Prompt Learning with Text-Augmented Cues for Out-of-Distribution Detection

Mingxu Feng, Dian Chao, Yuxuan Zhang, Yang Yang†, Weili Guo†
Nanjing University of Science and Technology
Nanjing, China
{mingxuf, chaodian, xuan_yuzhang, yyang, wlguo}@njust.edu.cn

***Abstract*—With the development of foundation models, advanced research explores the potential of vision-language models (VLMs) for out-of-distribution (OOD) detection. In particular, methods that generate outliers for regularized prompt learning have shown promising results in few-shot settings. However, existing methods that rely solely on visual modalities struggle to synthesize outliers semantically analogous to in-distribution (ID) data, neglecting near OOD scenarios. Notably, recent large language models (LLMs) exhibit a deep real-world understanding, enabling the generation of near textual outliers. Inspired by this, we propose an OOD detection framework named Text-Augmented Cues (TAC), which integrates expert knowledge from LLMs into the prompt learning of VLMs. Specifically, we first design LLMs query templates to generate outlier categories based on *Visual Similarity*. Then, LLMs are further leveraged to synthesize the semantic representations of ID and outlier categories based on *Feature Distinctiveness*. Subsequently, we incorporate visual-textual information of ID categories for prompt learning, regularized by textual outliers. Experimental results demonstrate that TAC significantly outperforms state-of-the-art (SOTA) VLM-based OOD detection methods in few-shot scenarios. The code is available at https://github.com/njustkmg/ICDM25-TAC.**

***Index Terms*—out-of-distribution detection, vision-language models, prompt learning, large language models.**

I. Introduction

Deep learning models perform well in a closed-world setting, assuming that both training and testing samples come from the same distribution. However, models fail to generalize effectively to out-of-distribution (OOD) data when deployed in open-world scenarios [1], [2], [3], [4]. This phenomenon may lead to severe consequences, particularly in critical domains such as autonomous driving [5], [6] and medical diagnosis [7]. Consequently, effective OOD detection is essential to preserve model reliability and ensure trustworthy decision-making.

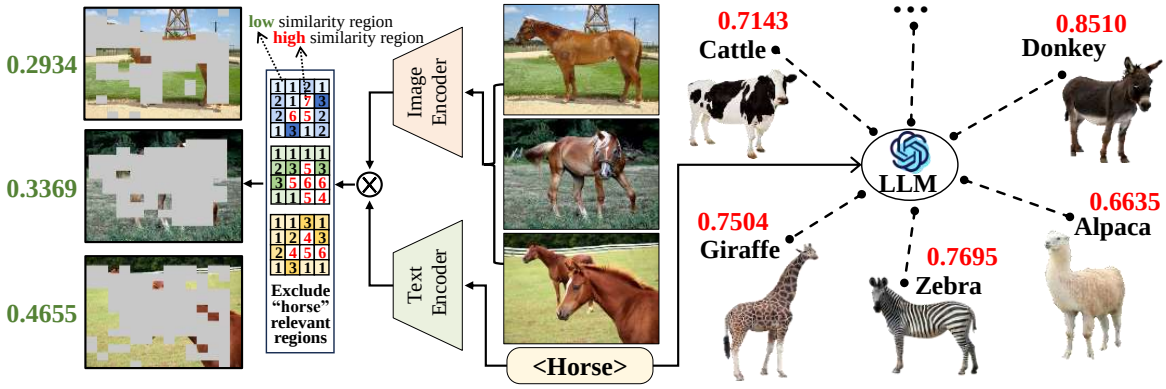
With the emergence of vision-language models (VLMs) renowned for their remarkable generalization across diverse tasks [8], [9], [10], recent studies have increasingly leveraged these models for few-shot OOD detection [11], [12]. In particular, several methods excel by generating visual outliers for regularized prompt learning [13], [14], [15], demonstrating superior detection efficacy. These methods generate outliers by exploiting redundant background cues in in-distribution (ID) visual data. However, such background information typically differs significantly from the semantic features of ID categories. As illustrated in Fig. 1 (a), the foreground depicts a horse, whereas the background consists of trees or grass, which are inherently easier for the model to distinguish. Consequently, these visual outliers generally attain lower predictive probabilities for ID classes. This leads to these methods primarily emphasizing the integration of far OOD knowledge while neglecting near OOD scenarios. Indeed, near OOD represents the most challenging aspect of OOD detection, as it necessitates distinguishing subtle semantic shifts that are more prone to misleading the model.

Recent research suggests that selecting more challenging outlier samples facilitates the model in establishing a more robust decision boundary between ID and OOD data [16], [17]. Meanwhile, pre-trained generative models, such as GAN [18] and diffusion model [19], have been explored for their potential in generating challenging outliers [20], [21], [22]. While these approaches perform well on small datasets, they become computationally expensive and resource-intensive when applied to large-scale datasets. Furthermore, the reliance on pre-trained models introduces inductive bias that leads to the generation of less diverse outliers, which hinders the model's ability to capture the full spectrum of real-world scenarios.

Leveraging the alignment capabilities of VLMs, a certain interchangeability exists between textual and visual data within their feature spaces. This naturally inspires our approach: *given the difficulty of generating challenging visual outliers, might challenging textual outliers be synthesized to enhance VLMs regularization training for OOD detection?* Recent advancements in large language models (LLMs) [23], [24], characterized by extensive pretraining on large datasets and autoregressive objectives, have endowed them with superior generative capabilities and a profound understanding of the real world [25], [26], [27]. This enables LLMs to generate high-quality outliers with nuanced semantic variations from ID categories. Fig. 1 (b) illustrates outliers generated by the LLMs for the horse category, exhibiting high predictive probabilities for ID class, indicating the model's diminished capacity to distinguish these outliers from horse instances.

Textual data not only addresses the challenge of generating near OOD outliers but also compensates for the limited number of ID visual data. Specifically, VLM-based prompt learning is often constrained in few-shot scenarios [13], where limited visual data impedes effective learning of category-specific features. This limitation leads to spurious correlations between category labels and background information,

† Corresponding author



(a). Visual Outliers Generation (b). Textual Outliers Generation

Fig. 1: Illustration of two types of outlier generation. The original images are from ImageNet-1k. Visual outliers generated using CLIP exhibit a significant semantic shift from the ID label "horse" and have a low prediction probability (left green numbers). In contrast, outlier instances in the text modality, generated by our method, demonstrate a subtle semantic shift and have a high prediction probability (right red numbers) for the ID label "horse".

undermining the model's robustness in downstream tasks and leading to low-confidence predictions for ID categories [28]. Additionally, expanding ID datasets by incorporating extensive image annotations incurs significant manual labeling costs. By leveraging LLMs, we can efficiently generate textual semantic information for ID categories to substitute visual data, guiding VLMs to focus on the intrinsic features of categories.

Building upon the observed insights, we propose a novel prompt learning framework named Text-Augmented Cues (TAC), which harnesses textual semantic information to refine ID and OOD representation learning. At the high level, TAC leverages the comprehensive understanding of the real world inherent in LLMs to inject authentic textual information into VLMs. Specifically, we first design query templates for LLMs based on the principle of *Visual Similarity*, enabling LLMs to identify outlier categories that share visual characteristics with the ID categories. Then, LLMs are further employed to synthesize the semantic representations of ID and outlier categories, guided by the principle of *Feature Distinctiveness*. It ensures that the generated semantic representations of the categories are well-defined and distinct from each other. Subsequently, TAC incorporates visual-textual information of ID categories for prompt learning, regularized by textual outliers, facilitating a more robust decision boundary. The main contributions of this work are summarized as follows:

- **Principally**, we emphasize the limitations of VLM-based prompt learning for OOD detection that rely solely on visual modalities and propose leveraging the expertise of LLMs to guide VLMs.
- **Technically**, we propose Text-Augmented Cues (TAC), a framework utilizing LLMs to generate challenging outliers, combined with a collaborative training strategy for ID-OOD differentiation.
- **Empirically**, TAC achieves **2.04%** and **1.90%** improvements on the far and near OOD benchmarks in FPR95 (in Section V-A) respectively, validating its effectiveness.

II. Related Work

A. Out-of-distribution detection

Traditional OOD detection grounded in the MSP [29] score as a baseline has been extended through various scoring methods, including Energy [30] scores and MaxLogit [31] scores, which were built on a single modality (e.g. ResNet [32] as backbone) to improve detection performance. With the remarkable performance of VLMs in downstream tasks gaining widespread attention, several studies have explored their application in enhancing OOD detection. MCM [33] leverages maximum concept matching scores between image and texture features based on CLIP [16] for OOD detection during inference. CLIPN [34] trains a negative text encoder using annotated auxiliary datasets, enabling CLIP to better understand the concept of "no" class prompts. Similarly, ZOC [35] develops a text decoder trained to generate candidate sets of unknown classes for images. EOE [36] leverages LLMs to generate potential OOD categories, mitigating the constraints of closed-set labels on the discrimination capabilities of CLIP. However, these methods mainly address zero-shot scenarios, leaving OOD detection in a few-shot setting underexplored. For this purpose, ID-like [13] and LoCoOp [14] adopt prompt learning techniques, extracting ID-irrelevant background features from ID images as visual outliers to regularize training. SCT [15] further introduces an adaptive modulation factor calibration strategy to address the issue of inaccurate foreground-background separation faced by these prompt learning methods. Additionally, NegPrompt [11] and LSN [12] rely exclusively on ID images to train a set of negative prompts that capture the non-class-related semantics.

B. Foundation Models

Foundation models, pre-trained on large-scale and diverse datasets, demonstrate remarkable performance across a wide range of downstream tasks. LLMs built on the Transformer architecture [37] and autoregressive training have achieved

breakthroughs in contextual understanding and text generation in natural language processing. In particular, GPT-4 [38] utilizes a Mixture-of-Experts (MoE) architecture and multimodal capabilities, setting new standards in text generation and cross-modal reasoning. LLa-MA [23] excels in inference through efficient training and low-bit quantization, while the Claude [39] series enhances alignment capabilities through constitutional AI. Deep-Seek-V3 [24] built on 14.8T tokens of pre-training data with an MoE architecture, reaches benchmark levels comparable to GPT-4. Meanwhile, VLMs leverage contrastive learning to effectively align visual and textual modalities, enabling cross-modal reasoning and representation learning. CLIP [8], FILIP [10], and ALIGN [9] employ contrastive loss to align text and image representations within a shared feature space. These methods adopt a dual-stream architecture to extract text and image features separately, maximizing the similarity of matching pairs and reducing their feature distance. Notably, CLIP trained on 400 million image-text pairs, demonstrates exceptional performance across multiple computer-vision benchmarks.

III. Preliminaries

A. Problem Definition

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ denote the ID training dataset, where $\boldsymbol{x}_i \in \mathcal{X}$ represents the image and $y_i \in \mathcal{Y}$ is its corresponding label. Additionally, we have an unlabeled outlier dataset $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{M}$ for regularized training. During testing, we encounter samples from both ID test set $\tilde{\mathcal{X}}$ and OOD test set $\mathcal{X}_{\text{out}}$, where the label space of $\tilde{\mathcal{X}}$ is consistent with $\mathcal{Y}$, while the label space of $\mathcal{X}_{\text{out}}$ is disjoint from $\mathcal{Y}$. Our objective is to train a robust classifier that accurately classifies ID categories while determining whether a test sample belongs to the label space $\mathcal{Y}$, leveraging the training datasets $\mathcal{D}$ and $\mathcal{Z}$.

B. Vanilla Prompt Learning Framework

In vanilla prompt learning, we employ a dual-encoder architecture consisting of a visual encoder and a text encoder. For a given ID image $\boldsymbol{x} \in \mathcal{X}$, the visual encoder $\mathcal{I}(\cdot)$ extracts its feature embedding $\mathcal{I}(\boldsymbol{x}) \in \mathbb{R}^d$, where d denotes the feature dimension. The text encoder $\mathcal{T}(\cdot)$ processes a prompt template $\mathbf{p}_i$, which consists of learnable context vectors $\boldsymbol{\omega}$ and the i-th class label y_i to generate the class-specific prompt embedding. The prediction probability of i-th class is formulated as:

$$p(y_i|\boldsymbol{x};\boldsymbol{\omega}) = \frac{\exp(\text{sim}(\mathcal{I}(\boldsymbol{x}), \mathcal{T}(\mathbf{p}_i))/\tau)}{\sum_{j=1}^{c} \exp(\text{sim}(\mathcal{I}(\boldsymbol{x}), \mathcal{T}(\mathbf{p}_j))/\tau)}, \quad (1)$$

where $\text{sim}(\cdot)$ denotes the cosine similarity function measuring feature alignment, $\tau > 0$ is a temperature scaling factor controlling the prediction sharpness, and c denotes the total number of classes in the training set.

C. Training Objective

During training, the model leverages the ID dataset $\mathcal{D}$ and the outlier dataset $\mathcal{Z}$ to learn the data distribution and establish decision boundaries. For ID samples $(\boldsymbol{x}_i, y_i) \in \mathcal{D}$, the model maximizes the prediction confidence for the correct label:

$$\arg\max_{y \in \mathcal{Y}} p(y|\boldsymbol{x}_i;\boldsymbol{\omega}) = y_i. \quad (2)$$

For each outlier sample $\boldsymbol{z}_i \in \mathcal{Z}$, the model minimizes the prediction confidence across all classes:

$$\max_{y \in \mathcal{Y}} p(y|\boldsymbol{z}_i;\boldsymbol{\omega}) \approx \frac{1}{c}. \quad (3)$$

This training strategy ensures effective discrimination between ID and OOD data, establishing a foundation for subsequent OOD detection tasks.

IV. Methodology

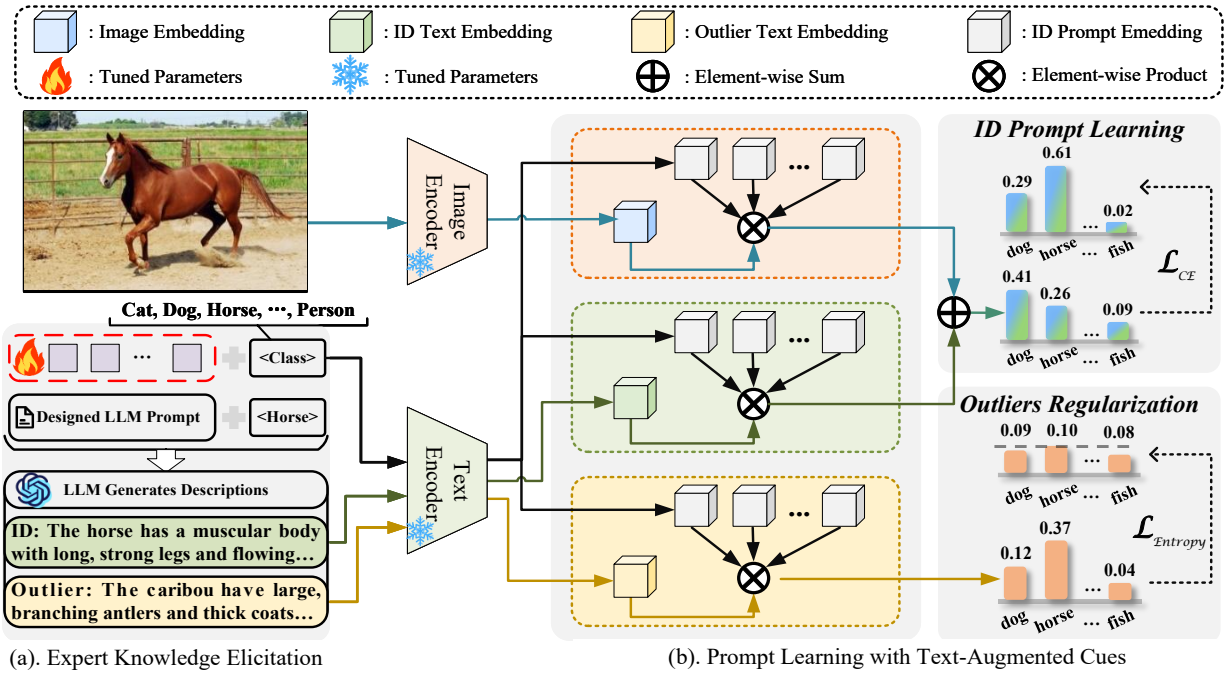
In this work, we focus on leveraging LLMs to generate representations in the text modality, which assist in training the ID prompts and enhance OOD detection performance. During this research, we encounter two primary challenges: 1) How to effectively utilize the capabilities of LLMs to generate ID and outlier representations? 2) How to design a training process that enables the model to distinguish the distribution between ID and OOD data? To address these challenges, we develop two key strategies: an effective semantic representation generation approach and a collaborative training framework that bridges modality and category distributions. The framework of our proposed method is depicted in Fig. 2.

A. Semantic Representation Acquisition

We propose a multi-stage approach that harnesses the generative capacity of LLMs to synthesize semantic descriptions. Through the meticulous design of prompts emphasizing both *Visual Similarity* and *Feature Distinctiveness*, we generate informative outlier labels and extract categorical features that capture nuanced semantic distinctions.

Outlier Label Acquisition. We initially design the LLMs query template based on the *Visual Similarity* principle to obtain the outlier labels set $\mathcal{Y}_{out}$. The principle aims to identify categories that are visually similar to known classes but semantically distinct. The LLMs query template, illustrated in Fig. 3, where $\mathcal{Y}$ indicates the set of ID class labels and $y_i \in \mathcal{Y}$. In this prompt, *visual resemblance* emphasizes perceptual similarity, while *taxonomically distinct* underscores semantic divergence. We then request the LLMs to provide an outlier labels set $\mathcal{Y}_{out}$ that meets the criteria.

VLM-based Similarity Filtering. During the acquisition process, we observe a few outlier labels semantically misaligned with their native classes yet congruent with other ID labels. For instance, LLMs generate "cargo ship" as an outlier label for "warplane", despite its semantic proximity to the ID class "container ship". Therefore, we implement a semantic similarity threshold filtering method, as illustrated in Fig. 3. Specifically, for each $\hat{y}_i \in \mathcal{Y}_{out}$, we compute similarities s_{ij} with all $y_j \in \mathcal{Y}$ using the VLM, and obtain the maximum similarity as its value $\tilde{s}_i$. We then obtain the similarity set $\tilde{S}$ between the outlier labels and the ID label set. δ is the predefined threshold. Outlier labels with $\tilde{s}_i > \delta$ are removed, while those with $\tilde{s}_i \leq \delta$ are retained. In the experiments, the



(a). Expert Knowledge Elicitation (b). Prompt Learning with Text-Augmented Cues

Fig. 2: The framework of TAC. It comprises two phases: (a). Expert Knowledge Elicitation: Leveraging designed LLMs query templates guided by *Visual Similarity* and *Feature Distinctiveness*, we extract discriminative textual descriptors for both ID and potential outliers through knowledge distillation from LLMs. (b). Prompt Learning with Text-Augmented Cues: We integrate visual-textual multimodal data for ID prompt learning, regularized by textual outliers.

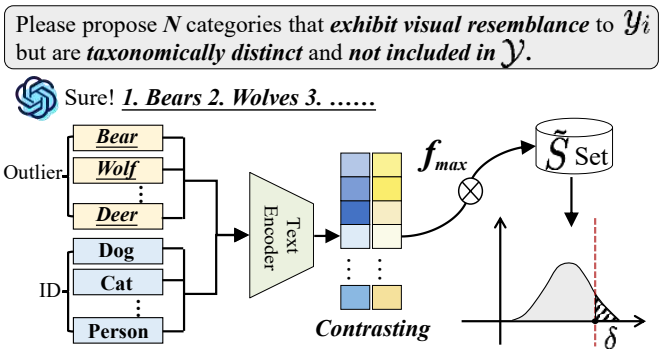


Fig. 3: LLMs query template for outlier label acquisition and VLM-based similarity filtering process.

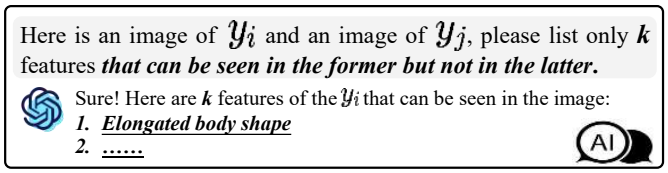


Fig. 4: LLMs query template for semantic representation generation.

retained label similarities generally ranging from 0.45 to 0.85 demonstrate the semantic diversity of the generated content.

Semantic Representation Generation. In the domain of feature-driven LLMs textual description generation, we propose a strategy based on *Feature Distinctiveness* principles, aimed at enhancing data generation representativeness through the precise capture of category-distinguishing information. Specifically, for each ID and outlier label, we design an LLM query template paradigm that emphasizes visual feature disparities between categories. During the generation process, we guide the LLMs to focus on critical features exclusively present within the target category.

As illustrated in Fig. 4, when generating ID textual descriptions, we select $y_i \in \mathcal{Y}$ and $y_j \in \mathcal{Y}_{out}$ to get the ID text dataset $\boldsymbol{T}_{\text{in}}$. Conversely, when generating outlier descriptions, we reverse their roles to get the outlier text dataset $\boldsymbol{T}_{\text{out}}$. By presenting *k features that can be seen in the former but not in the latter* within the prompt, we motivate the LLMs to generate text descriptions precisely centered on the distinctive visual characteristics of y_i. This feature-driven approach precisely distills the categorical essence, refines semantic boundaries, and enhances the discriminative fidelity of the generated textual representations.

B. Prompt Learning with Text-Augmented Cues

We propose a novel OOD detection framework that incorporates visual-textual information of ID categories for prompt learning, regularized by textual outliers. Building on the previous section, we incorporate the ID text dataset $\boldsymbol{T}_{\text{in}}$ and outlier text dataset $\boldsymbol{T}_{\text{out}}$ generated by LLMs as input, along with the ID image dataset $\mathcal{D}$. By meticulously engineering the loss function, this framework aims to optimize the model's ability to effectively distinguish between ID and OOD samples.

ID Feature Fusion. Let $\mathbf{p}$ represent the ID class prompt, and $\boldsymbol{t}_{\text{in}} \in \boldsymbol{T}_{\text{in}}$. Given an input image $\boldsymbol{x} \in \mathcal{X}$, we define a

semantic similarity score S that dynamically fuses features from different modalities:

$$S = \alpha \cdot \text{sim}(\mathcal{I}(\boldsymbol{x}), \mathcal{T}(\mathbf{p})) + (1-\alpha) \cdot \text{sim}(\mathcal{T}(\boldsymbol{t}_{\text{in}}), \mathcal{T}(\mathbf{p})), \quad (4)$$

where $\alpha \in [0,1]$ is an adjustable hyperparameter, $\mathcal{I}(\boldsymbol{x})$ is the image feature embedding, $\mathcal{T}(\mathbf{p})$ represents the ID prompt feature embedding, $\mathcal{T}(\boldsymbol{t}_{\text{in}})$ represents the text encoder output for the ID text description.

The semantic similarity score S quantifies multi-modal semantic alignment via a weighted similarity measure across image, text, and ID prompt embeddings. A weight parameter α is strategically introduced in S to mitigate potential modal bias in training, consequently enabling refined cross-modal representation learning.

ID Loss Function. We employ a standard cross-entropy loss function to measure the discrepancy between predicted probabilities and ground truth labels for ID samples. By minimizing this loss, we guide the model's parameter optimization, enhancing the discriminative capability across different classes. Formally, the cross-entropy loss $\mathcal{L}_{\text{CE}}$ is computed based on the semantic similarity score S, defined as follows:

$$\mathcal{L}_{\text{CE}} = \mathbb{E}_{[(\boldsymbol{x},y)\sim\mathcal{D},\boldsymbol{t}_{\text{in}}\sim\boldsymbol{T}_{\text{in}}]}\left[-\log\frac{\exp(S_{x,y}/\tau)}{\sum_{j=1}^{c}\exp(S_j/\tau)}\right], \quad (5)$$

where c represents the number of classes and $\tau > 0$ is the temperature scaling factor that controls the sharpness of probability distributions. The term $S_{x,y}$ indicates the semantic similarity score for the sample $\boldsymbol{x}$ and its corresponding ground-truth label y.

Outlier Loss Function. Given the outlier dataset $\boldsymbol{T}_{\text{out}}$, each outlier text feature should exhibit minimal semantic similarity with any ID prompt embedding. We thus introduce an entropy-based loss $\mathcal{L}_{\text{Entropy}}$ for OOD regularization:

$$\mathcal{L}_{\text{Entropy}} = \mathbb{E}_{\boldsymbol{t}_{\text{out}}\sim\boldsymbol{T}_{\text{out}}}\left[-\sum_{i=1}^{c} p(y_i|\boldsymbol{t}_{\text{out}};\boldsymbol{\omega})\log p(y_i|\boldsymbol{t}_{\text{out}};\boldsymbol{\omega})\right], \quad (6)$$

where $\boldsymbol{t}_{\text{out}}$ represents outlier text samples drawn from the dataset $\boldsymbol{T}_{\text{out}}$, and $p(y_i|\boldsymbol{t}_{\text{out}};\boldsymbol{\omega})$ denotes the predicted class probability for the i-th class given the outlier text and the learnable context vectors $\boldsymbol{\omega}$.

The entropy objective aims to enhance the semantic uncertainty of outlier feature representations. By minimizing the negative entropy, the loss function seeks to expand the semantic divergence between outlier features and ID prompt embeddings, thereby suppressing the model's overconfidence when confronted with OOD samples.

The overall loss function combines ID classification and OOD discrimination:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{Entropy}}, \quad (7)$$

where $\lambda \in \mathbb{R}^+$ is a hyperparameter controlling the weight of the OOD regularization term, thereby modulating the balance between accurately capturing ID features and enhancing the model's robustness to OOD samples.

Algorithm 1 Prompt Learning with Text-Augmented Cues

Input: Training dataset $\mathcal{D}$, ID class label $\mathcal{Y}$, ID text dataset $\boldsymbol{T}_{in}$, Outlier text dataset $\boldsymbol{T}_{out}$, Image encoder $\mathcal{I}(\cdot)$, Text encoder $\mathcal{T}(\cdot)$, ID Feature fusion weight α, regularization term weight λ, Training epochs M, Batch size n, Learning rate η, Logit scale τ

Output: Optimized prompt $\boldsymbol{\omega}$

Initialize learnable prompt $\boldsymbol{\omega}$
for epoch = 1 **to** M **do**
  for batch = 1 **to** N **do**
    $\{\boldsymbol{h}_j\}_{j=1}^{c} \leftarrow \{y_j \in \mathcal{Y} | \mathcal{T}(\boldsymbol{\omega}, y_j)\}_{j=1}^{c}$
    Sample a batch $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ from $\mathcal{D}$ and $\{\boldsymbol{t}_i\}_{i=1}^{n}$ from $\boldsymbol{T}_{in}$
    $\boldsymbol{f}_i \leftarrow \mathcal{I}(\boldsymbol{x}_i)$, $\boldsymbol{g}_i^{in} \leftarrow \mathcal{T}(\boldsymbol{t}_i)$
    $S_{i,j} \leftarrow \alpha \cdot \frac{\boldsymbol{f}_i \cdot \boldsymbol{h}_j}{\|\boldsymbol{f}_i\| \cdot \|\boldsymbol{h}_j\|} + (1-\alpha) \cdot \frac{\boldsymbol{g}_i^{in} \cdot \boldsymbol{h}_j}{\|\boldsymbol{g}_i^{in}\| \cdot \|\boldsymbol{h}_j\|}$
    $\mathcal{L}_{\text{CE}} \leftarrow -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp(S_{x,y}/\tau)}{\sum_{j=1}^{c}\exp(S_{i,j}/\tau)}$
    Sample a batch $\{\hat{\boldsymbol{t}}_i\}_{i=1}^{n}$ from $\boldsymbol{T}_{out}$
    $\boldsymbol{g}_i^{out} \leftarrow \mathcal{T}(\hat{\boldsymbol{t}}_i)$
    Compute $\mathcal{L}_{\text{Entropy}}$ by (6)
    Update: $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \eta\nabla_{\boldsymbol{\omega}}(\mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{Entropy}})$
  end for
end for

Algorithm 1 presents the procedure for prompt learning with Text-Augmented Cues.

C. Test-time OOD detection

In the context of test-time OOD detection, we adopt the Global-Local Maximum Concept Matching (GL-MCM) score proposed by [40]. The core innovation of GL-MCM lies in its dual-pathway architecture that synergistically integrates both global semantic patterns and localized discriminative features. Specifically, the score is formulated as follows:

$$S_{\text{GL-MCM}} = \max_{y\in\mathcal{Y}} p(y|\boldsymbol{x};\boldsymbol{\omega}) + \max_{y\in\mathcal{Y}} p(y|\boldsymbol{x}^{(k)};\boldsymbol{\omega}), \quad (8)$$

where $\boldsymbol{x}$ represents the global image, $\boldsymbol{x}^{(k)}$ denotes the k-th spatial region of image, $\mathcal{Y}$ is the ID class labels set, $\boldsymbol{\omega}$ represents the learnable context vectors.

The global component captures holistic semantic alignment between the input image and predefined class concepts, while the local component enhances sensitivity to anomalous patterns in sub-regions. The OOD detector $G_\theta(\boldsymbol{x};\mathcal{Y},\boldsymbol{\omega})$ can be defined through $S_{\text{GL-MCM}}$ as follows:

$$G_\theta(\boldsymbol{x};\mathcal{Y},\boldsymbol{\omega}) = \begin{cases} \text{ID} & S_{\text{GL-MCM}}(\boldsymbol{x}) \geq \theta \\ \text{OOD} & S_{\text{GL-MCM}}(\boldsymbol{x}) < \theta \end{cases}, \quad (9)$$

where θ is a predefined threshold, calibrated to retain a statistically significant percentile (e.g. 95%) of ID data scores, following prior works [15], [36].

V. EXPERIMENTS

A. Experimental Detail

Far OOD Detection. In this work, we follow the experimental setup of MOS [41] for far OOD detection on the ImageNet-1k [42] OOD benchmark. The ID dataset is ImageNet-1k, which consists of 1,000 categories of labeled images. For OOD samples, we utilize the iNaturalist [43], SUN [44], Places [45], and Texture [46] datasets. These OOD datasets contain no classes that overlap with the ImageNet-1k ID categories, ensuring a robust evaluation of the model's generalization ability across diverse semantic and visual categories.

Near OOD Detection. For near OOD experiments, we align with the MCM [33] setup, employing ImageNet-10 as ID dataset and ImageNet-20 as the OOD dataset, switching them interchangeably. ImageNet-10 consists of ten classes with high-resolution images, enabling evaluation on more detailed visual inputs. To facilitate near-OOD assessment, ImageNet-20 comprises twenty classes that are semantically related to those in ImageNet-10 (e.g., "dog" (ID) vs. "wolf" (OOD)).

Setups. In this experiment, we adopt the setup from prior research, employing CLIP-B/16 as the backbone, with ViT-B/16 [47] serving as the image encoder. A masked self-attention Transformer [37] is utilized as the text encoder. Furthermore, we integrate the LLaMA2-7B [23] model as the LLM, configured with a temperature parameter of 0.9. Following the hyperparameter configuration of CoOp [48], we train the model for 50 epochs with a learning rate of 0.002, a batch size of 32, an SGD optimizer, and a context token length of 16, ensuring methodological consistency. The sensitivity analysis of the unique parameters is provided in Section V-D for reference. In the main experiments, β and δ are consistently set to 0.05 and 0.85. For large-scale ID datasets, α is set to 0.98, respectively, while for small-scale ID datasets, it is set to 0.9. All experiments are conducted with Nvidia A6000 GPUs.

Comparison Methods. We conduct experiments on the CLIP [8] backbone, exploring two primary methodological paradigms: zero-shot and prompt learning methods. In the zero-shot realm, we select representative SOTA approaches, including MCM [33], CLIPN [34], and EOE [36], as foundational baseline methods. In the prompt learning methods, comparisons focused on SCT [15] and related techniques, including CoOp [48], LoCoOp [14], ID-like [11], LSN [12], NegPrompt [11], and SCT itself. Additionally, post-hoc methods (MSP [29], Energy [30], and MaxLogit [31]) were implemented as supplementary baselines on the CLIP backbone.

Evaluation Metrics. To evaluate OOD detection performance, two widely-used metrics are employed: 1) Area Under the Receiver Operating Characteristic Curve (AUROC), which quantifies the model's discriminative capability across different classification thresholds; 2) False Positive Rate at 95% True Positive Rate (FPR95), where lower values indicate better performance. To assess the impact on classification performance, the in-distribution testing accuracy (ID-ACC) is additionally reported.

B. Main Results

Comparisons on Far OOD Detection. Table I presents a comprehensive comparison of the state-of-the-art CLIP-based zero-shot and prompt learning methods, benchmarked on the large-scale ImageNet-1k as ID dataset. After 16-shot fine-tuning, TAC achieves remarkable performance across four OOD datasets, with an average FPR95 of 24.43% and AUROC of 93.70%. In contrast to SCT, which introduces modulation factors to calibrate pseudo-image outliers regularization training, TAC demonstrates superior performance by utilizing only textual outliers. TAC exhibits exceptional performance on the iNaturalist dataset, with FPR95 dramatically reduced to 7.65% and AUROC elevated to 98.44%. For Places365, our method surpasses the current SOTA level, while for SUN, TAC's performance closely aligns with the current SOTA approaches, demonstrating the method's robustness and generalizability across diverse benchmark datasets. Distinguishing from approaches like NegPrompt and LSN that incorporate negative prompts for OOD training enhancement, TAC focuses exclusively on positive prompt training. Compared to these methods, TAC yields significant improvements, with average gains of 2.18% in AUROC and 12.91% in FPR95. EOE represents the latest CLIP-based zero-shot method, and TAC's 1-shot approach demonstrates a breakthrough by improving the FPR95 by 1.92% compared to EOE. In the 1-shot setting, TAC achieves SOTA performance with an average FPR95 of 28.17% and AUROC of 92.87%, representing improvements of 2.92% and 0.83% over SCT, respectively.

The method's versatility is particularly noteworthy, as it can be integrated as a plugin to enhance existing approaches. When TAC is merged with SCT, leveraging multi-modal outlier data, it demonstrates significant potential for further performance optimization across different benchmark datasets. In the 16-shot fine-tuning scenario, this combined approach achieves a remarkably low average FPR95 of 23.10% and an impressive average AUROC of 94.19%. For more detailed results, please refer to Table IV and Section V-D.

Comparisons on Near OOD Detection. Table II presents the comparative results for near OOD detection tasks. TAC demonstrates exceptional performance across both FPR95 and AUROC metrics on the ImageNet-10 and ImageNet-20 detection benchmarks. In experiments using ImageNet-10 as ID dataset, TAC achieves a remarkably low FPR95 of 3.90% and a high AUROC of 98.73%, establishing strong performance metrics for near OOD detection. When utilizing ImageNet-20 as ID dataset, TAC maintains competitive performance, delivering improvements of 1.90% in FPR95 and 0.80% in AUROC compared to baseline methods. Note that δ is set to 0.85, we exclude outlier class labels whose similarity to ID class labels exceeds 85.00%, preventing ID sample misclassification. Overall, TAC achieves average performance across near OOD benchmarks, with an FPR95 of 5.55% and AUROC of 98.64%, significantly outperforming alternative approaches. These results demonstrate TAC's consistent effectiveness across different dataset configurations in the

TABLE I: Comparison results on Far OOD benchmarks. We use ImageNet-1k [42] as ID. We use CLIP-B/16 as a backbone. Bold values represent the superior performance. ↑ indicates larger values are better, and ↓ indicates smaller values are better.

	iNaturalist		SUN		Places		Texture		**Average**	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
CLIP-based zero-shot methods										
MSP	77.74	74.57	73.97	76.95	72.18	79.12	74.84	73.66	74.68	76.08
Energy	87.18	64.98	91.17	46.42	87.33	57.40	88.22	50.39	88.48	54.80
MaxLogit	88.03	60.88	91.16	44.83	87.45	55.54	88.63	48.72	88.82	52.49
MCM	94.61	30.91	92.57	37.59	89.77	44.69	86.11	57.77	90.77	42.74
CLIPN	95.27	23.94	93.92	26.17	92.28	33.45	**90.93**	**40.83**	**93.10**	31.10
EOE	**97.52**	**12.29**	**95.73**	**20.40**	**92.95**	**30.16**	85.64	57.53	92.96	**30.09**
Prompt learning methods										
1-shot										
CoOp	91.40	43.80	92.65	35.42	90.49	40.70	87.95	49.61	90.62	42.38
LoCoOp	94.05	28.81	94.51	25.76	91.59	33.68	86.85	51.53	91.75	34.95
ID-Like	**97.65**	**12.07**	91.07	40.55	88.31	47.94	**89.67**	**38.34**	91.68	34.73
LSN	87.20	59.28	91.47	40.15	88.74	46.11	83.92	60.34	87.83	51.47
NegPrompt	84.56	65.03	89.63	44.39	86.55	51.31	63.76	87.60	81.13	62.08
SCT	95.70	19.16	94.58	23.52	91.23	32.81	86.66	48.87	92.04	31.09
TAC	97.32	12.44	**94.63**	**23.46**	**92.31**	**30.23**	87.21	46.56	**92.87**	**28.17**
16-shot										
CoOp	93.92	28.25	93.13	31.15	90.50	39.12	90.40	41.86	91.99	35.10
LoCoOp	96.30	17.58	95.20	22.82	92.03	32.21	88.86	45.27	93.10	29.47
ID-Like	98.05	9.71	90.54	38.93	88.06	47.06	**91.89**	**32.82**	92.14	32.13
LSN	92.66	36.17	93.53	34.27	90.52	41.47	89.38	46.43	91.52	39.59
NegPrompt	90.49	37.79	92.25	32.11	91.16	35.52	88.38	43.93	90.57	37.34
SCT	95.86	13.94	**95.33**	**20.55**	92.24	29.86	89.06	41.51	93.12	26.47
TAC	**98.44**	**7.65**	94.72	21.79	**92.45**	**29.71**	89.18	38.56	**93.70**	**24.43**

TABLE II: Comparison results on Near OOD benchmarks. We use ImageNet-10 as ID.

Method	ID OOD	ImageNet-10 ImageNet-20		ImageNet-20 ImageNet-10		Average	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
CoOp		98.21	7.20	98.23	12.40	98.22	9.80
LoCoOp		98.29	5.90	98.45	8.80	98.37	7.35
SCT		97.69	6.50	98.45	8.40	98.07	7.45
TAC		**98.73**	**3.90**	**98.54**	**7.20**	**98.64**	**5.55**

challenging task of near OOD detection, with performance metrics significantly surpassing existing approaches.

C. Ablation Study

Influence of Textual Data on OOD Detection. To investigate the impact of ID textual data and outlier textual data on OOD detection performance, we conduct ablation studies. While maintaining the same configuration as the main experiments, we perform evaluations on the far OOD benchmark. Table III presents the ablation results, where α and β represent the weighting coefficients for ID textual data and outlier textual data in the training process, respectively. The experimental results demonstrate that without any textual data, the model achieves an average FPR95 of 31.16% and AUROC of 91.69% across the four datasets. Upon incorporating ID textual data, the average FPR95 significantly decreases to 27.95% while AUROC improves to 92.60%. Further integration of outlier textual data yields additional performance gains, reducing FPR95 to 24.43% and increasing AUROC to 93.70%.

These results indicate that incorporating textual data significantly enhances OOD detection performance. On the iNaturalist dataset, the configuration with complete textual data substantially improves performance compared to the baseline. This performance improvement likely stems from the rich semantic information provided by textual data, which enhances the model's ability to discriminate between ID and OOD samples. Notably, different types of textual data exhibit varying degrees of impact across datasets. Specifically, ID textual data demonstrates the most significant performance improvement on the iNaturalist dataset. Meanwhile, the introduction of outlier textual data shows the strongest effect on the Places dataset. This variability suggests that the effectiveness of textual data may be closely related to the inherent feature distributions of different datasets.

Cross-LLMs Performance Analysis. We conduct experiments across multiple LL-Ms, including LLaMA2-7B [23], Claude-3.5 [39], ChatGPT-4o [38], and Deepseek-V3 [24]. As shown in Fig. 5, the experiments use ImageNet-10 as ID dataset and ImageNet-20 as OOD dataset. The results demonstrate that our method consistently outperforms both the CoOp baseline and the image outlier detection method SCT across all LLM configurations, validating the method's robustness and generalizability across different language model architectures. Specifically, both ChatGPT-4o and Deepseek-V3 achieve superior FPR95 performance. ChatGPT-4o demonstrates the strongest performance with FPR95 of 3.90% and AUROC of 98.73%, likely attributed to its robust contextual understanding and precise grasp of visual concepts. Deepseek-V3 follows closely with FPR95 of 4.10% and AUROC of 98.47%.

TABLE III: Ablation study on the effect of ID and outlier textual data in the prompt learning process, where the weighting coefficients are either zero or non-zero. ✗ indicates that the weight coefficient is set to zero, while ✓ represents a non-zero coefficient.

		iNaturalist		SUN		Places		Texture		**Average**	
α	β	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
✗	✗	96.56	14.66	92.67	29.20	89.82	36.88	87.70	43.90	91.69	31.16
✓	✗	98.19	7.87	93.18	26.02	90.72	34.13	88.30	43.97	92.60	27.95
✗	✓	97.75	10.56	93.93	23.72	91.53	30.94	87.84	43.40	92.76	27.16
✓	✓	**98.44**	**7.65**	**94.72**	**21.79**	**92.45**	**29.71**	**89.18**	**38.56**	**93.70**	**24.43**

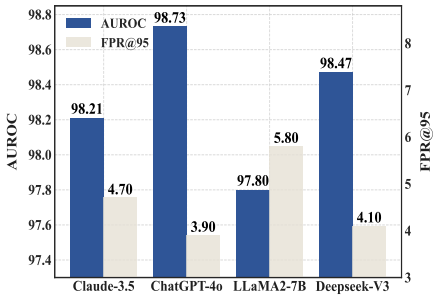


Fig. 5: Various LLMs Performance.

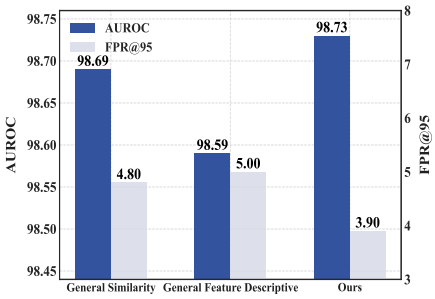


Fig. 6: LLM Prompt Ablation Study.

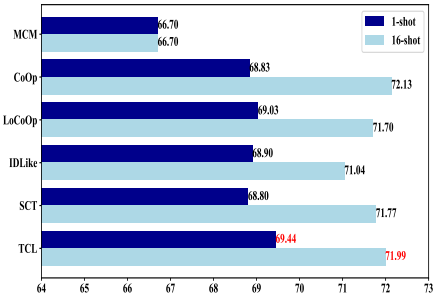


Fig. 7: ID Accuracy Comparison Results.

While LLaMA2-7B and Claude-3.5 exhibit relatively lower performance, they maintain effective detection capabilities. These findings indicate that various LLM architectures can effectively support OOD detection tasks, with model selection significantly impacting performance outcomes.

Evaluating Key Principles in LLMs query template. To validate the effectiveness of *Visual Similarity* and *Feature Distinctiveness* principles in LLMs query templates, we conduct controlled experiments. Based on our original LLMs query template design, we construct two variant prompts: a general similarity prompt and a general feature descriptive prompt. The general similarity prompt instructs LLMs to generate similar but different categories without emphasizing visual similarity constraints, while the general feature descriptive prompt asks LLMs to describe general features of categories without emphasizing feature distinctiveness. As shown in Fig. 6, experimental results on the ImageNet-10 benchmark demonstrate that without two principles, OOD detection performance degrades to varying degrees in both FPR95 and AUROC metrics. We posit that the visual similarity constraint helps bridge the gap between LLMs and VLMs in similarity perception, while feature distinctiveness enables LLMs to transfer more effective text-level classification knowledge to VLMs. The synergistic effect of these two principles provides substantial support for improving OOD detection performance.

D. Further Analysis

Impact of Multi-Modal Joint Training on ID Accuracy. To investigate the impact of different class distributions and multi-modal joint training methods on ID accuracy, we present the ACC comparison results of various methods in Fig. 7. It shows that the ACC obtained using our proposed training approach is almost identical to the baseline, with the baseline ACC in 16-shot at 72.13% and the ACC of our method at 71.99%. This minimal difference indicates that our method has only a little negative impact on ID classification performance. Furthermore, despite the negligible decline in ID accuracy, our method significantly outperforms other advanced OOD training methods in OOD detection on the large-scale ImageNet-1k benchmark. This demonstrates the effectiveness of our method in balancing ID classification and OOD detection, highlighting its unique advantages. Our method achieves an unprecedented 1-shot ACC of 69.44%, surpassing all comparative approaches and demonstrating exceptional fine-tuning efficiency under ultra-low data regime scenarios.

Enhancing OOD Detection via Plugin Integration. By integrating TAC into diverse baseline methods, we observe consistent improvements in OOD detection performance, as shown in Table IV. The plugin strategies across different approaches demonstrate TAC's remarkable adaptability: LoCoOp+TAC provides multimodal outlier regularization training, SCT applies adjustment factors to optimize visual outlier training, and EOE incorporates outlier data simultaneously in inference and training. Experimental results reveal that integrating the TAC plugin consistently enhances baseline methods' average AUROC and FPR95 metrics. Particularly, the SCT+TAC combination achieves optimal performance, with an average AUC of 94.19% and FPR95 reduced to 23.10%, substantiating the method's effectiveness. These findings not only validate TAC's potential as a universal plugin but also illuminate a novel paradigm for cross-modal anomaly detection.

Performance Evaluation in Various Few-Shot Scenarios. To comprehensively evaluate the performance of our method in few-shot learning scenarios, we conduct comparative exper-

TABLE IV: OOD detection performance by integrating TAC into various baseline methods in ImageNet-1k benchmark.

	iNaturalist		SUN		Places		Texture		**Average**	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
EOE	97.52	12.29	95.73	20.40	92.95	30.16	85.64	57.53	92.96	30.09
EOE+TAC	**98.45**	**7.34**	95.76	20.31	92.43	29.84	89.32	39.65	93.74	24.29
LoCoOp	96.30	17.58	95.20	22.82	92.03	32.21	88.86	45.27	93.10	29.47
LoCoOp+TAC	97.59	11.36	95.21	21.04	92.48	29.22	89.37	37.68	93.66	24.83
SCT	95.86	13.94	95.33	20.55	92.24	29.86	89.06	41.51	93.12	26.47
SCT+TAC	98.36	9.76	**95.94**	**18.66**	**92.99**	**27.27**	**89.76**	**36.72**	**94.19**	**23.10**

TABLE V: OOD detection performance across different few-shot settings with ImageNet-10 as ID and ImageNet-20 as OOD.

	1-shot		2-shot		4-shot		8-shot		**Average**	
Method	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
CoOp	97.92	9.40	98.12	6.10	98.18	7.00	98.16	7.10	98.10	7.40
LoCoOp	97.93	9.40	98.25	7.30	97.98	5.30	98.12	7.00	98.07	7.25
SCT	98.16	6.00	97.25	10.60	97.98	6.80	97.99	7.90	97.85	7.83
TAC	**98.68**	**3.50**	**98.48**	**5.70**	**98.37**	**3.60**	**98.21**	**6.50**	**98.44**	**4.83**

iments on the ImageNet-10 OOD benchmark using 1, 2, 4, and 8-shot settings. The experimental results are presented in Table V. In the 1-shot scenario, the TAC method demonstrates a significant performance advantage, achieving an FPR95 of only 3.50%, which is substantially lower than that of other methods. This indicates that our approach exhibits exceptional out-of-distribution detection capabilities under extremely limited data conditions. As the number of samples increases, the TAC method maintains a relatively stable performance. On average, our method achieves an FPR95 of 4.83% and an AUROC of 98.44%, showcasing more consistent and reliable performance compared to other baseline methods. These results provide strong evidence for the effectiveness and generalization potential of the proposed method.

VI. Conclusion

In this work, we propose a novel OOD detection paradigm by leveraging LLMs to augment VLMs' capacity to identify OOD samples in few-shot scenarios. Employing query templates grounded in *Visual Similarity* and *Feature Discriminative* principles, we extract semantically distinctive category representations from LLMs, constructing a cross-modal outlier representation space. Through a visual-textual collaborative optimization mechanism, we integrate LLM-generated textual outliers as regularization constraints, guiding VLMs to synchronously learn ID category-specific features and potential outlier distribution patterns during prompt learning. Experimental validation across multiple benchmarks demonstrates significant performance improvements over existing CLIP-based OOD detection approaches. However, the alignment between text and visual feature spaces in VLMs still has the potential for enhancement. Future research could focus on bridging the gap between text and image modalities to further enhance the effectiveness of textual information in OOD detection.

Acknowledgment

National Key RD Program of China (2022YFF0712100), NSFC (62276131, 62506168), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081), the Fundamental Research Funds for the Central Universities (No.30922010317, No.30923011007, No.30925010205).

References

[1] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.

[2] W. Xi, X. Song, W. Guo, and Y. Yang, "Robust semi-supervised learning for self-learning open-world classes," in *IEEE International Conference on Data Mining (ICDM)*, 2023, pp. 658–667.

[3] Y. Yang, N. Jiang, Y. Xu, and D.-C. Zhan, "Robust semi-supervised learning by wisely leveraging open-set data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8334–8347, 2024.

[4] H. Xu and Y. Yang, "Itp: Instance-aware test pruning for out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 2025, pp. 21 743–21 751.

[5] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[6] Y. Yang and H. Xu, "Strengthen out-of-distribution detection capability with progressive self-knowledge distillation," in *International Conference on Machine Learning*, 2025.

[7] J. Wei, G. Wang, and S. Zhang, "Fine-grained medical image out-of-distribution detection through multi-view feature uncertainty and adversarial sample generation," *Pattern Recognition*, p. 111401, 2025.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[9] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, 2021, pp. 4904–4916.

[10] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," in *International Conference on Learning Representations*, 2022.

[11] T. Li, G. Pang, X. Bai, W. Miao, and J. Zheng, "Learning transferable negative prompts for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 584–17 594.

[12] J. Nie, Y. Zhang, Z. Fang, T. Liu, B. Han, and X. Tian, "Out-of-distribution detection with negative prompts," in *International Conference on Learning Representations*, 2024.

[13] Y. Bai, Z. Han, B. Cao, X. Jiang, Q. Hu, and C. Zhang, "Id-like prompt learning for few-shot out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 480–17 489.

[14] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Locoop: Few-shot out-of-distribution detection via prompt learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 298–76 310, 2023.

[15] G. Yu, J. Zhu, J. Yao, and B. Han, "Self-calibrated tuning of vision-language models for out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 56 322–56 348, 2024.

[16] Y. Ming, Y. Fan, and Y. Li, "Poem: Out-of-distribution detection with posterior sampling," in *International Conference on Machine Learning*, 2022, pp. 15 650–15 665.

[17] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Atom: Robustifying out-of-distribution detection using outlier mining," in *Machine Learning and Knowledge Discovery in Databases*, 2021, pp. 430–445.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[19] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[20] H. Mirzaei, M. Jafari, H. R. Dehbashi, A. Ansari, S. Ghobadi, M. Hadi, A. S. Moakhar, M. Azizmalayeri, M. S. Baghshah, and M. H. Rohban, "Rodeo: Robust outlier detection via exposing adaptive out-of-distribution samples," in *International Conference on Machine Learning*, 2024.

[21] X. Du, Y. Sun, J. Zhu, and Y. Li, "Dream the impossible: Outlier imagination with diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 60 878–60 901, 2023.

[22] K. Kirchheim and F. Ortmeier, "On outlier exposure with generative models," *Advances in Neural Information Processing Systems*, 2022.

[23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[24] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[25] R. Xu and K. Ding, "Large language models for anomaly and out-of-distribution detection: A survey," in *Findings of the Association for Computational Linguistics: NAACL*, 2025, pp. 5992–6012.

[26] X. Wu, Q.-Y. Jiang, Y. Yang, Y.-F. Wu, Q.-G. Chen, and J. Lu, "Tai++: Text as image for multi-label image classification by co-learning transferable prompt," in *International Joint Conference on Artificial Intelligence*, 2024.

[27] D. Chao, Y. Zhang, L. Zhou, and Y. Yang, "Enriching category representations with llms towards robust zero-shot out-of-distribution detection," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2025.

[28] G. Zheng, W. Ye, and A. Zhang, "Benchmarking spurious bias in few-shot image classifiers," in *European Conference on Computer Vision*, 2024, pp. 346–364.

[29] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2016.

[30] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 464–21 475, 2020.

[31] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *International Conference on Machine Learning*, 2022, pp. 8759–8773.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 087–35 102, 2022.

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

[35] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu, "Zero-shot out-of-distribution detection based on the pre-trained model clip," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 6, 2022, pp. 6568–6576.

[36] C. Cao, Z. Zhong, Z. Zhou, Y. Liu, T. Liu, and B. Han, "Envisioning outlier exposure by large language models for out-of-distribution detection," *International Conference on Machine Learning*, 2024.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[38] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[39] "The claude 3 model family: Opus, sonnet, haiku." [Online]. Available: https://api.semanticscholar.org/CorpusID:268232499

[40] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection," *International Journal of Computer Vision*, pp. 1–11, 2025.

[41] R. Huang and Y. 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*, 2021, pp. 8710–8719.

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

[43] G. V. Horn, O. M. Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. J. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8769–8778.

[44] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.

[45] B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.

[46] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3606–3613.

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.