Learning Transferable Negative Prompts for Out-of-Distribution Detection

Tianqi Li[1], Guansong Pang*[2], Xiao Bai*[1], Wenjun Miao[1], and Jin Zheng[1]

[1]School of Computer Science and Engineering, State Key Laboratory of Complex & Critical Software Environment, Jiangxi Research Institute, Beihang University, China

[2]School of Computing and Information Systems, Singapore Management University

Abstract

Existing prompt learning methods have shown certain capabilities in Out-of-Distribution (OOD) detection, but the lack of OOD images in the target dataset in their training can lead to mismatches between OOD images and In-Distribution (ID) categories, resulting in a high false positive rate. To address this issue, we introduce a novel OOD detection method, named 'NegPrompt', to learn a set of negative prompts, each representing a negative connotation of a given class label, for delineating the boundaries between ID and OOD images. It learns such negative prompts with ID data only, without any reliance on external outlier data. Further, current methods assume the availability of samples of all ID classes, rendering them ineffective in open-vocabulary learning scenarios where the inference stage can contain novel ID classes not present during training. In contrast, our learned negative prompts are transferable to novel class labels. Experiments on various ImageNet benchmarks show that NegPrompt surpasses state-of-the-art prompt-learning-based OOD detection methods and maintains a consistent lead in hard OOD detection in closed- and open-vocabulary classification scenarios. Code is available at https://github.com/mala-lab/negprompt.

1. Introduction

Since the advent of deep learning, numerous image recognition models [5, 9, 29] have relied solely on image features for classification. However, in recent years large pre-trained vision-language models (VLMs), such as CLIP [37], integrated natural language processing into computer vision, enhancing the semantic understanding capabilities of computer vision models. It has been observed that these VLMs excel in image classification, particularly in zero-shot scenarios. This is attributed to their extensive self-supervised pre-training on web-scale image-text data, which has endowed them with robust semantic transfer abilities.

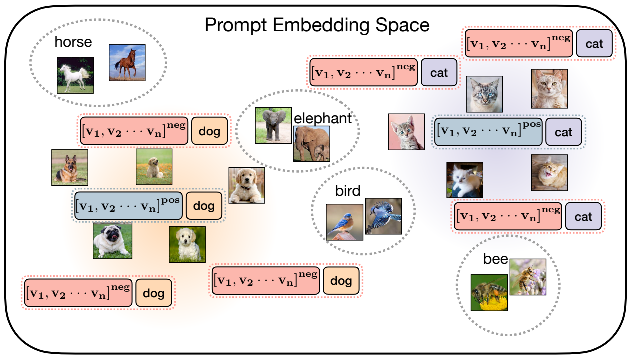
Prompt Embedding Space
horse
$[v_1, v_2 \cdots v_n]^{neg}$ cat
$[v_1, v_2 \cdots v_n]^{neg}$ cat
$[v_1, v_2 \cdots v_n]^{neg}$ dog
elephant
$[v_1, v_2 \cdots v_n]^{pos}$ cat
$[v_1, v_2 \cdots v_n]^{pos}$ dog
bird
$[v_1, v_2 \cdots v_n]^{neg}$ cat
bee
$[v_1, v_2 \cdots v_n]^{neg}$ dog
$[v_1, v_2 \cdots v_n]^{neg}$ dog

Figure 1. Illustration of the key intuition of NegPrompt. For each ID class, NegPrompt trains a small set of learnable prompts that have negative semantics to the learned positive prompt of the given class. As a result, OOD samples exhibit higher similarity to the negative prompts than the positive prompts.

Despite the strong zero-shot classification capabilities of VLMs, numerous research efforts are put to unleash their potential, *e.g.*, by investigating whether models like CLIP can achieve enhanced performance with training on downstream target datasets. This has led to the development of various techniques to fine-tune CLIP [8, 18], among which prompt learning has sparked widespread interest. Prompt learning (or prompt tuning) [19, 40, 51], a methodology originating from natural language processing, focuses on learning the prompt inputs into a large-scale pre-trained network, rather than learning or fine-tuning the parameters of the network. In CLIP, a common prompt is '`a photo of a [class name]`'. The aim of prompt learning, *e.g.*, in approaches like CoOp [55], is to learn a soft/differentiable context vector to replace the fixed text prompt like '`a photo of a`', thereby leveraging CLIP's powerful generalization in semantic understanding while also fine-tuning for specific target datasets [54, 55].

However, even though prompt learning enhances the target dataset perception capabilities of VLMs, it struggles with Out-Of-Distribution (OOD) detection [10, 26]. Un-

*Corresponding author: G. Pang (gspang@smu.edu.sg) and X. Bai (baixiao@buaa.edu.cn) der the OOD detection task, the test set comprises images from both the training classes – in-distribution (ID) data – and images from other unknown categories (OOD images). VLM-based classification is typically performed by first replacing the '[class name]' with the class label of each category in a prompt, which is then processed by the text encoder of CLIP to obtain the class embedding and assign the class label that has the highest cosine similarity with the embedding of a test image from the image encoder of CLIP. However in OOD detection tasks, the models do not have access to the class names of OOD images, thereby lacking the knowledge about OOD data.

This issue, exacerbated by the models' tendency towards overconfidence that often predicts OOD data as ID images with high confidence [34], undermines their ability to effectively detect OOD images.

There have been a number of CLIP-based methods designed specifically for VLM-driven OOD detection methods, but most of them [6, 32, 47] focused on a zero-shot setting where no training data of the target dataset is available. Due to the lack of adaptation to the target dataset, they tend to detect unusual ID images as OOD data. Among them, CLIPN [47], which trains an additional 'no' text encoder to provide prompt embeddings of not having specific classes, is the best detector, but it relies on a large-scale auxiliary dataset to train such a text encoder and it is computationally expensive.

The most related work is a very recent approach LoCoOp [33] that utilizes the training ID data to tune CLIP to capture local features of the ID classes in the prompt. It shows substantially improved performance compared to zero-shot methods [32], but it may compromise the ID classification accuracy due to an overemphasis on modeling local features. LoCoOp also lacks knowledge about OOD samples, making it difficult to differentiate boundary ID/OOD images.

In this work, we propose a novel CLIP-based OOD detection method, named **NegPrompt**. Inspired by [25], Negprompt is designed to learn a set of *negative prompts*, each representing a negative connotation of a given ID class label, to delineate the boundaries between ID and OOD images, as shown in Fig. 1. The negative prompts represent specialized ID-class-dependent concepts, guiding the models to pay attention to the characteristics that are contrary to or disjoint from the ID classes. NegPrompt aims to utilize ID training data and *positive prompts* (*i.e.*, the text prompt embeddings of ID classes) to learn such negative prompts in a way to which OOD images exhibits higher similarity than ID images.

Essentially, the learned negative prompts have similar semantics to the text prompts generated from the 'no' text encoder in CLIPN, but NegPrompt presents a fundamentally different approach: it capitalizes on the generalization ability of the CLIP model and learns the negative prompts with training ID data only, eliminating the reliance on external data and the extensive computation overhead as in CLIPN.

Furthermore, benefiting from the superior generalization ability of CLIP, our method does not require the exposure to all ID classes during training. In other words, the model learns *transferable negative prompts* by using only a small subset of the ID classes, after which we can obtain the negative prompts for the other ID classes by simply replacing the '[class name]' in the prompts with the name of those unexposed classes. This allows our model to work in open-vocabulary [7, 31] learning settings, where the models are required to classify images of novel classes that are not seen during training, in addition to a set of training base classes.

Our main contributions can be summarized as follows:

- We propose a prompt learning-based OOD detection approach NegPrompt, which is able to learn negative semantics relative to specific ID classes, thereby enhancing the VLMs' sensitivity to unknown samples. It is a lightweight method that does not require training extra encoders on external data as in related methods [47].
- NegPrompt possesses an open-vocabulary capability due to the transferability of its negative prompts. This means that with training images from just a small subset of ID classes and the class names of all IDs, we can achieve OOD detection on test data with all these ID classes. To the best of our knowledge, there have been no previous fine-tuning methods exploring such a capability.
- Extensive experiments on multiple ImageNet-based benchmarks show that NegPrompt consistently outperforms current state-of-the-art methods in both conventional and hard OOD Detection settings.

2. Related Work

2.1. Pre-trained Vision-Language Models

Understanding the semantic information of images remains a significant challenge in the field of computer vision. With the advent of Transformer [45] in computer vision tasks [5], CLIP [37] has been introduced as one of the most advanced pre-trained VLMs. Utilizing contrastive learning [21], large-scale models and datasets [38], CLIP employs image-text pairs as the training data for self-supervised learning. This approach has successfully trained the model to align visual and text signals in a latent space. Concurrently, other researchers [1, 24, 50] have also shown the remarkable generalization capabilities of CLIP and similar vision-language models [17] in various downstream tasks [7, 20, 57].

2.2. Prompt Learning

The concept of prompt learning was initially focused on automating the creation of templates/prompts for extracting knowledge from Bert [4] or GPT [36]. To bypass manual creation, prompt learning advocates the use of supervised learning to automate the prompt development. [40] proposed a gradient-based approach for identifying the best prompt, establishing the foundation of prompt learning. Later, CoOp [55] integrated prompt learning into computer vision. CoOp learns a segment of the context before it is fed into the text encoder of CLIP, thus tailoring the learned prompt to the specific target dataset. Many other prompt learning methods for different vision tasks [13, 18, 20, 41, 48, 54, 56] are subsequently introduced. However, they are not designed for OOD detection, so they struggle with dealing with the unknown OOD samples during inference.

2.3. Out-of-Distribution Detection

OOD detection is committed to identifying images in image classification tasks that belong to categories not present in the training dataset, typically originating from a different distribution. While traditional OOD detection methods often tackle the problem by either exploiting the prediction logits to define OOD scores [10, 12, 15, 26, 27] or focusing on the class-agnostic information in feature space that is not recoverable from logits [42, 46], recent methods [11, 14, 22, 28, 30, 43, 53] introduce extra or synthetic OOD data, employing fine-tuning to elevate their model's sensitivity towards unknown classes.

With the introduction of large pre-trained VLMs, OOD detection has embarked on a new trajectory driven by VLMs. MCM [32] aims to integrate the idea of maximum softmax probability [10] into the inference process of CLIP, while ZOC [6] enhances OOD detection in a zero-shot setting by learning an additional image interpreter and guessing the category of images. CLIPN [47] and LoCoOp [33], the most related methods to ours, are based on text prompts. However, CLIPN, during its pretraining phase, trains an additional negative text encoder using a large external dataset to improve its negative semantic prompt, which increases network parameters and deviates from prompt learning that is focused on tuning the target data. LoCoOp, on the other hand, uses prompt learning for matching text and image local features, which can compromise the global perception capability of CLIP and reduce classification accuracy for in-distribution (ID) samples. Also, its model lacks knowledge about OOD samples, which can often lead to high detection errors.

3. Method

In this paper, we propose an approach named NegPropmt that leverages pre-trained VLMs, specifically CLIP, to learn negative prompts relative to ID classes for the purpose of OOD detection. The negative prompts are learned in the CLIP's text-image-aligned embedding space with the support of training ID data and their positive prompts (*i.e.*, prompt embeddings of ID classes); no external outlier data is required. Due to its general effectiveness, the popular prompt learning method CoOp [55] is used by default to provide the positive prompts for training NegPrompt.

3.1. Preliminaries

Problem Statement. Formally, we assume that we have two datasets, namely ID dataset denoted as D^{in} and OOD dataset denoted as D^{out}. The ID dataset consists of image-label pairs (x^{in}, y^{in}), where $y^{in} \in Y^{in} = \{0, 1, 2, 3...k\}$ belong to the ID class set. Similarly, the OOD dataset contains image-label pairs (x^{out}, y^{out}), but all $y^{out} \in Y^{out} = \{k+1, k+2, ...\}$ belong to the OOD class set. It is important to note that these two sets do not intersect, meaning that $Y^{in} \cap Y^{out} = \varnothing$. As we have a test set X^{test} consists of images from ID and OOD, the goal of OOD detection is to train a classifier $\phi(x)$ that takes an image x as input and returns whether the image belongs to OOD. Unlike existing studies using full-/zero-shot ID training samples, we only use a few samples (16 samples per class) for D^{in}_{train} like [33] does. In the open-vocabulary detection setting, we only use a small part (10%) but not all the classes of D^{in}_{train}.

CLIP and CoOp. CLIP [37] is currently one of the most popular image-text models. During the pre-training phase, it uses large-scale image-text pairs for self-supervised contrastive learning, aligning images and texts into the same latent space. The main components of CLIP are an image encoder $Encoder^{image}(I)$ and a text encoder $Encoder^{text}(T)$, which respectively accept image and text inputs. In zero-shot image classification tasks, assume we have k class labels for classification, such as "cat", "dog", etc., CLIP first incorporates the class labels into pre-designed hard/unlearnable text prompts, such as "`a photo of a [class name]`", forming a prompt input set of "`a photo of a cat`", "`a photo of a dog`" and so on. These prompts are then individually fed into the text encoder $Encoder^{text}(T)$ to obtain k text features T^f. The testing image is then input into the image encoder to obtain an image feature I^f. The cosine similarity is calculated between the normalized image feature and all text features, formally, $Sim(T^f, I^f) = T^f \cdot I^f$, and the text feature T^f with the highest similarity to I^f is considered to be the category to which the image belongs.

Besides zero-shot classification, many have explored ways to improve CLIP's performance when the target data is accessible. CoOp [55] introduces prompt learning with CLIP by freezing its encoders and using backpropagation to learn dataset-specific soft/learnable prompts. The text prompts are represented as $t_i = \{\omega_1^{pos}, \omega_2^{pos}, ..., \omega_n^{pos}, c_i\}$, where c_i is the word embedding of the class name and ω^{pos}s are learnable vectors that have positive semantics w.r.t. the class i. The goal is to optimize these $\omega^{positive}$s. Specifically, t_i is processed by $Encoder^{text}$ to yield $T_i^{f,pos}$ as a positive prompt embedding of the class i, and then the pre-

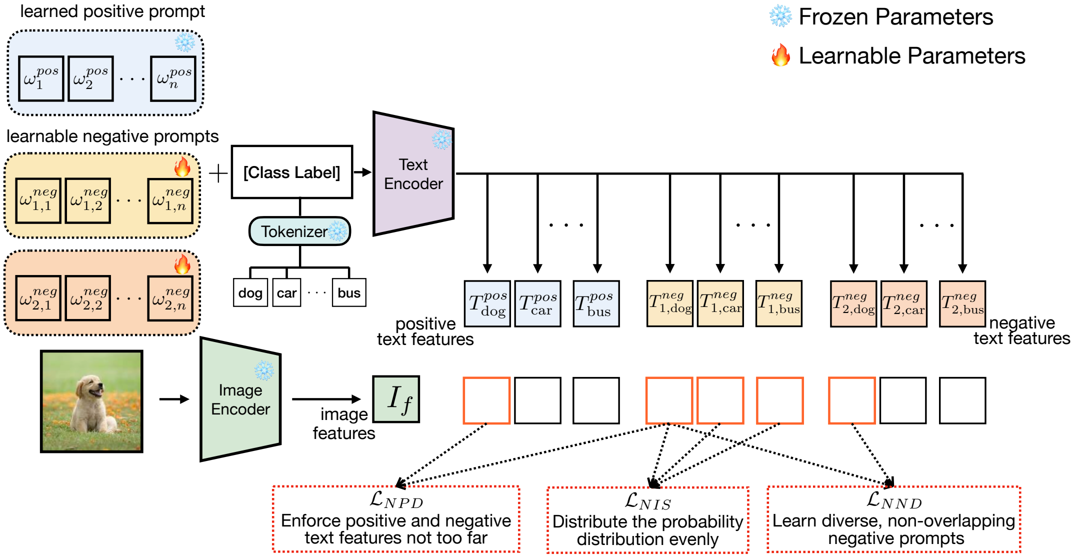
Frozen Parameters
Learnable Parameters
learned positive prompt
ω_1^{pos} ω_2^{pos} ... ω_n^{pos}
learnable negative prompts
$\omega_{1,1}^{neg}$ $\omega_{1,2}^{neg}$ ... $\omega_{1,n}^{neg}$
$\omega_{2,1}^{neg}$ $\omega_{2,2}^{neg}$ ... $\omega_{2,n}^{neg}$
+
[Class Label]
Tokenizer
dog
car
...
bus
Text Encoder
T_{dog}^{pos} T_{car}^{pos} T_{bus}^{pos}
$T_{1,\text{dog}}^{neg}$ $T_{1,\text{car}}^{neg}$ $T_{1,\text{bus}}^{neg}$
$T_{2,\text{dog}}^{neg}$ $T_{2,\text{car}}^{neg}$ $T_{2,\text{bus}}^{neg}$
positive text features
negative text features
Image Encoder
image features
I_f
$\mathcal{L}_{NPD}$
Enforce positive and negative text features not too far
$\mathcal{L}_{NIS}$
Distribute the probability distribution evenly
$\mathcal{L}_{NND}$
Learn diverse, non-overlapping negative prompts

Figure 2. Overview of NegPropmt. Given the CLIP model and positive prompts learned by existing prompt learning methods such as CoOp [55], NegPrompt learns a set of negative prompts relative to different ID class labels via three loss functions that enforce the separation between negative prompts and ID images, and between negative and positive prompts, as well as the diversity of the negative prompts.

diction probability is computed as:

$$p(y=i|x) = \frac{\exp(sim(T_i^{f,pos}, I^f)/\tau)}{\sum_{j=1}^{k}\exp(sim(T_j^{f,pos}, I^f)/\tau)}, \tag{1}$$

where $I^f = Encoder^{image}(x)$ and τ is a temperature parameter. Next, cross-entropy loss is used to maximize similarity between ID class text embeddings and images.

$$\mathcal{L}_{positive} = \mathbb{E}_{\mathbf{x}_{\text{in}} \sim D_{\text{train}}^{\text{in}}}[-\log(p(y=i|x)]. \tag{2}$$

This method largely improves CLIP's classification performance by adapting the learnable prompts to the target data. The resulting prompt embeddings of ID classes are used as positive prompts to support the accurate learning of negative prompts in our method below.

3.2. Proposed Approach

NegPrompt aims to learn a set of negative prompts, each representing a negative connotation of an ID class. As shown in Fig. 2, it involves generating a series of prompts that resemble the positive prompts obtained through CoOp, but with a negative class semantic, such as "A photo of not a [class label]". These negative prompts, combined with class labels, can generate a range of negative text features around the positive prompts, so that OOD images exhibit higher similarity to the negative prompts than ID images.

3.2.1 Learning Negative Prompts

To acquire accurate negative prompts representing negative class semantics, we first utilize CoOp to learn positive prompts ω^{pos}s, after which we consider positive prompts to accurately capture the class semantics of samples in the ID dataset. Thereafter, we freeze the positive prompts and focus solely on learning the negative prompts. A negative prompt is denoted as $\{\omega_1^{neg}, \omega_2^{neg}, ..., \omega_n^{neg}\}$, where n represents the number of context vectors we aim to learn. By utilizing the frozen CLIP text encoder, we can obtain negative prompt embeddings $T_{l,i}^{f,neg} = Encoder^{text}(t_{l,i}^{neg})$, where $t_{l,i}^{neg} = \{\omega_{l,1}^{neg}, \omega_{l,2}^{neg}, ..., \omega_{l,n}^{neg}, c_i\}$ is the l-th negative prompt relative the ID class i. Thus, due to the presence of the negative prompts, we turn the prediction probability into the following form:

$$p(y=i|x) = \frac{\exp(S_i^{f,pos})}{\sum_{j=1}^{k}\exp(S_j^{f,pos}) + \sum_{l=1}^{p}\sum_{j=1}^{k}\exp(S_{l,j}^{f,neg})} \tag{3}$$

where $S_j^{f,pos} = sim(T_j^{f,pos}, I^f)/\tau$, $S_{l,j}^{f,neg} = sim(T_{l,j}^{f,neg}, I^f)/\tau$ and p is the number of negative prompts we aim to learn for each ID class. The objective is to learn negative text prompt embeddings that can effectively separate ID and OOD samples around the positive text features. To achieve this, we introduce the following three loss functions:

Negative-Image Separation Loss. One of our objectives is to have the negative text prompt embeddings serve as the closest text features to OOD images. However, we only have images from the ID dataset; no OOD images are available. To remedy this problem, we take an alternative approach that aims to push the negative text features away from the ID images. In this regard, we draw inspiration from OE [11]. Contrary to the approach in OE, where the probability distribution of outlier images is evenly distributed among all ID labels, we distribute the probability distribution obtained from ID images evenly among all negative prompts. Since the network parameters are frozen, this drives the negative text features to move away from the ID images, resulting in the learning of prompts with a negative connotation. The formulation is as follows:

$$\mathcal{L}_{NIS} = \mathbb{E}_{x_{\text{in}} \sim D^{\text{in}}_{\text{train}}} [H(\mathbf{u}; F(x))], \tag{4}$$

where $F(x)$ is the probability vector computed as $Softmax(S^{f,neg})$, and $\mathbf{u}$ is a uniform distribution and H is the cross entropy loss.

Negative-Positive Distance Loss. To avoid learning trivial negative prompts that are distant from both ID and OOD images, we need to control the negative text feature within a certain range between the ID and OOD images. To this end, we devise a constraint, enforcing that the negative text feature does not deviate too far from the positive text feature. Therefore, we introduce the following loss function to guarantee a certain level of similarity between the negative and positive text feature in the latent space:

$$\mathcal{L}_{NPD} = -\frac{1}{k * p} \sum_{j=1}^{k} \sum_{i=1}^{p} sim(T^{f,neg}_{i,j}, T^{f,pos}_{j}). \tag{5}$$

Negative-Negative Distance Loss. Furthermore, to ensure that we are effectively learning diverse, non-overlapping negative prompts, we extend the distances between different negative text features within the same label via the following loss function:

$$\mathcal{L}_{NND} = \frac{1}{k * p * (p-1)} \sum_{j=1}^{k} \sum_{i=1}^{p} \sum_{l \neq i} sim(T^{f,neg}_{i,j}, T^{f,neg}_{l,j}). \tag{6}$$

Overall, our NegPrompt objective consists of the above three losses. During training, we are able to obtain a diverse set of non-trivial prompts that effectively convey negative meanings relative to the ID class labels by minimizing the following overall loss:

$$\mathcal{L}_{NegativePrompts} = \mathcal{L}_{NIS} + \beta * \mathcal{L}_{NPD} + \gamma * \mathcal{L}_{NND}, \tag{7}$$

where β and γ are hyper-parameters to balance the losses.

3.2.2 Open-Vocabulary Capability

Since the negative prompts we learn do not depend on specific class labels but are instead generic templates representing negative semantics of any given class labels, it is possible to utilize the generalization ability of CLIP to learn a set of transferable negative prompts. Specifically, instead of utilizing all of the ID classes D^{in}_{train}, NegPrompt may only employ its small subset $D^{in,sub}_{train}$ for training the negative prompts ω^{neg}, where $D^{in,sub}_{train} = \{(x, y^{in}_{sub}) \mid y^{in}_{sub} \subset y^{in}\}$. After obtaining the trained negative prompts, we combine them with the remaining ID class names, *i.e.*, replace c_i in $t^{neg}_{l,i}$ with the unseen ID class names, to obtain the corresponding negative prompts for the novel ID classes that are unseen during training. This approach, which achieves out-of-distribution detection by only exposing with a small portion of ID images, is unprecedented in prior research. We refer this as to be open-vocabulary OOD detection.

	ID	OOD
Split-1	All dog classes	Non-animal classes
Split-2	Half of hunting dog classes	Other 4-legged animal classes
Split-3	Mix of common classes	Mix of common classes

Table 1. Three ImageNet-1K splits for hard OOD detection.

3.2.3 Inference

During inference, we employ the MCM [32] scoring approach for OOD detection, but with the addition of our negative prompts into the softmax function. Particularly, MCM uses the inverse of the maximum softmax score in Eq. 1 as the OOD score. Our OOD scoring extends MCM and defines it as: $s(x) = \max(p(y = i|x))$, where $p(y = i|x)$ is defined in Eq. 3 that also includes the similarities of the test image to the negative prompts, in addition to the similarities to the positive prompts. The rationale behind this is that for ID images, they will be matched to one of the positive text features, leading to a higher $S^{f,pos}$ but lower $S^{f,neg}$, and thereby a higher maximum softmax score (*i.e.*, a lower OOD score). Conversely, for OOD data, it will be matched to one of the negative text features, resulting in a lower maximum softmax score (*i.e.*, a higher OOD score).

4. Experiment

4.1. Experimental Details

Datasets. For conventional OOD detection, we use a popular benchmark in which ImageNet-1K [3] with 1,000 classes is used as the ID dataset, and the same OOD datasets as in [32] are used, including subsets of Texture [2], iNaturalist [44], Places [52] and SUN [49]. In addressing the more challenging OOD scenarios, we partitioned the ImageNet1k dataset into two segments: one segment of the data serves as the ID, while the other serves as OOD. As shown in Table 1, three different splits are derived, following from [35]. We further create another ImageNet split, Split-4, in which the first 100 classes are used as ID data and the subsequent 900 classes are used as OOD samples. Following CoOp and LoCoOp [33, 55], during our training process, we utilized only few-shot training data for each category. Particularly, we only train the model with 16 images per ID class and without any exposure to OOD images. During testing, we employ the entire ID and OOD test set for evaluation.

Implementation Details. Following existing studies [47], we use CLIP based on CLIP-B/16 which is pre-trained from OpenCLIP [16]. NegPrompt is trained using 16-shot im-

Method	**Texture**		**iNaturalist**		**Places**		**SUN**		**Avg**	
	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓
Zero-shot methods										
MCM [32]†	86.11	57.77	94.61	30.91	89.77	44.69	92.57	34.59	90.76	42.74
CLIPN [47]†	90.93	40.83	95.27	23.94	92.28	33.45	93.92	26.17	93.10	31.10
CLIP-based posthoc methods										
MSP [10]†	74.84	73.66	77.74	74.57	72.18	79.12	73.97	76.95	74.98	76.22
MaxLogit [12]†	88.63	48.72	88.03	60.88	87.45	55.54	91.16	44.83	88.82	52.49
Energy [27]†	88.22	50.39	87.18	64.98	87.33	57.40	91.17	46.42	88.48	54.80
ReAct [42]†	88.13	49.88	86.87	65.57	87.42	56.85	91.04	46.17	88.37	54.62
ODIN [26] †	87.85	51.67	94.65	30.22	85.54	55.06	87.17	54.04	88.80	47.75
Prompt learning methods										
CoOp [55]	89.47	45.00	93.77	29.81	90.58	40.11	93.29	40.83	91.78	51.68
LoCoOp [33]†	90.19	42.28	96.86	16.05	91.98	32.87	95.07	23.44	93.52	28.66
NegPrompt (Ours)	**91.60**	**35.21**	**98.73**	**6.32**	**93.34**	**27.60**	**95.55**	**22.89**	**94.81**	**23.01**
Open-vocabulary OOD detection										
CoOp (10%)	87.58	50.55	91.08	42.53	89.56	46.12	91.52	41.92	89.94	45.28
LoCoOp (10%)	88.21	47.32	94.47	34.90	91.64	39.85	92.54	26.30	90.15	37.09
NegPrompt (Ours) (10%)	90.30	39.31	98.39	7.48	92.68	29.75	93.70	26.92	93.76	25.86

Table 2. Conventional OOD detection results. We trained using ImageNet1k as the ID and CLIP-B/16 as the CLIP backbone. The boldfaced results indicate the best performance. Results marked with † are taken from [47] and [33]. 'METHOD' (10%) in open-vocabulary OOD detection is to evaluate the performance of the 'METHOD' when only images from 10% ID classes are accessible during training.

ages of all ID classes under the normal OOD detection setting. For open-vocabulary OOD detection, we train our model using only the images of the first 10% classes from the ID dataset, withholding 90% ID classes that only appear together with OOD data during inference. We train a shared positive prompt and two shared negative prompts w.r.t. each training ID class. The hyperparameters β and γ are set to 0.1 and 0.05, respectively (see Appendix A for detail). In the first stage, CoOp is trained for 100 epochs to obtain the positive prompts. In the second stage, the positive prompts are frozen, and our model is trained for 10 epochs to learn the negative prompts. For all experiments, we report the averaged results over three runs with different random seeds.

Comparison Methods. To substantiate the effectiveness of NegPrompts, we conduct an empirical analysis of three distinct categories of methodologies employed for OOD detection utilizing Vision-language models. These categories encompass zero-shot pretraining approaches, the methods that combine the CLIP image encoder with classical approaches, and the methods grounded in prompt learning. In the context of zero-shot methods, we opted for the two recent methods, MCM [32] and CLIPN [47]. MCM employs the original CLIP, utilizing the maximum softmax probability operation on the similarities for detection, and CLIPN involves an additional training phase during pretraining, specifically training a negative text encoder using large external data. For the second group of methods, we adapt previous logits-based methodologies to the use of the CLIP image encoder, including MSP [10], Energy [27], MaxLogit [12], ReAct [42] and ODIN [26], to serve as the CLIP-adapted methods. For the prompt learning methods, NegPrompt is compared with CoOp [55] and LoCoOp [33].

Evaluation Metrics. Two OOD detection metrics are used. The first metric is the False Positive Rate at a 95% True Negative Rate (FPR95), which denotes the rate of falsely identified OOD instances when the true negative rate is maintained at 95%. The second metric is the Area Under the Receiver Operating Characteristic curve (AUROC), representing the measure of OOD ranking across various classification thresholds. We also check the classification accuracy of the ID data to evaluate how the OOD detectors affect the ID classification.

4.2. Comparison to State-of-the-art Models

Conventional OOD Detection. The results of conventional OOD detection are reported in Table 2. It is clear that our proposed NegPrompt achieves consistently superior performance in both individual OOD datasets and the averaged results. When compared with the zero-shot methods, on average, our approach surpasses the best competing method CLIPN by more than 1.5% in AUC and around 8% in FPR95, despite the fact that CLIPN requires the use of an additional large external dataset to train an additional negative text encoder. In other words, although NegPrompt is significantly more lightweight than CLIPN in model size, it can substantially and consistently outperform CLIPN in both metrics across all OOD datasets. The adapted post-hoc methods generally do not leverage the CLIP's capabilities well and thus perform less effectively.

NegPrompt also substantially surpasses both prompt learning-based methods, reducing the FPR95 by about 28% (CoOp) and 5% (LoCoOp). This indicates that the learned negative prompts provide informed knowledge about OOD data, which is lacking in the competing methods, helping largely reduce detection errors.

Hard OOD Detection. Hard OOD detection presents unique challenges as the OOD samples often exhibit some similar features as the ID samples. The results on the four

Method	**Split-1**		**Split-2**		**Split-3**		**Split-4**		**Avg**	
	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓
Zero-shot methods										
MCM	97.93	9.17	88.10	56.40	90.34	33.05	98.72	4.73	93.77	25.83
CLIPN	99.38	2.07	97.77	10.55	90.03	36.85	98.83	4.68	96.50	13.53
CLIP-based posthoc methods										
MSP	77.85	63.60	68.73	83.63	79.10	70.55	82.40	65.52	77.02	70.83
MaxLogit	99.87	0.49	98.06	8.69	90.96	34.34	99.35	2.66	97.06	11.55
Energy	**99.88**	**0.46**	98.18	8.40	90.65	35.02	99.36	2.83	97.02	11.68
ReAct	99.34	0.72	97.91	9.33	90.72	35.65	99.12	2.94	96.77	12.16
ODIN	98.78	1.12	98.23	8.18	89.92	37.20	98.76	13.20	96.42	14.92
Prompt learning methods										
CoOp	98.53	6.78	88.25	50.76	90.64	33.89	98.54	5.11	93.99	24.14
LoCoOp	98.64	6.29	84.63	61.09	91.30	27.79	98.83	41.44	93.35	34.15
NegPrompt (Ours)	99.85	0.62	**98.54**	**7.60**	**93.89**	**22.89**	**99.57**	**1.60**	**97.96**	**8.18**
Open-vocabulary OOD detection										
CoOp (10%)	97.97	12.217	80.11	74.62	87.92	46.00	96.59	16.60	90.65	37.36
LoCoOp (10%)	98.00	9.23	87.02	52.18	80.51	59.93	82.41	48.72	86.99	42.52
NegPrompt (Ours) (10%)	99.66	1.36	96.30	19.89	91.75	26.92	98.14	5.24	96.46	13.36

Table 3. Hard OOD detection results. We use the same notations here as those used in Table 2.

hard OOD datasets derived from ImageNet-1K are shown in Table 3. Similar empirical observations can be derived. Our method NegPropmt is consistently the best performer in the average performance, showcasing its general effectiveness across different dataset splits. The superiority of NegPrompt over the zero-shot and prompt learning-based methods is similar to that in Table 2. Although it is slightly less effective than Energy under the Split-1 setting, it outperforms Energy in all other metrics, achieving maximally over 3% AUC and 12% FPR95 improvement among the performance on the other OOD datasets.

Note that the results in Table 3 are generally more promising than that in Table 2. This is mainly because the OOD detection difficulty in all four ImageNet-1K splits is largely reduced since the number of their ID classes is significantly less than that in the full ImageNet-1K data.

Open-Vocabulary OOD Detection. The open-vocabulary OOD detection results are reported in both Tables 2 and 3. Impressively, even when training on only 10% ID classes, our approach can still perform better than the competing methods using the full ID classes, *e.g.*, the average AUC and FPR95 in Table 2. In this open-vocabulary setting, in general, both LoCoOp and CoOp exhibit a much larger performance decline than NegPrompt, especially on the results in Table 3, in which CoOp has over 3% AUC drop and LoCoOp has over 6% AUC drop while our method has only about 1.5% AUC drop. These results demonstrate that the negative prompts in NegPrompt have much better transferability than those in the two competing methods.

4.2.1 Classification Accuracy on ID Data

We also evaluate the classification accuracy on the ID data when using NegPrompt for OOD detection, with CoOp, LoCoOp, MCM and CLIPN as the baselines. The classification accuracy results on the full ImageNet-1K test data are shown in Table 4. When using the full ImageNet-1K training ID class data, our method NegPrompt can maintain the same classification accuracy as CoOp.

Method	**Top-1 Accuracy**
CoOp	72.1
LoCoOp[†]	71.7
CLIPN & MCM	67.0
NegPrompt(Ours)(10%)	71.9
NegPrompt(Ours)(Full)	72.1

Table 4. Top-1 Accuracy. Results with † are taken from [33].

Our accuracy is slightly compromised when using only 10% ID classes in our training. On the other hand, the OOD detection in LoCoOp compromises the ID classification accuracy, dropped from 72.1% in CoOp to 71.7%. This may be attributed to its focus on the localized regions within the background rather than the primary object of interest, resulting in the missing of some discriminative features for ID data classification. CLIPN and MCM, being zero-shot methods, have not been exposed to the target ID data, leading to a much lower accuracy than the other methods.

4.3. Analysis of NegPrompt

4.3.1 Why Does NegPrompt Work?

To better understand the effectiveness of NegPrompt, we summarize the average maximum similarity between ID/OOD images and positive/negative prompts, as shown in Fig. 3. Across all four OOD datasets, ID images have the highest similarity with positive prompts, and OOD images with negative prompts. This suggests positive prompts are closer to ID images and negative prompts to OOD images in latent space, ensuring OOD images receive lower softmax scores (*i.e.*, higher OOD scores) than ID images in Eq. 3.

Using TinyImageNet [23], a subset of ImageNet, we further visualized the learned negative text features with its test data [6]. Learning three negative prompts per ID class, the results in Fig. 4 demonstrate that positive text features

Ablation Study	Texture AUC ↑	 FPR95 ↓	iNaturalist AUC ↑	 FPR95 ↓	Places AUC ↑	 FPR95 ↓	SUN AUC ↑	 FPR95 ↓	Avg AUC ↑	 FPR95 ↓	 ACC ↑
Backbones											
ResNet-50	79.44	72.73	90.97	44.10	84.35	61.37	88.04	50.55	85.70	57.19	68.2
ResNet-101	82.97	70.83	93.96	31.3	86.41	51.66	88.81	47.61	88.04	50.35	70.3
ViT-B-32	**90.43**	**38.79**	97.40	12.64	**92.83**	32.79	92.82	**26.03**	93.37	27.56	**72.6**
# Negative Prompts											
1	90.04	40.67	98.24	9.23	90.37	32.62	92.26	27.33	92.73	27.46	72.1
Training Process											
One-stage	80.25	72.87	77.02	95.74	82.53	95.94	81.13	95.72	80.23	90.07	62.8
ViT-B-16 & 2 & Two-stage	90.30	39.31	**98.39**	**7.48**	92.68	**29.75**	**93.70**	26.92	**93.76**	**25.86**	72.1

Table 5. Results of our ablation experiments.

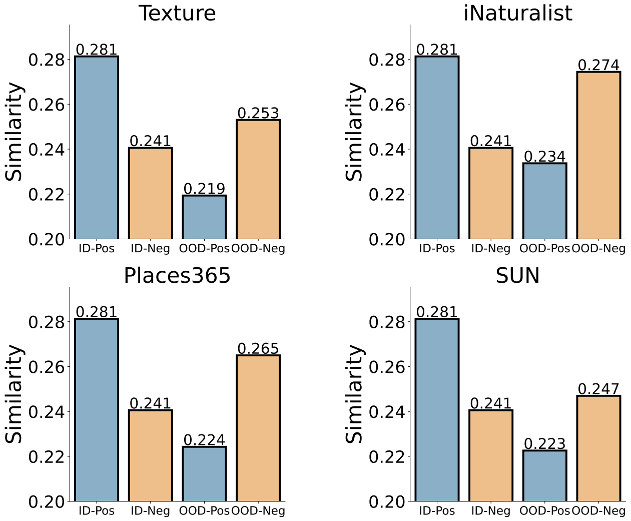

Figure 3. Similarity of ID/OOD and Positive/Negative Prompts.

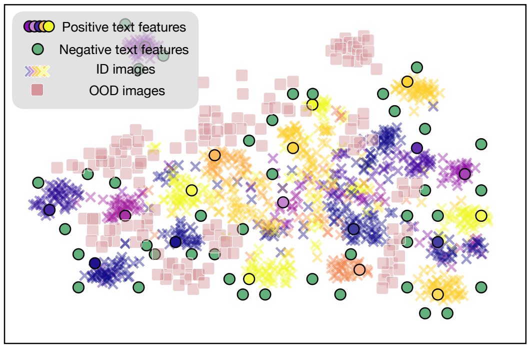

Figure 4. T-SNE visualization of NegPrompt, utilizing a subset of ImageNet - TinyImageNet as the dataset.

from positive prompts align closely with ID images in latent space. Conversely, the negative text features encoded from negative prompts lie outside of the ID data, with OOD images interspersed among them. This suggests negative text features effectively act as a fence aligning much better to the OOD images than the ID images, well supporting the OOD detection while preserving the ID classification accuracy.

4.3.2 Ablation Study

Table 5 shows our ablation study results on the backbone, the number of negative prompts, and the training process.

Backbones. We experiment with diverse CLIP backbones. The results reveal that for CNN-based backbones like ResNet50 and ResNet101, the OOF detection performance is not as proficient as the ViT-based backbones. Regarding ViT-B-32 and ViT-B-16, their performance is found to be uneven. Overall, the performance of OOD detection tends to increase with more advanced backbones.

The Number of Negative Prompts. The number of negative prompts also influences the results. It is observed that the OOD detection performance improves when increasing the number of negative prompts from one to two. Therefore, it is suggested to further increase the number of negative prompts for better detection accuracy. However, note that with an increase in the number of negative prompts, the computational cost also rises rapidly. A good balance between computational cost and detection accuracy is needed when determining the number of negative prompts.

Training Process. The training process is also important. As discussed before, due to the necessity of anchoring the positive prompts, a two-stage training process is used in our model. This process involves training the positive prompts in the first stage and freezing them before proceeding to train the negative prompts in the second stage. When simultaneously training both positive and negative prompts in a unified step, the model's ability to effectively learn positive prompts is significantly undermined, and consequently we obtain unstable negative prompts, leading to the largely decreased OOD detection performance.

5. Conclusion

We present NegPrompt, a novel approach for prompt learning-based OOD detection. It utilizes VLMs to learn a small set of negative prompts for conveying negative semantics relative to ID classes. Our empirical results reveal that NegPrompt 1) achieves superior OOD detection performance compared to the SOTA models across various OOD datasets in both conventional and hard OOD detection scenarios, and 2) learns transfer negative prompts that enable excellent open-vocabulary OOD detection performance.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China 62276016, 62372029.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5, 12

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[6] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6568–6576, 2022. 2, 3, 7

[7] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. *arXiv preprint arXiv:2312.10439*, 2023. 2

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 1, 3, 6

[11] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 3, 5

[12] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 3, 6

[13] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[14] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021. 3

[15] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021. 3

[16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 5, 13

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

[18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 1, 3

[19] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8: 423–438, 2020. 1

[20] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 3

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2

[22] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2021. 3

[23] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 7

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[25] Tianqi Li, Guansong Pang, Xiao Bai, Jin Zheng, Lei Zhou, and Xin Ning. Learning adversarial semantic embeddings for

zero-shot recognition in open worlds. *Pattern Recognition*, 149:110258, 2024. 2

[26] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1, 3, 6

[27] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 3, 6

[28] Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1151–1161, 2023. 3

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[30] Wenjun Miao, Guansong Pang, Tianqi Li, Xiao Bai, and Jin Zheng. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. *arXiv preprint arXiv:2312.10686*, 2023. 3

[31] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 2

[32] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems*, 35:35087–35102, 2022. 2, 3, 5, 6, 12, 13

[33] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023. 2, 3, 5, 6, 7, 12, 13

[34] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015. 2

[35] Andres Palechor, Annesha Bhoumik, and Manuel Günther. Large-scale open-set classification protocols for imagenet. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 42–51, 2023. 5, 12

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 13

[40] 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. 1, 3

[41] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems*, 35:30569–30582, 2022. 3

[42] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 3, 6

[43] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 3

[44] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 5, 12

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[46] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4911–4920, 2022. 3

[47] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 2, 3, 5, 6, 12

[48] Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *arXiv preprint arXiv:2308.11681*, 2023. 3

[49] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5, 12

[50] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[51] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*, 2021. 1

[52] 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*, 40(6):1452–1464, 2017. 5, 12

[53] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2021. 3

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

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3, 4, 5, 6, 13

[56] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 3

[57] Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. *arXiv preprint arXiv:2403.06495*, 2024. 2

Appendix

A. Ablation Study of Hyperparameters β, γ

As discussed in Section 3.2.1, the overall loss of our method comprises three components:

$$\mathcal{L}_{NegativePrompts} = \mathcal{L}_{NIS} + \beta * \mathcal{L}_{NPD} + \gamma * \mathcal{L}_{NND}, \tag{8}$$

where β and γ are two hyperparameters.

We conducted ablation studies on the values of β and γ, with the results presented in the table 6. It is observed that 1) setting either γ or β to zero, *i.e.*, removing $\mathcal{L}_{NPD}$ or $\mathcal{L}_{NND}$, can lead to a significant decrease in AUROC and/or FPR95; and 2) our method achieves the best performance when β and γ are set to 0.1 and 0.05, respectively, indicating a greater importance of $\mathcal{L}_{NPD}$ than $\mathcal{L}_{NND}$.

B. Dataset Information

In this section, we provide a detailed description of the datasets used.

For the conventional OOD detection, the ID images and OOD images are from different datasets: for ID, we utilized ImageNet-1k as our in-distribution datasets, consisting of 1,000 categories with 1,281,167 training images and 50,000 validation images. We train with 16 images per class for few-shot learning and employ all validation images as ID images for testing. Regarding OOD datasets, we followed [32, 33, 47], employing Texture [2], iNaturalist [44], Places [52] and SUN [49] as our OOD test datasets. Further details of these OOD test datasets are provided below.

Texture. Describable Textures Dataset [2] contains images of textures and abstract patterns. As no categories overlap with ImageNet-1k, we use the entire dataset as OOD images.

iNaturalist. Containing images from the real world, iNaturalist [44] has 13 super-categories and 5,089 subcategories covering plants, insects, birds, mammals, and so on. We use the subset that contains 110 plant classes not overlapping with ImageNet-1k.

Places365. As a large scene photograph dataset, Places365 [52] contains photos that are labeled with scene semantic categories from three macro-classes: Indoor, Nature, and Urban. The subset we use is sampled from 50 categories that are not present in ImageNet-1k.

SUN. Scene Understanding Dataset [49] contains 899 categories that cover more than indoor, urban, and natural places with or without human beings appearing. We use the subset which contains 50 natural objects not showing in ImageNet-1k.

Regarding hard Out-of-Distribution detection, both the ID and OOD classes are derived from ImageNet-1k. We adhered to the three splits proposed in [35] and introduce another split: designating the first 100 classes of ImageNet as ID and the remaining 900 classes as OOD, which aims to facilitate comprehensive OOD detection across the entire ImageNet-1k dataset. The number of ID and OOD categories in each split is illustrated in Table 7.

β	γ	Texture		iNaturalist		Places		SUN		Avg	
		AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓
ablation of β											
0	0.05	89.35	37.73	96.85	11.65	90.50	32.86	89.48	26.01	91.55	27.06
0.01	0.05	89.86	36.64	98.53	8.23	91.88	29.13	93.3	24.04	93.39	24.51
0.05	0.05	**91.74**	36.12	97.15	7.23	92.67	28.17	94.36	23.24	93.98	23.69
0.2	0.05	90.94	35.41	98.27	7.16	92.61	27.80	93.66	22.93	93.87	23.33
0.5	0.05	90.69	36.21	97.79	8.73	91.58	28.32	92.85	25.84	93.23	24.78
1	0.05	90.39	36.45	97.13	10.27	91.63	30.92	92.55	25.55	92.93	25.80
	ablation of γ										
0.1	0	89.52	37.11	98.02	10.86	91.4	33.08	92.6	25.08	92.89	26.53
0.1	0.01	89.87	36.13	98.08	9.04	92.44	29.14	93.44	24.85	93.46	24.79
0.1	0.1	91.56	35.84	98.25	7.36	93.17	27.63	95.28	23.33	94.57	23.54
0.1	0.2	90.87	35.41	97.36	7.89	92.15	28.40	93.42	22.97	93.45	23.67
0.1	0.5	91.35	36.39	98.18	9.05	92.35	30.29	92.14	25.91	93.51	25.41
0.1	1	90.77	37.37	98.40	9.46	91.23	31.77	92.84	23.98	93.31	25.65
0.1	0.05	91.60	**35.21**	**98.73**	**6.32**	**93.34**	**27.60**	**95.55**	**22.89**	**94.81**	**23.01**

Table 6. Ablation experiments for hyperparameters β and γ. We fixed the values of β and γ at 0.1 and 0.05, respectively, and conducted controlled experiments. The best results are highlighted in **bold**, and each result was averaged over three trials.

	ID	**OOD**
Split-1	All dog classes 116: 1856 / 5800	Non-animal classes 166: — / 8300
Split-2	Half of hunting dog classes 30: 480 / 1500	Other 4-legged animal classes 55: — / 2750
Split-3	Mix of common classes 151: 2416 / 7550	Mix of common classes 164: — / 8200
Split-4	First 100 classes 100: 1600 / 5000	Remaining 900 classes 900: — / 45000

Table 7. All ImageNet-1K splits for hard OOD detection. Given are the numbers of *classes* : *training* / *test* samples.

C. Implementation Details

All methods are implemented in Pytorch 1.10. We run all OOD detection experiments on an NVIDIA V100 GPU. For CLIP, we follow LoCoOp [33] and employ the implementation of OpenClip [16], using parameters pre-trained with LAION-2B [39]. A two-stage training method is adopted: during the first stage, we train only positive prompts, following the same training scheme as CoOp [55]. A single positive prompt is trained for all classes, with the number of context tokens set to 16. In the second stage, we freeze the positive prompts and initialize the negative prompt parameters as the positive prompt, followed by training for 15 epochs. During the testing phase, we employ the MCM [32], where the highest similarity between the positive text feature and the image feature, post-Softmax of all similarities, is taken as the result.

D. Computational Time

A comparative analysis of the computational time for prompt learning-based approaches on ImageNet-1k/Texture is also performed. As indicated in Table 8, our approach has a higher training overhead compared to CoOp, yet it is much lower than that of LoCoOp. Regarding the inference time, NegPrompt aligns with CoOp and surpasses LoCoOp.

Method	**Training Time**	**Inference Time**
LoCoOp [33]	475 min	20 min
CoOp [55]	341 min	3 min
NegPrompt(ours)	443 min	3 min

Table 8. The computation time of prompt learning-based methods. For LoCoOp [33], we use its official implementation. The training time and inference time are averaged from three separate experiments.