

# FA: Forced Prompt Learning of Vision-Language Models for Out-of-Distribution Detection

Xinhua Lu<sup>1,2,4</sup>, Runhe Lai<sup>1,2,4</sup>, Yanqi Wu<sup>1,2,4</sup>, Kanghao Chen<sup>3†</sup>, Wei-Shi Zheng<sup>1,2,4</sup>, Ruixuan Wang<sup>1,2,4†</sup>

<sup>1</sup>Sun Yat-sen University <sup>2</sup>Peng Cheng Laboratory

<sup>3</sup>Hong Kong University of Science and Technology (Guangzhou)

<sup>4</sup>Key Laboratory of Machine Intelligence and Advanced Computing

{luxh55, lairh5, wuyq268}@mail2.sysu.edu.cn, kchen879@connect.hkust-gz.edu.cn,

wszheng@ieee.org, wangruix5@mail.sysu.edu.cn

## Abstract

Pre-trained vision-language models (VLMs) have advanced out-of-distribution (OOD) detection recently. However, existing CLIP-based methods often focus on learning OOD-related knowledge to improve OOD detection, showing limited generalization or reliance on external large-scale auxiliary datasets. In this study, instead of delving into the intricate OOD-related knowledge, we propose an innovative CLIP-based framework based on **Forced prompt leArning (FA)**, designed to make full use of the In-Distribution (ID) knowledge and ultimately boost the effectiveness of OOD detection. Our key insight is to learn a prompt (i.e. forced prompt) that contains more diversified and richer descriptions of the ID classes beyond the textual semantics of class labels. Specifically, it promotes better discernment for ID images, by forcing more notable semantic similarity between ID images and the learnable forced prompt. Moreover, we introduce a forced coefficient, encouraging the forced prompt to learn more comprehensive and nuanced descriptions of the ID classes. In this way, FA is capable of achieving notable improvements in OOD detection, even when trained without any external auxiliary datasets, while maintaining an identical number of trainable parameters as CoOp. Extensive empirical evaluations confirm our method consistently outperforms current state-of-the-art methods. Code is available at <https://github.com/0xFAFA/FA>.

## 1. Introduction

AI models often encounter Out-of-Distribution (OOD) samples [14, 25, 28, 38], which differ from the distribution of the training data, when deployed in real-world applications.

† Corresponding author.

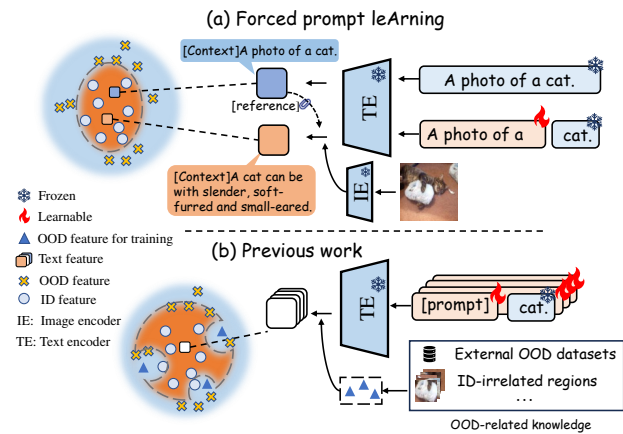


Figure 1. Comparison of CLIP-based OOD detection methods. Existing methods focus on OOD-related knowledge to learn a complex ID/OOD decision boundary as shown in the orange area in (b). In contrast, our FA aims to fully exploit ID knowledge by forcing the model to learn richer descriptions of the ID classes beyond the textual semantics of class labels. These richer descriptions of the ID classes compared to the reference text allow better discernment of ID/OOD samples (orange vs. blue area in (a)).

Detecting OOD data is vital for the reliability of AI systems. This problem becomes harder in fields like intelligent diagnosis, where only a small amount of labeled data is available, motivating the problem of few-shot OOD detection.

To tackle the OOD detection problem, previous single-modal post-hoc methods [14, 16, 24, 27, 41] achieve notable success based on a model pre-trained on the ID dataset. However, they come with limitations, such as requiring substantial computational resources and annotation costs for training, which hinder the performance of few-shot OOD detection. With the development of large pre-trained vision-language models (VLMs) like CLIP [37], few-shot OOD detection has achieved remarkable performance. Most of the CLIP-based OOD detection methods [30, 32] focus on

leveraging the powerful generalization capability of CLIP to improve OOD detection.

Recent CLIP-based OOD detection methods [1, 11, 23, 33, 34, 46, 48] aim to learn OOD-related knowledge to improve OOD detection performance. For example, recent CLIP-based works [11, 46] improve OOD detection by relying on training on external large-scale auxiliary datasets, which entails significant resource overhead. To eliminate dependence on external datasets, existing methods [1, 33, 48] learn OOD-related knowledge from exposed outliers based only on ID training data, such as background regions in ID training images. However, these OOD features extracted from specific regions are difficult to match with infinite OOD data encountered in practice. Moreover, some methods [23, 34] aim to learn negative prompts semantically opposite to the ID class labels, while these limited negative prompts are often insufficient to capture the distinctions between the diverse and numerous OOD data and the ID data. Overall, improving OOD detection by learning OOD-related knowledge shows inherent limitations or requires significant computational resources as well as labor costs. Inspired by the method [44] for open-set recognition, which demonstrates that enhancing closed-set accuracy can typically improve open-set recognition capabilities, we are motivated to explore ways to improve the identification of ID images, thereby enhancing OOD detection performance.

In this paper, rather than focusing on the intricate OOD-related knowledge, we propose a novel CLIP-based framework based on Forced prompt leArning (FA), which aims to fully exploit the ID knowledge and ultimately improve OOD detection. Our key insight is to learn a prompt that contains richer knowledge beyond the textual semantics of class labels. It facilitates advanced discernment for ID images, by forcing higher semantic similarity between ID images and the learnable prompt (*i.e.* forced prompt), which provides more diversified and richer descriptions of the ID classes. In this way, FA can achieve significant improvements in OOD detection performance, even when trained without any additional external data, while maintaining the same number of learnable parameters as CoOp [54].

Specifically, we introduce a novel forced prompt along with the original prompt, both of which are initialized identically. FA preserves the generalization capability of the VLMs (*i.e.* CLIP) by freezing the original prompt and optimizing a trainable copy (*i.e.* forced prompt). In particular, the forced prompt treats the original prompt as a reference, while forcing the text features associated with the forced prompt to be more salient compared to those associated with the original prompt. In this way, it effectively improves the OOD performance even without sacrificing the ID classification capability of the model. Additionally, we introduce a forced coefficient, encouraging the forced

prompt to learn more comprehensive connotations. Experimentally, FA achieves superior few-shot OOD detection performance across diverse OOD benchmarks. Compared to current SOTA methods [48] that focus on learning OOD-related knowledge, our method significantly reduces the average FPR95 score from 31.62% to 27.81% and improves the average AUROC from 92.01% to 93.26% even in 1-shot OOD detection on ImageNet-1k. The main contributions are summarized as follows:

- We propose a simple yet effective framework, which fully exploits the ID knowledge to improve OOD detection without focusing on intricate OOD-related knowledge.
- We present a Forced prompt leArning (FA) strategy to exploit richer knowledge beyond the textual semantics of class labels.
- We evaluate our method on diverse OOD benchmarks, showing that our model consistently outperforms current state-of-the-art methods.

## 2. Related work

**Prompt Learning.** In recent years, pre-trained vision-language models (VLMs) such as CLIP [37] have demonstrated powerful few-shot learning capabilities in both visual and textual domains. However, the design of the prompt greatly affects the performance of VLMs for downstream tasks, which may require manually crafting numerous prompts. Inspired by prompt learning studies in natural language processing (NLP) [26], CoOp [54] uses a set of learnable context vectors to transfer CLIP for specific downstream tasks, becoming a pioneering method for subsequent studies [4, 33, 53]. Currently, prompt learning is widely used in OOD detection mainly to represent various OOD-related knowledge. In this work, we learn the forced prompt, which contains richer descriptions of the ID classes to improve OOD detection.

**Out-of-Distribution Detection.** Conventional methods explore OOD detection for single-modal models. One line of studies designs score functions to differentiate ID and OOD data based on output from the pre-trained models, such as logit output [14, 16, 24, 27] or outputs from the penultimate layer [22, 41, 42, 50]. These methods are called post-hoc methods. Another line of studies [9, 10, 18, 31, 39, 43, 55] adopts various training strategies to learn a reliable decision boundary between ID and OOD data.

Recently, the development of OOD detectors based on VLMs, especially CLIP, has received much attention as VLMs have demonstrated their remarkable generalization capability in both visual and textual domains. Some studies leverage real outlier information from external auxiliary OOD datasets [11, 46] or extensive corpora [5, 20] to promote OOD detection. However, this is impractical in real-world scenarios, where outliers are infinite and agnostic.

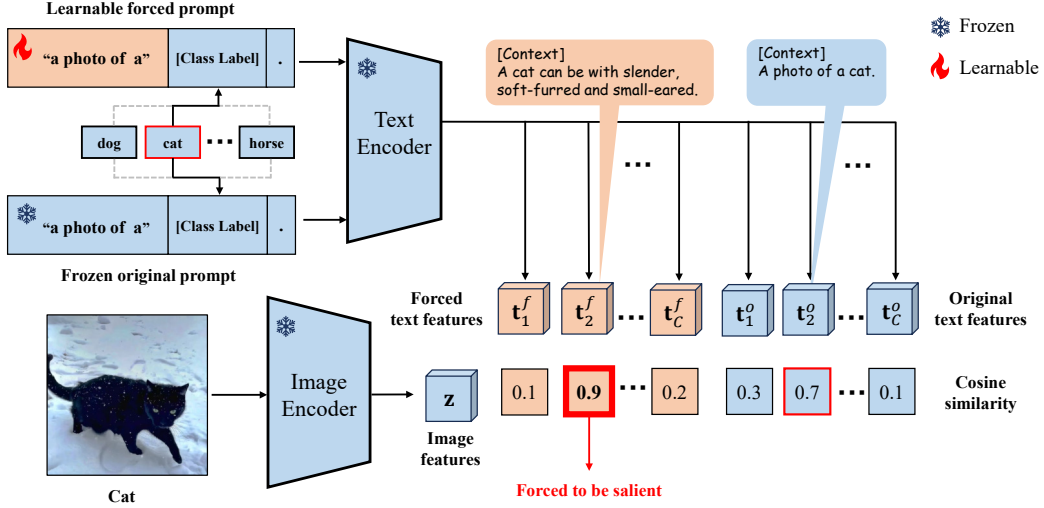


Figure 2. Overview of the proposed FA framework. Our framework includes the learnable forced prompt and the frozen original prompt, both of which are initialized identically by a manual template “a photo of a [class- $c$ ]”. The learnable forced prompt treats the frozen original prompt as a reference, forcing its text features to become more salient compared to those generated by the original prompt. The richer and more diversified description of ID classes learned by FA can ultimately improve OOD detection.

To eliminate the dependence on external data, recent methods [1, 23, 32, 33, 48] were developed under the assumption that only ID data are available. As a representative zero-shot method that only uses available ID labels, MCM [30] employs softmax scaling to align visual features with ID text concepts for OOD detection. GL-MCM [32] further introduces a local maximum concept matching (L-MCM) score to improve the separability of the local information. Compared with zero-shot methods, which may undergo a domain gap with ID downstream data, prompt learning methods achieve better OOD detection performance with access to few-shot ID samples. For prompt learning OOD detection methods, LoCoOp [33] and SCT [48] keep the textual embeddings of ID classes away from ID-irrelevant local region embeddings in the multi-modal embedding space by their proposed entropy maximization strategy, which ensures that OOD embeddings can be dissimilar to any textual embedding of ID classes. To detect challenging OOD samples to improve OOD detection, ID-like [1] explores the vicinity of ID samples to construct OOD samples correlated to the ID and refines the additional “ID-like” text embeddings to fine-grained differences. In addition, Neg-Prompt [23] and LSN [34] introduce the negative connotations of ID categories which can be represented by additional negative prompts, enabling more accurate detection of OOD samples. Overall, various OOD-related knowledge is commonly adopted to address the model’s overconfidence in current research. In this work, we focus on leveraging ID knowledge to improve OOD detection, instead of exploring complex OOD-related knowledge.

### 3. Preliminaries

**Prompt Learning with CLIP.** CLIP contains an image encoder  $f(\cdot)$  and a text encoder  $g(\cdot)$ , designed to extract features from images and text descriptions respectively. Generally, for an ID dataset which contains  $C$  categories, the hand-crafted prompt  $\hat{\mathbf{u}}_c$  “a photo of a [class- $c$ ]” is designed to match images, where  $class-c$  represents the class name. Formally, these prompts are individually encoded in  $\hat{\mathbf{t}}_c = g(\hat{\mathbf{u}}_c) \in \mathbb{R}^{d \times 1}$ ,  $c = 1, \dots, C$ . Given an image  $\mathbf{x}$ , it can be encoded in  $\mathbf{z} = f(\mathbf{x}) \in \mathbb{R}^{d \times 1}$ . Because image features and corresponding text features are aligned on the multi-modal embedding space, the cosine similarity  $\cos(\mathbf{z}, \hat{\mathbf{t}}_c)$  of  $\mathbf{z}$  and  $\hat{\mathbf{t}}_c$  can represent the matching degree between the image  $\mathbf{x}$  and text description  $\hat{\mathbf{u}}_c$ . Therefore, if the text feature  $\hat{\mathbf{t}}_c$  of class  $c$  has the highest similarity to  $\mathbf{z}$ , the image  $\mathbf{x}$  will be considered to be class  $c$ .

To effectively transfer CLIP to downstream image recognition tasks, CoOp [54] sets a portion of tokens in the text prompt as continuous learnable parameters. Concretely, CoOp initializes the text prompts as  $\mathbf{u}_c = [\mathbf{v}_1, \dots, \mathbf{v}_L, \mathbf{w}_c]$ ,  $c \in \{1, \dots, C\}$ , where  $L$  is the length of the prompt’s tokens,  $\mathbf{v}_i$  ( $i \in \{1, \dots, L\}$ ) is the learnable vector with the same dimension as the word embedding and  $\mathbf{w}_c$  is the word embedding for the class name of class  $c$ . The text prompt  $\mathbf{u}_c$  is encoded into features  $\mathbf{t}_c = g(\mathbf{u}_c)$ . Together with the image feature  $\mathbf{z}$ , the probability for the image  $\mathbf{x}$  to be classified into class  $c$  is defined as

$$p(y = c | \mathbf{x}) = \frac{e^{\cos(\mathbf{z}, \mathbf{t}_c)/\tau}}{\sum_{j=1}^C e^{\cos(\mathbf{z}, \mathbf{t}_j)/\tau}}, \quad (1)$$

where  $\tau$  is a fixed temperature scaling hyper-parameter.

Overall, the prompt’s tokens can be trained to align with training data by minimizing the cross-entropy loss with the prediction probability from Eq. (1).

**OOD Detection.** In the OOD detection task, the model is expected to determine whether a test image belongs to one of the learned ID classes. Therefore, OOD detection can be seen as a binary classification problem to distinguish ID images from OOD images in a test set  $\mathcal{D}_{test}$  as follows

$$D(\mathbf{x}) = \begin{cases} 1, & \text{if } S(\mathbf{x}) \geq \mu \\ 0, & \text{if } S(\mathbf{x}) < \mu \end{cases}, \quad (2)$$

where  $\mathbf{x} \in \mathcal{D}_{test}$ , 1 and 0 respectively indicate that  $\mathbf{x}$  is classified as the ID class and the OOD class by the OOD detector  $D(\cdot)$ ,  $S(\cdot)$  is certain score function, and  $\mu$  is a pre-defined threshold constant.

## 4. Methodology

### 4.1. Overview

In this study, we propose a novel CLIP-based framework based on Forced prompt leArning (FA), designed to fully exploit the ID knowledge from few-shot ID samples by forcing the prompt to learn richer knowledge beyond the textual semantics of class labels. As shown in Fig. 2, we introduce a forced prompt along with the original prompt, both of which are initialized with the same semantics. By freezing the original prompt and optimizing a trainable copy, ID images will show higher semantic similarity to the forced prompt compared to the original prompt. Moreover, to encourage descriptions of ID classes learned by the forced prompt to become more detailed and comprehensive, we also introduce a forced coefficient. During testing, the model is capable of effectively distinguishing between ID data and OOD data, benefiting from the diversified descriptions of the ID classes provided by the forced prompt.

### 4.2. Forced prompt learning

Merely the semantic information conveyed by class names is insufficient to comprehensively encompass the discriminative information of each class. In order to explore richer semantic descriptions of the ID classes beyond the textual semantics of class labels, a novel forced prompt along with the original prompt is proposed. To implement the above objective, we introduce a simple yet effective training strategy using the forced cross-entropy (FCE) loss based on the forced prompt as follows

$$\mathcal{L}_{FCE} = \mathbb{E}_{(\mathbf{x}, y_c) \sim \mathcal{D}_{train}^{ID}} \left[ -\log \frac{e^{s_c^f / \tau}}{\sum_{j=1}^C e^{s_j^f / \tau} + \sum_{j=1}^C e^{s_j^o / \tau}} \right], \quad (3)$$

where the ID training dataset  $\mathcal{D}_{train}^{ID}$  consists of ID image-label pairs  $(\mathbf{x}, y_c)$ ,  $s_j^f = \cos(\mathbf{z}, \mathbf{t}_j^f)$  and  $s_j^o = \cos(\mathbf{z}, \mathbf{t}_j^o)$

represent the similarity between the image feature and the prompt feature corresponding to the forced prompt and the original prompt, respectively.

Based on this loss function, the ID image features will be forced to show higher cosine similarity to the learnable forced prompt compared to the original text prompt. This is because the cosine similarity between the ID image features and the original text prompt is already high before any further prompt learning. Consequently, the forced prompt is forced to uncover richer discriminative information of ID classes, so that the text features of the forced prompt achieve a higher cosine similarity with the image features compared to those of the original prompt. Benefiting from this information specific to the ID classes, the cosine similarity between the image features of an OOD image and the text features of both prompts may exhibit less distinction compared to those of the ID data. Therefore, even when trained without any reliance on external auxiliary datasets, the detector will have a notable capability to distinguish ID data from OOD data.

However, this naive FCE loss may not work effectively because of its inherent limitations. Concretely, employing random initialization for both prompts fails to fully leverage the semantic information provided by CLIP’s prior knowledge, resulting in limited performance. To solve this limitation, we propose to initialize the forced prompt using the manual template (*i.e.* “a photo of a [class-c]”) identified to the original prompt. In this way, both prompts with manual initialization will have clear semantic information compared to random initialization, thereby improving the model’s generalization capability. Specifically, the class labels are utilized in both prompts by concatenating with the prompt embedding to provide foundational semantic information related to the category names, based on CLIP’s prior knowledge. In particular, for the forced prompt, we adopt the shared learnable vector across all classes rather than the independent learnable vector for each class. This choice is inspired by CoOp [54], which demonstrates that using an independent learnable vector mostly underperforms the shared learnable vector in challenging low-data scenarios since the former has more parameters and requires more data for training. Formally, we utilize the embeddings of the manual template to initialize the prompt’s embeddings of both prompts, formulated as  $\mathbf{u}_c = [\mathbf{v}_1, \dots, \mathbf{v}_L, \mathbf{w}_c]$ ,  $c \in \{1, \dots, C\}$ , which will be fed to the text encoder to obtain prompt feature  $\mathbf{t}_c$  (see Section 3). Here,  $L$  is the length of the token (*e.g.*, for the “a photo of a”,  $L = 4$ ). Then we freeze the original prompt and the class label part of both the prompts to preserve the generalization capability of CLIP, while only making the forced prompt learnable. Note that our method effectively improves OOD detection performance while maintaining consistency with CoOp in the number of learnable parameters without leveraging ad-

ditional learnable prompts, unlike existing work [1, 23, 34].

Although the above prompt design achieves moderate performance, relying solely on a single original prompt as a reference is somewhat insufficient to develop a comprehensive capability, particularly with the limitations of ID data (*e.g.*, few-shot OOD). To encourage the forced prompt to capture more comprehensive and nuanced descriptions of the ID classes, we introduce a forced coefficient  $K$  ( $K \geq 0, K \in \mathbb{N}$ ), which indicates the intensity with which the model is compelled to learn from the data. Formally, we first derive the original text features  $\{\mathbf{t}_1^o, \dots, \mathbf{t}_C^o\}$  from the original prompt and the image features  $\mathbf{z}$  from the corresponding image. Then, we compute the cosine similarity between the original text features and the image features for  $K$  iterations, which will be used in the subsequent computation of the cross-entropy loss function. Notably, when  $K = 0$ , no original prompt is used, and the model operates equivalently to CoOp. Based on the forced coefficient, we refine Eq. (3) as follows

$$\mathcal{L}_{FCE-K} = \mathbb{E}_{(\mathbf{x}, y_c) \sim \mathcal{D}_{train}^{ID}} \left[ -\log \frac{e^{s_c^f/\tau}}{\sum_{j=1}^C e^{s_j^f/\tau} + K \sum_{j=1}^C e^{s_j^o/\tau}} \right]. \quad (4)$$

Based on this function, the cosine similarity between the ID image features and the text features of the forced prompt must become more salient as the coefficient  $K$  increases.

### 4.3. Model inference

During model inference, for the downstream classification task, we adopt the same strategy as CLIP, depending solely on the forced prompt [1]. For the OOD detection task, our method can be flexibly combined with different score functions, such as the MCM [30] score and the GL-MCM [32] score. The MCM score is defined as the maximum similarity between the global image features  $\mathbf{z}^g$  (*i.e.*  $\mathbf{z}$ ) and all text features  $\mathbf{t}_c^a$  (*i.e.* the concatenation of text features  $\mathbf{t}_c^f$  and  $\mathbf{t}_c^o$ ) after applying softmax with temperature  $\tau_0$ , *i.e.*,

$$S_{\text{MCM}}(\mathbf{x}) = \max_c \frac{e^{\cos(\mathbf{z}^g, \mathbf{t}_c^a)/\tau_0}}{\sum_{j=1}^C e^{\cos(\mathbf{z}^g, \mathbf{t}_j^f)/\tau_0} + K e^{\cos(\mathbf{z}^g, \mathbf{t}_j^o)/\tau_0}}, \quad (5)$$

while the GL-MCM score simultaneously considers both global and local features, which can be expressed as

$$S_{\text{GL-MCM}}(\mathbf{x}) = S_{\text{MCM}}(\mathbf{x}) + S_{\text{L-MCM}}(\mathbf{x}), \quad (6)$$

$$S_{\text{L-MCM}}(\mathbf{x}) = \max_{i,c} \frac{e^{\cos(\mathbf{z}_i^l, \mathbf{t}_c^a)/\tau_0}}{\sum_{j=1}^C e^{\cos(\mathbf{z}_i^l, \mathbf{t}_j^f)/\tau_0} + K e^{\cos(\mathbf{z}_i^l, \mathbf{t}_j^o)/\tau_0}}, \quad (7)$$

where  $\mathbf{z}_i^l$  ( $i \in \{1, \dots, N\}$ ) represents  $N$  extracted local features generated by CLIP’s image encoder [32, 33]. We set  $\tau_0 = 1$  during model inference.

## 5. Experiments

### 5.1. Experimental details

**Datasets.** We use a popular benchmark for conventional OOD detection [19, 30, 32, 33], where ImageNet-1k [7] serves as the ID dataset, and the OOD datasets are iNaturalist [17], SUN [47], Places [52], and Texture [6]. Moreover, inspired by existing studies [1, 2], we use datasets like OpenImage-O [45], NINCO [2], and ImageNet-O [15], which are cleaner and more realistic, to simulate the more challenging OOD detection. Besides, we also utilize other widely adopted datasets for few-shot settings as ID datasets [49, 51, 54], which include Stanford-Cars [21], UCF101 [40], Caltech101 [12], Flowers102 [35], EuroSAT [13], FGVCAircraft [29], OxfordPets [36], and Food101 [3], considering factors such as image resolution and the number of classes.

**Setup.** Following previous studies [33], we use the ViT-B/16 [8] as the backbone model. For our model, when using ImageNet-1k as the ID dataset, the epochs are respectively set to 30 and 50 for the 1-shot and 16-shot settings, while for other ID datasets, we set the epoch to 200 following CoOp [54]. The value of forced coefficient  $K$  is uniformly set to 3, which will be further discussed in the sensitivity study. Other hyperparameters are as follows: learning rate =  $2e-3$ , batch size = 160, SGD (momentum = 0.9, weight decay =  $5e-4$ ) as the optimizer with a cosine scheduler,  $\tau = 1$  in Eq. (3) and Eq. (4). All experiments on our model can be conducted on a single Nvidia A30 GPU. The average experiment results (including our reproductions) over four runs are reported for comparison.

**Comparison Methods.** To validate the effectiveness of our FA fairly, we mainly compare it with CLIP-based OOD detection methods that do not use real outliers (*e.g.* OOD labels). For previous post-hoc methods, including MSP [14], ODIN [24], Energy [27], ReAct [41], and MaxLogit [16], we adapt these methods with the CLIP image encoder as the CLIP-based post-hoc methods. For zero-shot methods, we select MCM [30], GL-MCM [32] and CLIPN [46] as baselines. For prompt learning methods, we adopt CoOp [54], LoCoOp [33], IDLike [1], and SCT [48] as baselines.

**Metrics.** For evaluation, we adopt the following metrics: (1) The False Positive Rate at 95% True Positive Rate for in-distribution samples (FPR95); (2) The Area Under the Receiver Operating Characteristic Curve (AUROC); (3) In-distribution data classification Top-1 accuracy (ID ACC).

### 5.2. Main results

**Conventional OOD Detection.** Table 1 summarizes our comparison results on the ImageNet-1k benchmarks, which show that our FA achieves state-of-the-art OOD detection performance among other CLIP-based methods under different few-shot settings (More details under the 4-shot set-

Method	iNaturalist		SUN		Places		Textures		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
<i>Zero-shot methods</i>										
MCM	31.95	94.16	37.22	92.55	42.98	90.10	58.35	85.83	42.63	90.66
GL-MCM	15.09	96.72	29.08	93.41	37.07	90.37	58.94	83.11	35.04	90.90
CLIPN	19.17	96.17	26.43	94.02	32.26	92.62	41.23	90.12	30.21	93.19
<i>CLIP-based post-hoc methods</i>										
MSP <sup>†</sup>	74.57	77.74	76.95	73.97	79.72	72.18	73.66	74.84	76.22	74.68
ODIN <sup>†</sup>	98.93	57.73	88.72	78.42	87.80	76.88	85.47	71.49	90.23	71.13
Energy <sup>†</sup>	64.98	87.18	46.42	91.17	57.40	87.33	50.39	88.22	54.80	88.48
ReAct <sup>†</sup>	65.57	86.87	46.17	91.04	56.85	87.42	49.88	88.13	54.62	88.37
MaxLogit <sup>†</sup>	60.88	88.03	44.83	91.16	55.54	87.45	48.72	88.63	52.49	88.82
<i>Prompt learning based methods</i>										
					1-shot					
CoOp <sub>MCM</sub>	41.14 <sup>±9.39</sup>	91.47 <sup>±2.12</sup>	39.06 <sup>±4.17</sup>	91.68 <sup>±0.83</sup>	45.38 <sup>±4.88</sup>	89.16 <sup>±1.28</sup>	51.37 <sup>±3.09</sup>	87.82 <sup>±1.12</sup>	44.24 <sup>±1.39</sup>	90.03 <sup>±0.32</sup>
CoOp <sub>GL</sub>	23.30 <sup>±6.37</sup>	94.59 <sup>±1.51</sup>	32.08 <sup>±3.87</sup>	92.21 <sup>±0.95</sup>	39.22 <sup>±4.55</sup>	89.59 <sup>±1.58</sup>	55.78 <sup>±3.46</sup>	83.26 <sup>±1.38</sup>	37.59 <sup>±0.34</sup>	89.91 <sup>±0.31</sup>
LoCoOp <sub>MCM</sub>	36.64 <sup>±4.87</sup>	92.78 <sup>±0.93</sup>	31.86 <sup>±3.79</sup>	93.50 <sup>±0.92</sup>	38.81 <sup>±3.37</sup>	90.70 <sup>±0.97</sup>	48.11 <sup>±2.31</sup>	89.43 <sup>±0.68</sup>	38.85 <sup>±2.67</sup>	91.59 <sup>±0.58</sup>
LoCoOp <sub>GL</sub>	21.97 <sup>±2.91</sup>	95.39 <sup>±0.59</sup>	<u>24.95</u> <sup>±2.37</sup>	<b>94.42</b> <sup>±0.59</sup>	34.14 <sup>±2.66</sup>	91.14 <sup>±0.69</sup>	49.04 <sup>±2.77</sup>	87.73 <sup>±0.96</sup>	32.53 <sup>±2.19</sup>	92.17 <sup>±0.47</sup>
IDLike	<u>17.73</u> <sup>±1.91</sup>	96.68 <sup>±0.31</sup>	48.17 <sup>±1.39</sup>	89.53 <sup>±0.58</sup>	50.43 <sup>±6.58</sup>	88.27 <sup>±2.27</sup>	<u>29.12</u> <sup>±7.64</sup>	<u>93.25</u> <sup>±2.16</sup>	36.36 <sup>±3.86</sup>	91.93 <sup>±1.18</sup>
LSN <sup>†</sup>	59.28 <sup>±7.02</sup>	87.20 <sup>±3.15</sup>	40.15 <sup>±0.82</sup>	91.47 <sup>±0.14</sup>	46.11 <sup>±1.86</sup>	88.74 <sup>±0.57</sup>	60.34 <sup>±0.14</sup>	83.92 <sup>±0.42</sup>	51.47 <sup>±1.53</sup>	87.84 <sup>±0.58</sup>
NegPrompt <sup>†</sup>	65.03 <sup>±8.69</sup>	84.56 <sup>±2.52</sup>	44.39 <sup>±1.66</sup>	89.63 <sup>±0.66</sup>	51.31 <sup>±6.21</sup>	86.55 <sup>±2.19</sup>	87.60 <sup>±1.61</sup>	63.76 <sup>±3.02</sup>	62.08 <sup>±3.71</sup>	81.13 <sup>±1.78</sup>
SCT <sub>MCM</sub>	41.93 <sup>±12.17</sup>	91.77 <sup>±2.40</sup>	30.39 <sup>±2.17</sup>	93.76 <sup>±0.39</sup>	38.73 <sup>±2.54</sup>	90.78 <sup>±0.34</sup>	46.78 <sup>±3.51</sup>	88.96 <sup>±1.16</sup>	39.46 <sup>±3.39</sup>	91.32 <sup>±0.90</sup>
SCT <sub>GL</sub>	20.57 <sup>±10.2</sup>	<u>95.63</u> <sup>±1.89</sup>	<b>24.56</b> <sup>±3.03</sup>	<u>94.39</u> <sup>±0.53</sup>	<u>33.27</u> <sup>±2.96</sup>	91.27 <sup>±0.57</sup>	48.12 <sup>±2.97</sup>	86.76 <sup>±0.89</sup>	31.62 <sup>±3.19</sup>	92.01 <sup>±0.77</sup>
FA <sub>MCM</sub> (Ours)	25.50 <sup>±2.72</sup>	94.72 <sup>±0.27</sup>	36.24 <sup>±3.15</sup>	92.40 <sup>±0.95</sup>	35.38 <sup>±2.75</sup>	<b>91.99</b> <sup>±0.68</sup>	<b>28.34</b> <sup>±0.87</sup>	<b>93.95</b> <sup>±0.19</sup>	<u>31.37</u> <sup>±1.07</sup>	<u>93.25</u> <sup>±0.32</sup>
FA <sub>GL</sub> (Ours)	<b>14.12</b> <sup>±1.32</sup>	<b>96.76</b> <sup>±0.10</sup>	29.99 <sup>±1.57</sup>	92.95 <sup>±0.66</sup>	<b>32.48</b> <sup>±1.48</sup>	<u>91.83</u> <sup>±0.49</sup>	34.66 <sup>±1.21</sup>	91.50 <sup>±0.36</sup>	<b>27.81</b> <sup>±0.44</sup>	<b>93.26</b> <sup>±0.27</sup>
					16-shot					
CoOp <sub>MCM</sub>	30.26 <sup>±1.98</sup>	93.43 <sup>±0.81</sup>	34.69 <sup>±0.43</sup>	92.59 <sup>±0.14</sup>	41.91 <sup>±0.71</sup>	90.11 <sup>±0.23</sup>	44.68 <sup>±2.11</sup>	89.95 <sup>±0.48</sup>	37.89 <sup>±0.71</sup>	91.52 <sup>±0.29</sup>
CoOp <sub>GL</sub>	15.96 <sup>±1.67</sup>	96.11 <sup>±0.55</sup>	27.26 <sup>±1.99</sup>	93.29 <sup>±0.46</sup>	35.36 <sup>±2.08</sup>	90.58 <sup>±0.64</sup>	48.63 <sup>±2.11</sup>	86.11 <sup>±0.59</sup>	31.81 <sup>±1.27</sup>	91.51 <sup>±0.39</sup>
LoCoOp <sub>MCM</sub>	27.35 <sup>±3.19</sup>	94.12 <sup>±0.80</sup>	30.93 <sup>±1.19</sup>	93.75 <sup>±0.26</sup>	38.26 <sup>±1.53</sup>	91.12 <sup>±0.27</sup>	41.36 <sup>±2.56</sup>	90.99 <sup>±0.53</sup>	34.47 <sup>±0.73</sup>	92.49 <sup>±0.14</sup>
LoCoOp <sub>GL</sub>	18.46 <sup>±1.38</sup>	95.85 <sup>±0.63</sup>	<u>22.43</u> <sup>±0.96</sup>	<u>95.15</u> <sup>±0.22</sup>	31.53 <sup>±1.52</sup>	92.15 <sup>±0.22</sup>	43.35 <sup>±3.12</sup>	89.38 <sup>±0.78</sup>	28.94 <sup>±1.29</sup>	93.13 <sup>±0.17</sup>
IDLike	19.23 <sup>±10.5</sup>	<u>96.70</u> <sup>±1.58</sup>	54.15 <sup>±2.88</sup>	87.64 <sup>±1.19</sup>	56.63 <sup>±0.07</sup>	85.86 <sup>±0.44</sup>	34.69 <sup>±6.41</sup>	91.90 <sup>±2.31</sup>	41.18 <sup>±1.73</sup>	90.53 <sup>±0.01</sup>
LSN <sup>†</sup>	36.17 <sup>±4.81</sup>	92.66 <sup>±1.16</sup>	34.27 <sup>±0.44</sup>	93.53 <sup>±0.20</sup>	41.47 <sup>±0.85</sup>	90.52 <sup>±0.37</sup>	46.43 <sup>±0.60</sup>	89.38 <sup>±0.24</sup>	39.58 <sup>±0.73</sup>	91.53 <sup>±0.09</sup>
NegPrompt <sup>†</sup>	37.79 <sup>±0.11</sup>	90.49 <sup>±0.01</sup>	32.11 <sup>±3.77</sup>	92.25 <sup>±1.00</sup>	35.52 <sup>±0.41</sup>	91.16 <sup>±0.03</sup>	43.93 <sup>±9.09</sup>	88.38 <sup>±3.31</sup>	37.34 <sup>±1.41</sup>	90.57 <sup>±0.59</sup>
SCT <sub>MCM</sub>	29.41 <sup>±2.19</sup>	93.76 <sup>±0.55</sup>	27.28 <sup>±2.80</sup>	94.22 <sup>±0.40</sup>	36.35 <sup>±2.13</sup>	91.16 <sup>±0.33</sup>	42.25 <sup>±1.89</sup>	90.52 <sup>±0.48</sup>	33.82 <sup>±1.78</sup>	92.42 <sup>±0.35</sup>
SCT <sub>GL</sub>	15.19 <sup>±2.16</sup>	<b>96.71</b> <sup>±0.44</sup>	<b>20.00</b> <sup>±0.61</sup>	<b>95.57</b> <sup>±0.11</sup>	<b>29.71</b> <sup>±0.85</sup>	92.37 <sup>±0.12</sup>	44.17 <sup>±0.82</sup>	88.59 <sup>±0.39</sup>	<u>27.27</u> <sup>±0.44</sup>	93.31 <sup>±0.17</sup>
FA <sub>MCM</sub> (Ours)	25.79 <sup>±1.48</sup>	94.29 <sup>±0.35</sup>	33.54 <sup>±1.17</sup>	93.02 <sup>±0.19</sup>	33.77 <sup>±1.64</sup>	<b>92.64</b> <sup>±0.42</sup>	<b>23.17</b> <sup>±1.31</sup>	<b>95.14</b> <sup>±0.26</sup>	29.07 <sup>±1.11</sup>	<u>93.77</u> <sup>±0.19</sup>
FA <sub>GL</sub> (Ours)	<b>14.49</b> <sup>±1.27</sup>	96.48 <sup>±0.29</sup>	27.65 <sup>±1.08</sup>	93.46 <sup>±0.18</sup>	<u>31.09</u> <sup>±1.38</sup>	<u>92.44</u> <sup>±0.34</sup>	<u>29.50</u> <sup>±0.89</sup>	<u>92.93</u> <sup>±0.18</sup>	<b>25.68</b> <sup>±0.58</sup>	<b>93.82</b> <sup>±0.11</sup>

Table 1. Comparison results on ImageNet-1k OOD benchmarks. All results using the same backbone ViT-B/16. The results marked with <sup>†</sup> are taken from [48]; the others are our reproductions. The prompt learning based methods are run under four trials, reporting the mean and standard deviation of the performance. The subscripts <sub>MCM</sub> and <sub>GL</sub> indicate the use of the MCM score and the GL-MCM score. The best and second-best results are indicated in bold and underline. ↑ indicates larger values are better; ↓ indicates smaller values are better. All values are percentages.

Method	1-shot	4-shot	16-shot
CoOp	67.44 <sup>±0.50</sup>	69.71 <sup>±0.07</sup>	70.99 <sup>±0.14</sup>
LoCoOp	67.40 <sup>±0.64</sup>	69.55 <sup>±0.10</sup>	<u>71.53</u> <sup>±0.17</sup>
IDLike	68.17 <sup>±0.57</sup>	68.91 <sup>±0.14</sup>	69.46 <sup>±0.02</sup>
SCT	<u>68.63</u> <sup>±0.13</sup>	<u>69.93</u> <sup>±0.22</sup>	<b>71.78</b> <sup>±0.05</sup>
FA(Ours)	<b>68.67</b> <sup>±0.39</sup>	<b>69.96</b> <sup>±0.04</sup>	71.02 <sup>±0.09</sup>

Table 2. Comparison results in ID Top-1 accuracy on ImageNet-1k under different few-shot settings. Other notations are the same as Tab. 1.

ting are provided in the Appendix A). Specifically, in the 16-shot setting, FA with GL-MCM score (FA<sub>GL</sub>) outperforms the best baseline SCT<sub>GL</sub> (average FPR95 of 25.68% vs. 27.27%, and average AUROC of 93.82% vs. 93.31%).

Additionally, in the 1-shot setting, FA<sub>GL</sub> surpasses SCT<sub>GL</sub> by a large margin, showing improvements of at least 1.25% and 3.81% on average AUROC and FPR95. More importantly, even the average score of FA<sub>MCM</sub> has exceeded SCT<sub>GL</sub> in the 1-shot setting. Moreover, our FA effectively improves the OOD performance without sacrificing the ID classification capability of the model, as shown in Tab. 2.

In particular, we can also observe that our method performs not well on the SUN dataset, as shown in Tab. 1. This is related to the presence of a significant number of samples belonging to ID category in these datasets, as demonstrated by recent studies [1]. In other words, the SUN dataset in the popular OOD benchmark requires more thorough cleaning to be effectively used as an OOD dataset.

**Challenging OOD Detection.** Recent studies [1, 2] have

Method	OpenImage-O		NINCO		ImageNet-O		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CoOp <sub>MCM</sub>	37.95±0.59	91.99±0.22	78.00±0.23	73.00±0.59	70.14±0.86	81.34±0.83	62.03±0.38	82.11±0.48
CoOp <sub>GL</sub>	32.23±1.00	92.22±0.25	72.04±0.81	75.35±0.88	67.35±1.19	78.44±0.94	57.21±0.96	81.99±0.58
LoCoOp <sub>MCM</sub>	37.03±0.94	92.24±0.21	77.96±0.63	72.78±0.79	70.96±0.75	81.25±0.39	61.98±0.55	82.09±0.35
LoCoOp <sub>GL</sub>	33.87±1.23	92.53±0.07	75.16±0.92	72.97±0.39	68.06±1.92	81.19±0.25	59.03±1.27	82.23±0.17
IDLike	54.43±2.35	87.89±0.45	78.93±1.79	69.32±0.42	84.08±1.45	67.45±0.86	72.48±0.89	74.89±0.28
SCT <sub>MCM</sub>	37.28±1.16	92.04±0.33	78.51±0.86	71.09±1.19	71.45±0.54	81.29±0.15	62.41±0.73	81.48±0.52
SCT <sub>GL</sub>	32.24±0.34	92.75±0.17	74.15±0.53	73.14±0.94	68.35±0.62	80.84±0.45	58.25±0.39	82.24±0.44
FA <sub>MCM</sub> (Ours)	37.93±1.65	91.84±0.45	70.44±1.24	77.77±0.94	63.84±1.36	82.71±0.30	57.40±1.34	84.08±0.47
FA <sub>GL</sub> (Ours)	32.10±0.89	92.39±0.28	65.61±1.32	79.01±0.71	63.13±0.98	80.39±0.21	53.61±0.89	83.93±0.31

Table 3. Challenging OOD detection results on cleaner OOD datasets in the 16-shot setting. We use the same notation as Tab. 1.

ID Dataset	Method	iNaturalist		SUN		Places		Textures		Average	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
UCF101	CoOp <sub>MCM</sub>	14.40±6.61	97.41±1.36	37.09±3.94	93.14±0.48	40.07±2.77	91.89±0.65	33.24±4.21	93.19±1.22	31.21±3.58	93.91±0.45
	CoOp <sub>GL</sub>	9.65±6.35	98.18±0.98	28.54±5.07	94.51±0.92	32.94±1.74	93.28±0.32	35.96±2.65	91.83±1.42	26.77±3.22	94.45±0.42
	LoCoOp <sub>MCM</sub>	2.38±0.98	99.21±0.22	19.73±4.24	96.11±0.78	22.57±3.01	94.96±0.77	15.89±3.12	96.12±0.67	15.14±1.91	96.60±0.33
	LoCoOp <sub>GL</sub>	1.13±0.22	99.65±0.09	17.93±4.09	96.51±0.89	22.34±4.58	95.03±1.06	18.17±3.01	95.25±0.71	14.89±2.49	96.61±0.58
	IDLike	66.63±27.6	89.75±5.52	87.88±6.56	78.04±6.14	84.47±12.1	76.04±8.53	49.07±11.2	91.19±2.74	72.01±14.0	83.76±5.64
	SCT <sub>MCM</sub>	1.51±0.51	99.44±0.10	24.17±5.54	95.21±0.88	26.62±4.41	93.75±0.81	15.34±3.84	95.98±0.64	16.91±3.20	96.09±0.48
	SCT <sub>GL</sub>	0.79±0.27	99.75±0.04	18.69±2.41	95.86±0.56	23.77±2.47	93.98±0.47	17.44±2.62	94.52±0.54	15.17±1.56	96.03±0.25
	FA <sub>MCM</sub> (Ours)	0.12±0.11	99.94±0.01	3.06±1.33	99.34±0.23	4.68±1.00	99.08±0.26	2.68±0.51	99.31±0.43	2.63±0.58	99.42±0.22
	FA <sub>GL</sub> (Ours)	0.13±0.02	99.94±0.02	4.98±1.02	99.01±0.22	7.46±0.86	98.44±0.27	5.67±0.83	98.75±0.19	4.56±0.47	99.03±0.13
EuroSAT	CoOp <sub>MCM</sub>	99.14±0.73	31.95±9.17	98.29±0.49	44.36±4.19	97.28±0.85	47.51±4.63	86.69±4.37	69.17±3.15	95.35±1.03	48.25±3.09
	CoOp <sub>GL</sub>	92.45±4.26	57.59±10.4	90.97±3.49	61.89±4.83	88.69±3.85	64.00±4.92	62.37±1.93	83.58±2.31	83.62±1.18	66.77±2.83
	LoCoOp <sub>MCM</sub>	96.51±3.33	53.49±8.06	93.52±5.68	55.59±9.19	92.88±4.05	57.89±5.19	76.45±7.48	78.59±2.66	89.84±4.29	61.39±3.90
	LoCoOp <sub>GL</sub>	86.13±12.2	72.68±7.97	87.81±6.37	68.58±6.40	85.93±4.15	69.90±3.40	54.89±4.67	87.00±1.01	78.69±4.65	74.54±3.59
	IDLike	98.02±2.13	73.32±5.83	95.74±2.84	65.19±6.86	94.79±3.67	69.44±5.06	63.34±9.10	86.68±2.98	87.97±4.21	73.66±5.04
	SCT <sub>MCM</sub>	99.79±0.22	32.90±6.73	98.36±0.84	44.30±3.41	97.77±0.86	48.08±1.64	83.98±6.92	73.25±3.58	94.97±2.19	49.63±2.39
	SCT <sub>GL</sub>	96.57±5.41	58.21±5.31	92.41±4.55	61.54±3.47	90.49±5.02	63.33±2.71	55.29±14.7	85.48±3.68	83.69±7.02	67.14±2.20
	FA <sub>MCM</sub> (Ours)	85.24±6.88	78.62±3.18	36.24±10.8	92.13±2.44	28.17±10.2	94.05±1.68	10.36±3.97	97.71±0.69	39.99±7.79	90.63±1.91
	FA <sub>GL</sub> (Ours)	77.82±7.28	86.88±2.15	56.10±8.03	88.43±2.26	49.87±7.69	89.07±1.49	8.99±2.51	97.74±0.36	48.19±6.23	90.53±1.42

Table 4. Representative OOD detection results with various ID datasets in the 16-shot setting. We use the same notation as Tab. 1.

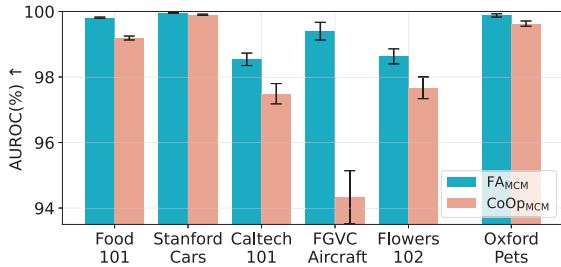


Figure 3. A portion of OOD detection average performance using other ID datasets under the 16-shot setting. We report the mean and standard deviation of the AUROC for our FA (blue) and CoOp (orange) using the MCM score. More details can be found in the Appendix B.

shown that some of the used OOD datasets contain more or less images of objects that belong to ID classes in ImageNet-1k [7]. Hence, probing the true performance of OOD detectors for ImageNet-1k demands OOD datasets that are both challenging and genuinely OOD. As shown in Tab. 3, to fully substantiate the effectiveness of our

FA, we leverage 3 additional cleaner and more challenging OOD datasets, including OpenImage-O [45], NINCO [2], and ImageNet-O [15], inspired by [2]. Notably, both our FA<sub>MCM</sub> and FA<sub>GL</sub> consistently outperform current SOTA methods in terms of average FPR95 and AUROC under different few-shot settings (More details under the 1-shot and 4-shot settings are provided in the Appendix A).

**Various ID Datasets.** Considering factors like dataset variety, image resolution, and the number of classes [30], we also leverage 8 additional widely used datasets for few-shot setting as ID datasets for comparison [30, 34, 54]. Overall, despite our model is simple, it achieves superior average OOD performance on these ID datasets compared to other methods. We selected representative results using UCF101 [40] and EuroSAT [13] as ID datasets in the 16-shot setting, as shown in Tab. 4. Since using the other datasets (*i.e.* Food101 [3], StanfordCars [21], Caltech101 [12], FGVC Aircraft [29], Flowers102 [35], and OxfordPets [36]) as ID datasets for OOD detection presents relatively lower difficulty as shown in Fig. 3, we leave more

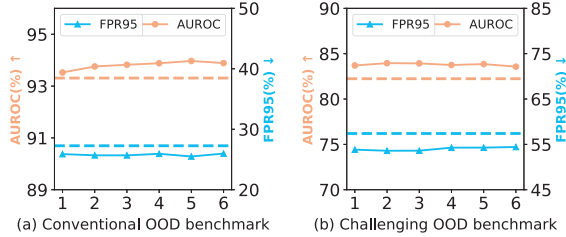


Figure 4. Sensitivity study of hyperparameter  $K$ . We report the average FPR95 (Blue) and AUROC (Orange) of our  $FA_{GL}$  on different OOD benchmarks using ImageNet-1k as the ID dataset under 16-shot setting. Dashed lines represent the performance of the best baseline (*i.e.*  $SCT_{GL}$ ).

experimental details in Appendix B.

In particular, for the ID dataset EuroSAT,  $FA_{MCM}$  surpasses the best-competing method  $LoCoOp_{GL}$  by 16.09% in average AUROC and 38.7% in average FPR95, underscoring its effectiveness and versatility even when the ID dataset is of low resolution (64x64 pixels).

### 5.3. Ablation study

**Influence of the  $\mathcal{L}_{FCE-K}$ .** As shown in Tab. 5, we trained the model  $FA_{CE}$  and  $FA_{FCE-K}$  using the standard cross-entropy loss  $\mathcal{L}_{CE}$  and our  $\mathcal{L}_{FCE-K}$ , respectively. Our  $FA_{FCE-K}$  significantly outperforms  $FA_{CE}$  under the same inference setting, which demonstrate the effectiveness of our  $\mathcal{L}_{FCE-K}$ .

**Initialization of the forced and original prompt.** As shown in Tab. 6, we perform experiments comparing different initialization scenarios for both prompts. Manual initialization uses the embeddings of “a photo of a [class-c]” to initialize the prompt’s embeddings, while the random initialization setting is based on CoOp [54]. The results show that employing consistent manual initialization for both prompts enables the forced prompt to effectively learn more specific descriptions of the ID classes.

**Influence of the shared learnable vector.** We also explore the influence of the shared learnable vector and the independent learnable vector for the forced prompts. As shown in Tab. 7, we can observe that using the shared learnable vector performs well in the few-shot setting since the independent learnable vector for each class has more parameters and requires more training data [54].

### 5.4. Sensitivity study

As shown in Fig. 4, we explore the influence of the hyperparameter  $K$ . Overall, with  $K$  ranging from 1 to 6, the results indicate that our proposed framework is insensitive to the choice of  $K$ . Concretely, our method is stable and outperforms the best baseline  $SCT_{GL}$  in average AUROC and FPR95 on both conventional and challenging OOD detection benchmarks. Specially, as  $K$  increases, the average AUROC shows an upward trend and then tends to be sta-

ble (see Fig. 4 (a)), showing a bottleneck in capturing more comprehensive semantic representations of the ID classes.

Method	MCM		GL-MCM	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑
$FA_{CE}$	41.43 <sup>±3.11</sup>	91.01 <sup>±0.67</sup>	34.29 <sup>±1.94</sup>	90.99 <sup>±0.56</sup>
$FA_{FCE-K}$	29.07 <sup>±1.11</sup>	93.77 <sup>±0.19</sup>	<b>25.68<sup>±0.58</sup></b>	<b>93.82<sup>±0.11</sup></b>

Table 5. Influence of the  $\mathcal{L}_{FCE-K}$ . MCM and GL-MCM refer to the score functions used during model inference.

Forced-M-Init	Original-M-Init	$FA_{MCM}$		$FA_{GL}$	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
-	-	40.71 <sup>±4.00</sup>	91.28 <sup>±0.82</sup>	34.88 <sup>±2.92</sup>	92.12 <sup>±0.76</sup>
✓	-	40.81 <sup>±2.25</sup>	91.25 <sup>±0.64</sup>	36.03 <sup>±2.84</sup>	91.62 <sup>±1.01</sup>
-	✓	31.35 <sup>±1.34</sup>	93.35 <sup>±0.34</sup>	26.39 <sup>±1.02</sup>	93.64 <sup>±0.19</sup>
✓	✓	29.07 <sup>±1.11</sup>	93.77 <sup>±0.19</sup>	<b>25.68<sup>±0.58</sup></b>	<b>93.82<sup>±0.11</sup></b>

Table 6. Ablation study of different initialization scenarios for the forced and original prompts. “M-Init” represents manual initialization. ✓ represents using manual initialization; - represents using random initialization.

Shared-Vector	M-Init	$FA_{MCM}$		$FA_{GL}$	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑
-	-	44.02 <sup>±2.24</sup>	90.17 <sup>±0.42</sup>	46.29 <sup>±1.33</sup>	88.24 <sup>±0.41</sup>
✓	-	40.71 <sup>±4.00</sup>	91.28 <sup>±0.82</sup>	34.88 <sup>±2.92</sup>	92.12 <sup>±0.76</sup>
-	✓	34.69 <sup>±0.62</sup>	92.57 <sup>±0.12</sup>	29.27 <sup>±0.55</sup>	92.82 <sup>±0.80</sup>
✓	✓	29.07 <sup>±1.11</sup>	93.77 <sup>±0.19</sup>	<b>25.68<sup>±0.58</sup></b>	<b>93.82<sup>±0.11</sup></b>

Table 7. Ablation study of the shared learnable vector. ✓ / - for “Shared-vector” represents using the shared / independent learnable vector for the forced prompt; ✓ / - for “M-Init” represents using manual / random initialization for the forced prompt.

## 6. Conclusion

In this study, we propose a novel CLIP-based framework for OOD detection based on Forced prompt learning (FA), which focus on fully exploiting the ID knowledge to effectively improve OOD detection without exploring complex OOD-related knowledge. We introduce a learnable forced prompt in addition to the frozen original prompt, both of which are using the same manual initialization. The forced prompt treats the original prompt as a reference, forcing itself to learn more diversified semantic descriptions of the ID classes rather than being limited to the textual semantics of class labels. We also introduce a forced coefficient to facilitate the forced prompt in learning more nuanced descriptions of the ID classes. Comprehensive experimental evaluations demonstrate that our method consistently surpasses current state-of-the-art methods on diverse benchmarks. We expect that our study can bring new insight on VLMs-based OOD detection methods and inspire more future research.

## 7. Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (grant No. 62071502), the Major Key Project of PCL (grant No. PCL2023A09), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

## References

- [1] Yichen Bai, Zongbo Han, Bing Cao, Xiaoheng Jiang, Qinghua Hu, and Changqing Zhang. Id-like prompt learning for few-shot out-of-distribution detection. In *CVPR*, pages 17480–17489, 2024. 2, 3, 5, 6
- [2] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, pages 2471–2506, 2023. 5, 6, 7
- [3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 5, 7, 11
- [4] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: prompt learning with optimal transport for vision-language models. In *ICLR*, 2023. 2
- [5] Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves OOD detection with pre-trained vision-language models. In *NeurIPS*, 2024. 2
- [6] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 5
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5, 7
- [8] 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. 5
- [9] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. VOS: learning what you don’t know by virtual outlier synthesis. In *ICLR*, 2022. 2
- [10] Xuefeng Du, Yiyun Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *NeurIPS*, 36:60878–60901, 2023. 2
- [11] Sepideh Esmailpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI*, pages 6568–6576, 2022. 2
- [12] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, page 178, 2004. 5, 7, 11
- [13] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 7
- [14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 2, 5
- [15] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 5, 7
- [16] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, pages 8759–8773, 2022. 1, 2, 5
- [17] Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 5
- [18] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, pages 10948–10957, 2020. 2
- [19] Rui Huang and Yixuan Li. MOS: towards scaling out-of-distribution detection for large semantic space. In *CVPR*, pages 8710–8719, 2021. 5
- [20] 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. In *ICLR*, 2024. 2
- [21] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. 5, 7, 11
- [22] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, pages 7167–7177, 2018. 2
- [23] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *CVPR*, pages 17584–17594, 2024. 2, 3, 5
- [24] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 1, 2, 5
- [25] Jiashuo Liu, Zheyang Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 1
- [26] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9): 1–35, 2023. 2
- [27] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 2, 5
- [28] Shuo Lu, Yingsheng Wang, Lijun Sheng, Aihua Zheng, Lingxiao He, and Jian Liang. Recent advances in ood detection: Problems and approaches. *arXiv preprint arXiv:2409.11884*, 2024. 1

- [29] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. [5](#), [7](#), [11](#)
- [30] 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*, 2022. [1](#), [3](#), [5](#), [7](#)
- [31] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *ICLR*, 2023. [2](#)
- [32] 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. [1](#), [3](#), [5](#)
- [33] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *NeurIPS*, 2023. [2](#), [3](#), [5](#)
- [34] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *ICLR*, 2024. [2](#), [3](#), [5](#), [7](#)
- [35] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *In 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. [5](#), [7](#), [11](#)
- [36] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. [5](#), [7](#), [11](#)
- [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. In *ICML*, pages 8748–8763, 2021. [1](#), [2](#)
- [38] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *arXiv preprint arXiv:2110.14051*, 2021. [1](#)
- [39] Vikash Sehwal, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *ICLR*, 2021. [2](#)
- [40] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. [5](#), [7](#)
- [41] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, pages 144–157, 2021. [1](#), [2](#), [5](#)
- [42] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, pages 20827–20840, 2022. [2](#)
- [43] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICLR*, 2023. [2](#)
- [44] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022. [2](#)
- [45] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pages 4911–4920, 2022. [5](#), [7](#)
- [46] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In *ICCV*, pages 1802–1812, 2023. [2](#), [5](#)
- [47] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. [5](#)
- [48] Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. In *NeurIPS*, 2024. [2](#), [3](#), [5](#), [6](#)
- [49] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *CVPR*, pages 10899–10909, 2023. [5](#)
- [50] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *CVPR*, pages 15701–15711, 2023. [2](#)
- [51] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kun-chang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of CLIP for few-shot classification. In *ECCV*, pages 493–510, 2022. [5](#)
- [52] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2018. [5](#)
- [53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16795–16804, 2022. [2](#)
- [54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. [2](#), [3](#), [4](#), [5](#), [7](#), [8](#)
- [55] Yibo Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *CVPR*, pages 7379–7387, 2022. [2](#)