMs) such as contrastive language-image pretraining (CLIP) [20] have shown strong zero-shot performance on downstream tasks by aligning visual and textual modalities in a shared embedding space. This opens a new direction for OOD detection, particularly in low-resource or few-shot settings. However, CLIP's zero-shot approach depends heavily on manually crafted prompts, where even slight variations (e.g., "*a flower*" vs. "*a type of a flower*") can significantly impact performance. To reduce this sensitivity, a class of prompt tuning methods called *context optimization* has been introduced. For example, CoOp [31] and CoCoOp [30] replace hand-crafted textual embeddings with learnable context vectors that are optimized to enhance alignment between in-distribution (ID) image features and class text embeddings, leading to improved the classification accuracy.

However, context optimization methods face a significant limitation in their direct applicability towards OOD detection tasks. By focusing on bringing ID image features closer to their text embeddings, they may inadvertently include background or semantically irrelevant regions, corrupting the learned representations. LoCoOp [16] addresses this limitation by leveraging CLIP's spatially-aware local features. It identifies ID-irrelevant regions—those where the true class is not among the top predictions—and treats them as proxy OOD features. By applying an entropy-maximization strategy to the predictions associated with ID-irrelevant features, this approach enhances the separation between ID and OOD samples without relying on any specific OOD data. A related method was proposed in [27], where adaptive weighting is incorporated into the LoCoOp optimization framework to dynamically balance ID- and OOD-specific loss terms based on the model's prediction confidence.

Our contributions. LoCoOp and its variants [16, 27] have made promising progress in prompt learning-based OOD detection. However, they rely solely on softmax probabilities computed from the cosine similarities between image and text embeddings. While softmax outputs are known to be overly confident on OOD inputs, feature embeddings typically preserve more calibrated and discriminative information [3, 7, 22, 23]. Motivated by this, our approach aims to improve ID-OOD separability by leveraging the rich discriminative structure of the feature embeddings. To achieve this,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD 2025, Toronto, ON, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

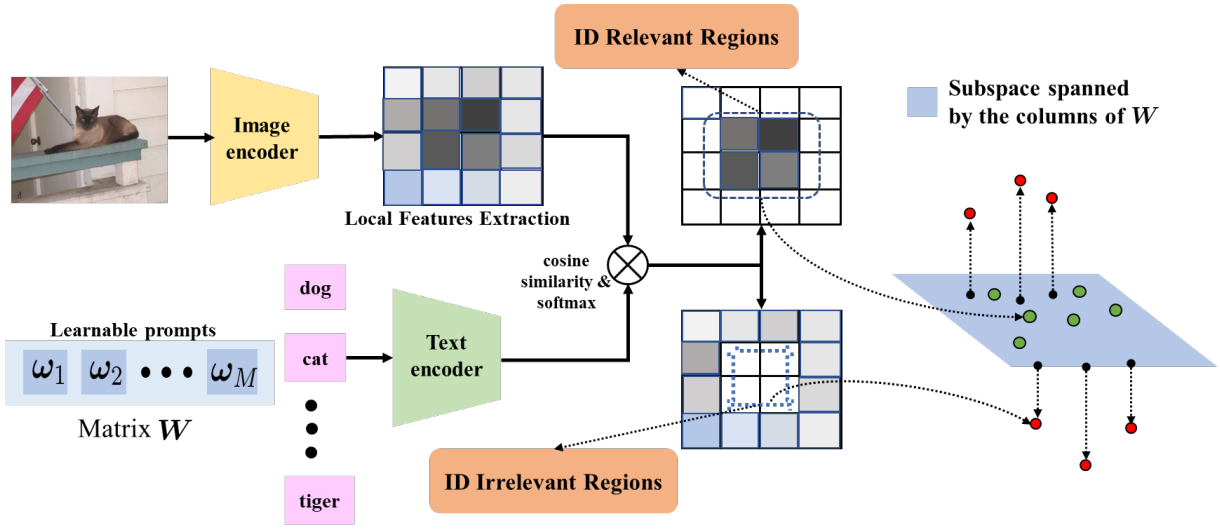

Figure 1: The proposed Subspace learning-based Context Optimization (SubCoOp) framework for prompt learning-based OOD detection.

we exploit the *subspace geometry* in context optimization framework: ID features are projected into a subspace spanned by the prompt vectors, while ID-irrelevant features are projected into the orthogonal null space. We design an end-to-end learning objective that is easy to implement and enables improved OOD detection performance without compromising ID classification accuracy. Experiments on large-scale real-world datasets such as ImageNet-1K [2] demonstrate the effectiveness of our method.

2 Proposed Method

Our goal is to design a prompt learning-based OOD detection framework that works under few-shot settings to effectively detect the OOD samples from ID samples.

Problem Statement. We consider an ID dataset $\mathcal{D}^{\text{in}} = \{(\boldsymbol{x}^{\text{in}}, y^{\text{in}})\}$, where $\boldsymbol{x}^{\text{in}} \in \mathbb{R}^L$ denotes the input features of an image and $y^{\text{in}} \in \mathcal{Y}^{\text{in}} := \{1, \ldots, K\}$ is its corresponding class label. AI models are typically trained under the closed-world assumption, where test samples are expected to come from the same distribution as the ID data. In practice, however, models frequently encounter OOD samples—data that deviates from the training distribution [8]. In classification settings, there may occur a semantic-shift such that test samples may belong to an *unknown* label space $\mathcal{Y}^{\text{out}}$, where $\mathcal{Y}^{\text{in}} \cap \mathcal{Y}^{\text{out}} = \emptyset$. The objective of OOD detection is to build a classifier that, given a test sample $\boldsymbol{x}$, predicts whether it belongs to an ID class or not, thereby preventing models from assigning high-confidence predictions to OOD samples. OOD detection can be framed as a binary classification problem. Formally, this is achieved through a detection function $\boldsymbol{d}_\eta : \mathbb{R}^L \to \{\text{ID}, \text{OOD}\}$ such that

$$\boldsymbol{d}_\eta(\boldsymbol{x}) = \begin{cases} \text{ID} & s(\boldsymbol{x}) \geq \eta \\ \text{OOD} & s(\boldsymbol{x}) < \eta, \end{cases} \tag{1}$$

where $s(\boldsymbol{x})$ is a scoring function associated with the input feature $\boldsymbol{x}$ and η is the threshold. In this work, we focus on a *few-shot* setting, using only 16 annotated examples per class from $\mathcal{D}^{\text{in}}$ for training. The model's effectiveness is evaluated on a test set which consists of both ID and OOD samples.

Prompt Learning with Positive Class Labels. Context optimization (CoOp) [31] leverages pre-trained VLMs, such as CLIP [20], for open-vocabulary visual recognition tasks. While CLIP typically uses static, hand-crafted prompts, CoOp learns a set of positive prompt vectors in a data-driven manner. These vectors are optimized as part of the model parameters during training, enabling few-shot learning for the downstream task.

Let $\boldsymbol{x}^{\text{in}} \in \mathbb{R}^L$ be an ID input image. The image is processed by the visual encoder $\boldsymbol{f} : \mathbb{R}^L \to \mathbb{R}^D$ of CLIP to extract the visual feature vector $\boldsymbol{f}^{\text{in}} = \boldsymbol{f}(\boldsymbol{x}^{\text{in}})$. The textual prompt is composed as $\boldsymbol{t}_k = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_M, \boldsymbol{c}_k\}$, where each $\boldsymbol{\omega}_m \in \mathbb{R}^D$ is a learnable context vector, $c_k \in \mathbb{R}^D$ is the class name embedding of the image, for each class $k \in [K]$, and M is the number of positive prompt vectors. The textual encoder $\boldsymbol{g}$ processes the prompt $\boldsymbol{t}_k$ to yield the textual feature $\boldsymbol{g}_k = \boldsymbol{g}(\boldsymbol{t}_k)$. With these notations, we can define the probability of the input being classified as class k as follows:

$$p(y = k \mid \boldsymbol{x}^{\text{in}}) = \frac{\exp\left(\text{sim}(\boldsymbol{f}^{\text{in}}, \boldsymbol{g}_k)/\tau\right)}{\sum_{k'=1}^{K} \exp\left(\text{sim}(\boldsymbol{f}^{\text{in}}, \boldsymbol{g}_{k'})/\tau\right)}, \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and $\tau > 0$ is a temperature parameter. In the final objective function, we will use the cross-entropy loss $\mathcal{L}_{\text{CE}}$ to match between the probabilities in (2) and the true label y^{in}.

Prompt Vector-induced Subspace. In order to utilize the feature embedding space to efficiently enable ID-OOD separability, we consider the matrix formed by the prompt vectors $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M$. Let $\boldsymbol{W} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_M]$. Our idea is to project the visual feature embeddings of the ID data to a low-dimensional subspace spanned by the columns of $\boldsymbol{W} \in \mathbb{R}^{D \times M}$, denoted as $\mathcal{R}(\boldsymbol{W})$. At the same time, if the embeddings corresponding to the OOD data lie outside

this subspace, such criterion can enhance the ID-OOD separability in the feature space. However, achieving this in practice poses several challenges. First, we operate in a few-shot setting, where only a limited number of labeled ID samples are available, making it difficult to reliably estimate a representative subspace. Second, without access to OOD samples during training, the model struggles to learn a well-defined subspace boundary due to the absence of explicit supervision.

OOD Local Features Extraction. Recently, LoCoOp [16] introduced a novel perspective to prompt optimization-based OOD detection by extracting local features that serve as proxy OOD signals, thereby preventing the model from assigning high ID confidence scores to OOD features. To detect local features not corresponding to ID classes (i.e., ID-irrelevant features), the method in [16] examine a set of spatial indices from the feature map: $\mathcal{I} = \{0, 1, 2, \ldots, H \times W - 1\}$, where H and W are the height and width of the feature map, respectively. Following a strategy inspired by semantic segmentation [20], the class probabilities associated to each region $i \in \mathcal{I}$ can be computed based on the similarity between local visual features and text embeddings:

$$p_i(y = k \mid \boldsymbol{x}^{\text{in}}) = \frac{\exp\left(\text{sim}(\boldsymbol{f}_i^{\text{in}}, \boldsymbol{g}_k)/\tau\right)}{\sum_{k'=1}^{K} \exp\left(\text{sim}(\boldsymbol{f}_i^{\text{in}}, \boldsymbol{g}_{k'})/\tau\right)}, \tag{3}$$

where $\boldsymbol{f}_i^{\text{in}} \in \mathbb{R}^D$ denotes the feature extracted from the ith local region of the image and $\boldsymbol{g}_k$ corresponds to the text prompt embedding for the kth class as defined in (2).

For any region i, if it corresponds to an ID class, its ground-truth label y^{in} is expected to appear among the top-C predicted classes. Conversely, if the region is unrelated to any ID class (e.g., background noise), the true class is unlikely to rank within the top-C, due to the lack of strong semantic alignment. Leveraging this observation, one can define an index set $\mathcal{J}$ to identify such ID-irrelevant regions:

$$\mathcal{J} = \left\{ i \in \mathcal{I} : \text{rank}\left(p_i(y = y^{\text{in}} \mid \mathbf{x}^{\text{in}})\right) > C \right\}. \tag{4}$$

Here, $\text{rank}(p_i(y = y^{\text{in}} \mid \boldsymbol{x}^{\text{in}}))$ denotes the rank of the true class y^{in} among the predicted scores over all ID classes and C is a hyperparameter or can be fixed based on prior knowledge about the number of fine-grained classes or semantic relationships in the dataset.

Subspace-based Regularization Loss. Based on the extracted local features for both ID and OOD data, we design regularization losses to improve the separability between them. Specifically, the local feature vectors corresponding to ID data are projected onto an M-dimensional subspace spanned by the column vectors of $\boldsymbol{W} \in \mathbb{R}^{D \times M}$, denoted by $\mathcal{R}(\boldsymbol{W})$. On the other hand, the features from ID-irrelevant or OOD regions are projected to lie in the null space $\mathcal{N}(\boldsymbol{W})$ orthogonal to $\mathcal{R}(\boldsymbol{W})$, defined as $\mathcal{N}(\boldsymbol{W}) = \left\{\boldsymbol{f} \in \mathbb{R}^D : \boldsymbol{W}^\top \boldsymbol{f} = \boldsymbol{0}\right\}$, which has dimension $D - M$. It is important to keep $M < D$, since when $M = D$, the null space becomes trivial (containing only the zero vector), thus limiting our ability to separate ID and OOD samples effectively. This condition is typically satisfied in practice, as the number of prompt vectors M is usually small (e.g., $M \approx 16$ as suggested in [16, 31]), whereas the dimensionality of CLIP embeddings is relatively large (e.g., $D = 512$). The proposed regularization losses for the ID and OOD regions are defined as follows:

$$\mathcal{L}_{\text{Sub-ID}} = \sum_{j \in \mathcal{J}'} \sum_{n=1}^{N} \frac{\left\|\text{Proj}_{\boldsymbol{W}^\perp}\left(\boldsymbol{f}_{n,j}^{\text{in}}\right)\right\|_2}{\left\|\boldsymbol{f}_{n,j}^{\text{in}}\right\|_2}, \tag{5a}$$

$$\mathcal{L}_{\text{Sub-OOD}} = \sum_{j \in \mathcal{J}} \sum_{n=1}^{N} \frac{\left\|\text{Proj}_{\boldsymbol{W}}\left(\boldsymbol{f}_{n,j}^{\text{in}}\right)\right\|_2}{\left\|\boldsymbol{f}_{n,j}^{\text{in}}\right\|_2}, \tag{5b}$$

where N denotes the total number of training samples, $\boldsymbol{f}_{n,j}^{\text{in}}$ denotes the ith local region feature for the nth image, $\mathcal{J}'$ is the complement of the set $\mathcal{J}$, i.e., $\mathcal{J}' = \mathcal{I} \setminus \mathcal{J} = \{i \in \mathcal{I} \mid i \notin \mathcal{J}\}$, and the projections $\text{Proj}_{\boldsymbol{W}^\perp}$ and $\text{Proj}_{\boldsymbol{W}}$ are given by:

$$\text{Proj}_{\boldsymbol{W}^\perp}(\boldsymbol{f}) = \left(\boldsymbol{I}_M - \boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top\right) \boldsymbol{f},$$

$$\text{Proj}_{\boldsymbol{W}}(\boldsymbol{f}) = \left(\boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top\right) \boldsymbol{f}.$$

Here, the loss term $\mathcal{L}_{\text{Sub-ID}}$ encourages ID features to lie within the column space $\mathcal{R}(\boldsymbol{W})$ by minimizing their projected components in the orthogonal complement, $\mathcal{N}(\boldsymbol{W})$. Conversely, the loss term $\mathcal{L}_{\text{Sub-OOD}}$ promotes OOD features to lie in $\mathcal{N}(\boldsymbol{W})$ by suppressing their projections onto $\mathcal{R}(\boldsymbol{W})$.

To regularize ID-irrelevant feature regions, we further apply an entropy-based loss over the set $\mathcal{J}$. For each ID-irrelevant region $j \in \mathcal{J}$, the goal is to encourage the model to exhibit low confidence in its predictions based on the local image feature $\boldsymbol{f}_j^{\text{in}}$. To achieve this, inspired by entropy maximization techniques [16, 28], we enforce high entropy on the predicted class distribution $\boldsymbol{p}_j(\boldsymbol{x}_n^{\text{in}}),\ j \in \mathcal{J}$, which is a K-dimensional probability vector where each entry represents $p_j(y = k \mid \boldsymbol{x}_n^{\text{in}})$, as defined in (3). The entropy-based regularization loss is defined as:

$$\mathcal{L}_{\text{Ent-OOD}} = -\sum_{n=1}^{N} \sum_{j \in \mathcal{J}} H\big(\boldsymbol{p}_j(\boldsymbol{x}_n^{\text{in}})\big), \tag{6}$$

where $H(\cdot)$ denotes the entropy function.

Overall Loss Function. The overall training loss combines the cross-entropy loss with the above discussed regularization losses as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{EReg}} + \lambda_2 \mathcal{L}_{\text{SReg_ID}} + \lambda_3 \mathcal{L}_{\text{SReg_OOD}}, \tag{7}$$

where $\mathcal{L}_{\text{CE}}$ is discussed after (2), other regularization terms are defined in (5) and (6), and $\lambda_1, \lambda_2, \lambda_3 > 0$ are the regularization parameters. We name our approach as **Sub**space learning-based **Co**ntext **Op**timization (SubCoOp).

3 Experiments

Dataset. For our experiments, we utilize ImageNet-1k dataset [2] as ID data. For OOD data, we use a number of commonly used benchmark datasets such as iNaturalist [24], SUN [26], Places [29], and Texture [25]. For the few-shot training, we use 16 images per ID class, and evaluate the model using the whole OOD datasets and the test ID dataset.

Implementation Details. We adopt the CLIP ViT-B/16 model [4] as the backbone of the visual encoder for the pretrained network. For ID-irrelevant feature extraction, we set the rank threshold parameter to $C = 200$. In addition, we fix $M = 16$, $\lambda_1 = 0.25$, $\lambda_2 = 2$,

Table 1: Comparison of FPR95 and AUROC scores (%) on various OOD datasets with ID dataset ImageNet-1k.

Method	iNaturalist		SUN		Places365		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-shot methods										
MCM	30.94	94.61	37.67	92.56	44.76	89.76	57.91	86.10	42.82	90.76
GL-MCM	15.18	96.71	30.42	93.09	38.85	89.90	57.93	83.63	35.47	90.83
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 (16-shot)										
LSN	$46.40^{\pm 1.76}$	$91.91^{\pm 2.73}$	$31.86^{\pm 1.56}$	$93.21^{\pm 1.32}$	$40.61^{\pm 0.65}$	$90.05^{\pm 1.53}$	$47.21^{\pm 0.88}$	$88.98^{\pm 0.97}$	$41.52^{\pm 1.21}$	$91.04^{\pm 1.64}$
NegPrompt	$38.11^{\pm 1.15}$	$90.22^{\pm 0.78}$	$31.44^{\pm 0.29}$	$92.59^{\pm 0.18}$	$36.15^{\pm 2.05}$	$90.97^{\pm 0.78}$	$44.64^{\pm 1.34}$	$87.49^{\pm 0.52}$	$37.59^{\pm 1.21}$	$90.32^{\pm 0.57}$
CoOp	$26.72^{\pm 2.09}$	$94.53^{\pm 0.36}$	$36.96^{\pm 0.87}$	$92.34^{\pm 0.15}$	$45.01^{\pm 1.45}$	$89.43^{\pm 0.15}$	$40.38^{\pm 1.45}$	$90.95^{\pm 0.18}$	$37.27^{\pm 1.47}$	$91.81^{\pm 0.21}$
LoCoOp	$18.70^{\pm 2.12}$	$96.09^{\pm 0.38}$	$22.83^{\pm 0.98}$	$95.12^{\pm 0.07}$	$34.78^{\pm 3.47}$	$91.52^{\pm 0.63}$	$43.75^{\pm 0.22}$	$89.81^{\pm 0.33}$	$30.02^{\pm 1.70}$	$93.14^{\pm 0.35}$
SubCoOp	$14.33^{\pm 0.76}$	$96.99^{\pm 0.08}$	$22.14^{\pm 1.96}$	$95.10^{\pm 0.44}$	$32.04^{\pm 2.82}$	$92.07^{\pm 0.61}$	$42.35^{\pm 3.04}$	$89.87^{\pm 0.53}$	$27.72^{\pm 2.15}$	$93.51^{\pm 0.42}$

and $\lambda_3 = 5$. We employ the SGD optimizer with a learning rate of 0.002, a batch size of 32, and train the model for 50 epochs. We use Nvidia 3090 Ti GPU for all the experiments.

OOD Detection Score: While testing, we adopt the global-local maximum concept matching (GL-MCM) score [18] for OOD detection (i.e., the score function $s(\boldsymbol{x})$ as employed in (1)). This metric integrates the maximum softmax probability scores from both whole image feature and local image features and is defined as follows:

$$s_{\text{GL-MCM}}(\boldsymbol{x}) = \max_k \frac{\exp\left(\text{sim}(\boldsymbol{f}, \boldsymbol{g}_k)/\tau\right)}{\sum_{k'=1}^{K} \exp\left(\text{sim}(\boldsymbol{f}, \boldsymbol{g}_{k'})/\tau\right)} + \max_{k,i} \frac{\exp\left(\text{sim}(\boldsymbol{f}_i, \boldsymbol{g}_k)/\tau\right)}{\sum_{k'=1}^{K} \exp\left(\text{sim}(\boldsymbol{f}_i, \boldsymbol{g}_{k'})/\tau\right)} \tag{8}$$

where $\boldsymbol{f}$ is the vision encoder output for the test image $\boldsymbol{x}$ and $\boldsymbol{f}_i$'s are its features corresponding to the local regions.

Evaluation Metrics. We evaluate the OOD detection performance using the following metrics: *(i)* false positive rate (FPR) at %95 refers to the false positive rate of OOD samples when the true positive rate of ID samples is at 95%; *(ii)* area under the receiver operating characteristic curve (AUROC), measures the model's ability to distinguish between ID and OOD samples; and *(iii)* classification accuracy on ID data.

Baselines. To evaluate our proposed method, we compare it with zero-shot learning approaches, post-hoc CLIP based methods, and few-shot prompt learning methods. For the zero shot baseline, we use the state-of-the-art MCM [15] and GL-MCM [17] methods. For the post-hoc methods, we adopt MSP [8], ReAcT [22], ODIN [12], Maxlogit [1], and Energy Score [14] as baselines. These methods leverage CLIP's pretrained representations and combining them with simple post-processing techniques/scores for OOD detection. For few-shot prompt tuning based baselines, we consider CoOp [31], LSN [19], NegPrompt [12], and LoCoOp [16]. CoOp, LoCoOp, and our approach SubCoOp are based on learning a set of positive prompts. On the other hand, NegPrompt and LSN each learn a set of negative prompts per ID class in addition to the positive prompt vectors.

Results. We report the OOD detection performance of our method and the baselines, as shown in Table 1, averaged over three random trials. Prompt tuning–based methods outperform other line of approaches as they encourage the model to align visual features with more discriminative and dynamically learned text prompts. Our subspace-based prompt tuning strategy, SubCoOp, consistently outperforms competing methods, achieving superior results with a notable margin. On the ImageNet-1K dataset, SubCoOp attains the best OOD detection performance, with an average FPR95 of 27.72%, AUROC of 93.51%, while maintaining high ID classification accuracy (see Table **??** in the supplementary material). Similarly, strong OOD performance is observed on the ImageNet-100 dataset, as reported in Table **??** of the supplementary. These results demonstrate that when combined with proxy OOD feature extraction, our subspace regularization significantly enhances the separability between ID and OOD samples.

4 Conclusion

In this work, we propose a novel approach that integrates subspace representation learning with prompt optimization for few-shot OOD detection using VLMs. Our method induces a distinctive geometry in the feature embedding space by projecting ID features onto a subspace spanned by learnable prompt vectors, while pushing ID-irrelevant features toward the orthogonal null space. Experiments on several OOD benchmarks based on ImageNet-1K demonstrate that our prompt tuning framework, SubCoOp, consistently outperforms state-of-the-art methods in OOD detection, without sacrificing ID classification accuracy. As future work, we plan to develop a subspace-projection-inspired OOD detection score, which can serve as either a surrogate or a complement to the current state-of-the-art GL-MCM score.

References

[1] Steven Basart, Mazeika Mantas, Mostajabi Mohammadreza, Steinhardt Jacob, and Song Dawn. 2022. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848

[3] Xin Dong, Junfeng Guo, Ang Li, Wei-Te Ting, Cong Liu, and HT Kung. 2022. Neural mean discrepancy for efficient out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19217–19227.

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

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 3354–3361.

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

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*. PMLR, 8759–8773. https://proceedings.mlr.press/v162/hendrycks22a.html

[8] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[9] Rui Huang, Andrew Geng, and Yixuan Li. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems* 34 (2021), 677–689.

[10] Rui Huang and Yixuan Li. 2021. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8710–8719.

[11] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* 31 (2018).

[12] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690* (2017).

[13] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems* 33 (2020), 21464–21475.

[14] Weitang Liu, Xiaoyun Wang, John Douglas Owens, and Yixuan Li. 2020. Energy-based Out-of-distribution Detection. *ArXiv* abs/2010.03759 (2020). https://api.semanticscholar.org/CorpusID:222208700

[15] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems* 35 (2022), 35087–35102.

[16] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. 2023. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems* 36 (2023), 76298–76310.

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

[18] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. 2025. GL-MCM: Global and Local Maximum Concept Matching for Zero-Shot Out-of-Distribution Detection. *International Journal of Computer Vision* (2025), 1–11.

[19] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. 2024. Out-of-distribution detection with negative prompts. In *The twelfth international conference on learning representations*.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[21] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. 2017. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In *Information Processing in Medical Imaging*. Springer International Publishing, 146–157.

[22] Yiyou Sun, Chuan Guo, and Yixuan Li. 2021. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems* 34 (2021), 144–157.

[23] Yiyou Sun and Yixuan Li. 2022. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 691–708.

[24] Grant Van Horn. 2018. 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 2. 5.

[25] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8769–8778.

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

[27] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. 2024. Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection. In *NeurIPS*.

[28] Qing Yu, Atsushi Hashimoto, and Yoshitaka Ushiku. 2021. Divergence optimization for noisy universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2515–2524.

[29] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

[30] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16816–16825.

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.