Out-of-Distribution Detection with Positive and Negative Prompt Supervision Using Large Language Models

Zhixia He[1], Chen Zhao[2], Minglai Shao[1*], Xintao Wu[3], Xujiang Zhao[4], Dong Li[2], Qin Tian[5], Linlin Yu[6]

[1]School of New Media and Communication, Tianjin University, Tianjin, China
[2]Department of Computer Science, Baylor University, Waco, Texas, USA
[3]Department of Electrical Engineering and Computer Science, University of Arkansas, Fayetteville
[4]NEC Laboratories America
[5]College of Intelligence and Computing, Tianjin University, Tianjin, China
[6]Department of Computer Science, Augusta University, Augusta, GA, USA
{2023245033, shaoml, tianqin123}@tju.edu.cn, {chen_zhao, dong_li1}@baylor.edu, xintaowu@uark.edu, xuzhao@nec-labs.com, linyu@augusta.edu

Abstract

Out-of-distribution (OOD) detection is committed to delineating the classification boundaries between in-distribution (ID) and OOD images. Recent advances in vision-language models (VLMs) have demonstrated remarkable OOD detection performance by integrating both visual and textual modalities. In this context, negative prompts are introduced to emphasize the *dissimilarity* between image features and prompt content. However, these prompts often include a broad range of non-ID features, which may result in suboptimal outcomes due to the capture of overlapping or misleading information. To address this issue, we propose *Positive and Negative Prompt Supervision*, which encourages negative prompts to capture inter-class features and transfers this semantic knowledge to the visual modality to enhance OOD detection performance. Our method begins with class-specific positive and negative prompts initialized by large language models (LLMs). These prompts are subsequently optimized, with positive prompts focusing on features within each class, while negative prompts highlight features around category boundaries. Additionally, a graph-based architecture is employed to aggregate semantic-aware supervision from the optimized prompt representations and propagate it to the visual branch, thereby enhancing the performance of the energy-based OOD detector. Extensive experiments on two benchmarks, `CIFAR-100` and `ImageNet-1K`, across eight OOD datasets and five different LLMs, demonstrate that our method outperforms state-of-the-art baselines.

Introduction

When deploying machine learning models in open-world scenarios, it is inevitable to encounter samples from previously unseen classes, commonly referred to as *out-of-distribution* (OOD) data (Hendrycks and Gimpel 2016; Shao et al. 2024; Lin et al. 2025a; Zhao, Chen, and Thuraisingham 2021; Wu et al. 2025; Lin et al. 2025b). This issue is particularly critical in high stakes applications, such as autonomous driving (Bogdoll, Nitsche, and Zöllner 2022) and medical diagnostics (Li et al. 2025a; Zimmerer et al. 2022; Li et al. 2025b), where the misclassification of OOD samples can pose significant safety hazards. Moreover, recent studies have revealed that even state-of-the-art deep neural networks often make overconfident predictions on OOD data (Parmar et al. 2023; Li et al. 2024a). Consequently, there is a growing need for OOD detection methods that can effectively identify OOD samples.

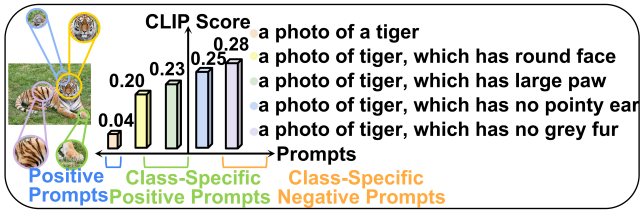


Figure 1: Illustration of the CLIP score using prompts and an image of a tiger. Compared to the prompt "*a photo of a tiger*", positive prompts enriched with visual features significantly enhances the CLIP score. Furthermore, the introduction of negative features can further increase the score.

Traditional OOD detection methods primarily rely on a single visual modality, overlooking the rich semantic content carried by labels (Liu et al. 2020). With the emergence of vision-language models (VLMs), a paradigm shift has occurred—from relying on vision-only information to integrating both visual and textual modalities (Zhang et al. 2024). In the context of VLM-based OOD detection, the textual modality is typically represented by positive prompts (e.g., *"a photo of a {class}"*), which are used to estimate the probability of an image belonging to the given class. However, these fixed-format prompts do not account for the distinctions between categories. To address this challenge, class-specific positive prompts enriched with visual features generated by large language models (LLMs) have been introduced, demonstrating superior performance compared to those using only category names (Menon and Vondrick 2022). More recently, CLIPN (Wang et al. 2023) incorporates *"no class"* into prompts to express negative concepts, which are referred to as negative prompts. Despite these ad-

*Corresponding authors

vances, the use of negative prompts may still result in the learning of overlapping non-ID features or noisy information, potentially leading to suboptimal performance. Given the limitations of existing endeavors, our twofold objective is to: (1) encourage positive and negative prompts to comprehensively capture ID category features and to clearly delineate the category boundaries; and (2) aggregate semantic-aware supervision from prompt representations and transfer it to the visual branch to improve OOD detection.

One promising direction is to construct class-specific negative prompts by augmenting positive prompts with negative features. To preliminarily explore its effectiveness, we select an image of a tiger, prompt LLMs to generate class-specific descriptions, and calculate the corresponding CLIP score, as illustrated in Figure 1. The results indicate that, compared to positive prompts, class-specific negative prompts achieve a better match with the image. As will be discussed later, these augmented prompts primarily capture inter-class features and guide the model focus on category distinctions, thus better matching the corresponding image. In this paper, we propose the **P**ositive and **N**egative **P**rompt **S**upervision (**PNPS**) framework. This framework consists of three phases: prompt construction with large language models, alignment of visual and textual modalities, and cross-modal graph neural networks. We first construct positive and negative prompts by prompting LLMs to generate class-specific visual descriptions. To further enhance the expressiveness of these prompts, we introduce learnable textual parameter matrices, which enable positive and negative prompts to effectively capture intra-class and inter-class features, respectively. However, a key challenge for improving OOD detection lies in how to aggregate semantic-aware supervision from the optimized prompts and propagate it to the image representations. To address this, we build multi-modal graphs to facilitate the aggregation and propagation of semantic supervision within and across modalities. Extensive experiments on two ID and eight OOD datasets using five different LLMs demonstrate the effectiveness of our proposed PNPS method. Our main contributions are:

- We employ a graph-based structure to aggregate both positive and negative prompt representations as semantic-aware supervision, which is then propagated to the visual modality to enhance OOD detection performance. To our knowledge, we are the first to utilize a graph-based framework for OOD detection in the image domain.
- We introduce a novel three-phase PNPS framework that optimizes class-specific positive and negative prompts to capture intra-class and inter-class features, facilitating a more profound understanding of ID features as well as the delineation of more clearly defined category boundaries.
- On the `CIFAR-100` benchmark, our PNPS improves AUROC by 1.06%, 2.10%, 4.41%, and 3.97% over the best baseline on the `CIFAR-10`, `SVHN`, `Texture`, and `Places365`, respectively. In addition, our approach also achieves superior performance on `ImageNet-1K`.

Related Work

Out-of-Distribution Detection. OOD detection aims to identify images that do not belong to any category in the training dataset. Early research primarily focuses on designing score functions based on the predicted logits, such as the energy score (Liu et al. 2020). Data augmentation techniques are also adopted to enhance data diversity and improve model generalization, by applying effective transformations to the training data (Goodfellow et al. 2020; Nie et al. 2024). Furthermore, later studies explore extracting class-agnostic information from the feature space, which is not accessible from predicted logits alone (Sun, Guo, and Li 2021). For instance, ViM (Wang et al. 2022) enhances OOD detection performance by combining class-agnostic features with logits. Moreover, with recent advancements in VLMs, leveraging textual information becomes a new and promising direction to further improve image OOD detection.

VLM-Based Out-of-Distribution Detection. With the advancements in VLMs, a series of VLM-based OOD detection methods have rapidly emerged. As an early application, MCM (Ming et al. 2022) utilizes the maximum logit of scaled softmax to detect OOD images. Subsequently, the interaction between NLP and CV has facilitated the application of prompt learning from NLP to OOD detection, notably exemplified by CoOp (Zhou et al. 2022). More recently, with the emergence of LLMs, prompt-based LLMs have enabled the automatic generation of class-specific visual features. Methods such as DCLIP (Menon and Vondrick 2022) and CuPL (Pratt et al. 2023) demonstrate that prompts with detailed descriptions can further enhance the matching between images and prompts. Building on these studies, a promising direction involves simulating non-ID scenarios to improve OOD detection performance. For instance, CLIPN (Wang et al. 2023) introduces negative prompts, such as *" a photo of no* $\{class\}$*"*, to capture negation semantics within images, while EOE (Cao et al. 2024) leverages LLMs to generate pseudo-OOD labels. Although the aforementioned methods have shown promising results, explorations that combine prompts with negative features remain limited, leaving significant potential for further research.

Preliminaries

Let $\mathcal{X}$ denotes the image space and $\mathcal{Y} = \{y_1, y_2, ..., y_{|\mathcal{C}|}\}$ represents the set of ID class labels, where $|\mathcal{C}|$ is the total number of classes. We define x as the random variable sampled from $\mathcal{X}$, with each sample associated with a label $y \in \mathcal{Y}$. Notably, the training and testing data are drawn from different distributions, *i.e.*, $\mathbb{P}^{tr}(\mathcal{X}, \mathcal{Y}) \neq \mathbb{P}^{te}(\mathcal{X}, \mathcal{Y})$, where test set may include OOD instances with labels $y \notin \mathcal{Y}$.

CLIP and DCLIP. CLIP (Radford et al. 2021) is trained on 400 million image-text pairs and utilizes contrastive learning to align visual and textual representations. It comprises a text encoder $\mathcal{T} : t \rightarrow \mathcal{R}^d$ and an image encoder $\mathcal{I} : x \rightarrow \mathcal{R}^d$. During inference, given an image-label pair $(x, y) \in (\mathcal{X}, \mathcal{Y})$, a prompt like *"a photo of a* $\{y\}$*"* is fed to the text encoder to obtain $\mathcal{T}(t)$. The predicted probability

for the image feature $\mathcal{I}(x)$ is then computed as follows:

$$p(y=i|x)=\frac{e^{<\mathcal{I}(x),\mathcal{T}(t^i)>/\tau}}{\sum_{c=1}^{|\mathcal{C}|}e^{<\mathcal{I}(x),\mathcal{T}(t^c)>/\tau}}, \quad (1)$$

where τ is the temperature parameter. The following introduces DCLIP (Menon and Vondrick 2022), which leverages LLMs to generate visual features that describe the object category in a photo. Specifically, it prompts the LLMs with the following query $\boldsymbol{Q}$ and obtains the corresponding answer $\boldsymbol{A}$:

***Q**: What are useful features for distinguishing a {class} in a photo?*
***A**: There are {visual features} to tell there is a {class} in a photo.*

The generated visual features are embedded into a fixed-format template to construct class-specific positive prompts, which outperform those using only the category names. For fair comparison with the template *"a photo of a {class}"*, we standardize the class-specific positive prompt format as *"a photo of a {class}, which has {visual features}"*.

Methodology

In this section, we describe our proposed PNPS framework. As shown in Figure 2, the architecture consists of three main phases: 1) utilizing LLMs to generate category-discriminative features for constructing prompts; 2) optimizing prompts to capture both inter-class and intra-class features; and 3) employing a graph-based structure to aggregate and propagate semantic knowledge extracted from prompts.

Prompt Construction with Large Language Models

Traditional OOD detection methods typically represent semantic categories (e.g., tiger) as numerical labels (e.g., 0), which capture only index-based relationships and overlook rich semantic content. To bridge this gap, positive textual prompt templates, such as *"a photo of a {class}"*, have been introduced; however, these prompts still fail to capture the intrinsic differences between categories. To emphasize the distinctions between categories, recent LLM-enriched approaches construct class-specific positive prompts by appending discriminative visual descriptions (Menon and Vondrick 2022). Likewise, negative prompts are introduced to learn the dissimilarity for each categories (Wang et al. 2023).

This naturally raises a question: could incorporating negative category features into prompt templates yield better performance? We refer to such prompts as *"class-specific negative prompts"*, and evaluate their effectiveness with pre-trained CLIP based on ViT/B-16, as shown in Figure 1. The results indicate that the use of these negative prompts leads to a better match with the image. Building on this finding, we further enhance the match by encouraging LLMs to generate more discriminative descriptions. Specifically, we divide categories in the training set into super-classes based on species similarity, as exemplified by the five classes {*"tiger", "wolf", "bear", "leopard", "lion"*} in the `CIFAR-100` dataset are grouped into *"large carnivores"* super-class. Take the tiger class as an example, we modify the original query from $\boldsymbol{Q}$ to $\boldsymbol{Q'}$ as follows:

***Q'**: What are useful features for distinguishing a {tiger} from {wolf, bear, leopard, lion} in a photo?*

The reason for the above is to encourage LLMs to generate non-overlapping features, thus offering more discriminative clues for category differentiation. As shown in Figure 2, we incorporate the generated feature *"striped fur"* into the template to construct the class-specific positive prompt *"a photo of a tiger, which has striped fur"*, symbolized as t^{tiger+}. To exemplify, consider the task of distinguishing a tiger from a wolf, bear, leopard, and lion. Since *"striped fur"* is unique to the tiger, its negation—*"no striped fur"* serves as a complementary descriptor for the other four categories. By doing so, the model is guided to attend to fur-related features when distinguishing between categories. For instance, the class-specific negative prompt for the wolf is *"a photo of a wolf, which has no striped fur"*, represented as t^{wolf-}.

All categories in the training dataset are grouped into $|\mathcal{C}^{super}|$ super-classes, where $\mathcal{C}^{super}$={*"large carnivores", "trees",* $\cdots$}. The super-class containing the tiger category is denoted as $\mathcal{C}^{super}_{tiger}$={*"large carnivores"*}. If N visual features are generated for each category, then each category will have N class-specific positive prompts and $(|\mathcal{C}^{super}_y|-1)\cdot N$ class-specific negative prompts. To illustrate, for the tiger category, negative prompts can be constructed by taking the negations of the N visual features from each of the other four categories. These prompts are then fed into the text encoder $\mathcal{T}(\cdot)$ to obtain representations $\mathbf{H}^T=\{\mathbf{H}^{T+},\mathbf{H}^{T-}\}$.

$$\mathbf{H}^T=\left\{\boldsymbol{h}_1^T,\cdots,\boldsymbol{h}_N^T,\boldsymbol{h}_{N+1}^T,\cdots,\boldsymbol{h}_{|\mathcal{C}^{super}_y|\cdot N}^T\right\}, \quad (2)$$

where $\mathbf{H}^{T+}=\left\{\boldsymbol{h}_n^T\right\}_{n=1}^{N}$ and $\mathbf{H}^{T-}=\left\{\boldsymbol{h}_n^T\right\}_{n=N+1}^{|\mathcal{C}^{super}_y|\cdot N}$. Ideally, positive prompts capture intra-class features, whereas negative prompts, constructed by combining a category name with the description of another class, should have their representations positioned at the boundaries between the two categories. However, due to the hallucination in LLMs-generated features, and the limited expressiveness of the pre-trained text encoder, the resulting representations may fail to convey the semantic meaning of *"no"* (Nie et al. 2024). The subsequent step is to explore an effective method for optimizing the aforementioned prompt representations.

Alignment of Visual and Textual Modalities

Inspired by CLIP-Adapter (Gao et al. 2021), which achieves content optimization by adding an additional linear layer after the pre-trained text encoder in few-shot learning setting, we introduce a learnable prompt parameter matrix $\mathbf{W}^T$, while keeping the pre-trained text encoder frozen. The resulting transformed textual representations are given by:

$$\hat{\mathbf{H}}^T=\left\{\hat{\boldsymbol{h}}_1^T,\cdots,\hat{\boldsymbol{h}}_N^T,\hat{\boldsymbol{h}}_{N+1}^T,\cdots,\hat{\boldsymbol{h}}_{|\mathcal{C}^{super}_y|\cdot N}^T\right\}=\mathbf{W}^T\mathbf{H}^T. \quad (3)$$

To achieve better alignment between the visual and textual modalities, we also introduce a learnable visual parameter matrix $\mathbf{W}^I$. The subsequent challenge lies in designing loss functions that enable class-specific positive prompts to effectively capture intra-class features, while negative prompts encode inter-class distinctions. The optimization process involves three components: positive loss $\mathcal{L}^+$, negative loss

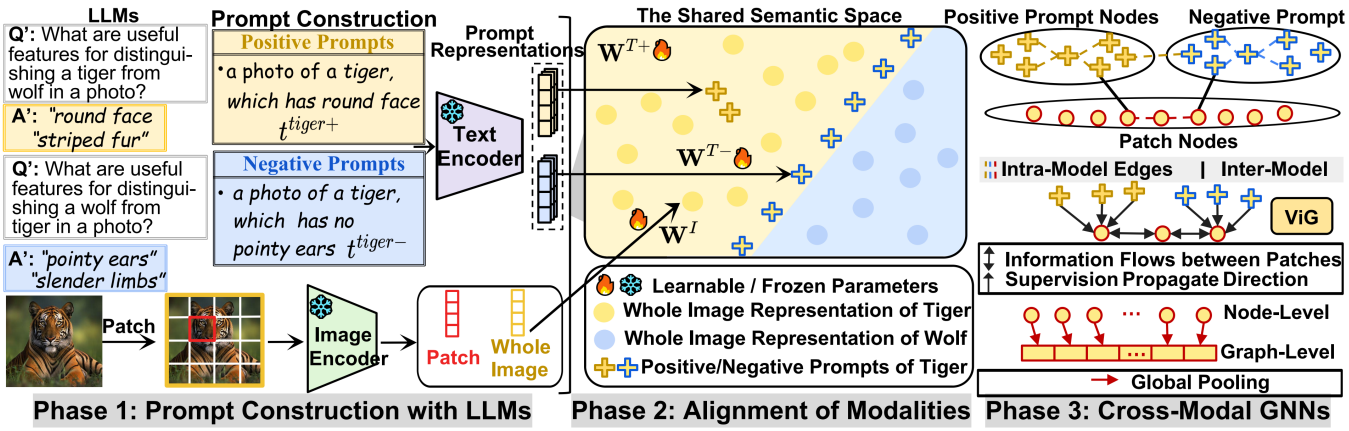


Figure 2: An overview of the PNPS framework. To begin with, we employ LLMs to generate discriminative features, which are then filled into templates to construct class-specific positive and negative prompts. These prompts, together with image patches and full images, are subsequently encoded into their respective representations. To enhance the expressiveness of the prompt representations, we introduce learnable textual and visual parameter matrices, $\mathbf{W}^T$ and $\mathbf{W}^I$, for further optimization. Building on these optimized textual representations, we then construct cross-modal graph connections to aggregate semantic supervision from the prompts and propagate it to the visual branch, thereby improving the performance of image OOD detection.

$\mathcal{L}^-$, and the negative-positive distant loss $\mathcal{L}_{npd}$, which optimizes the relative distance between positive prompts and their corresponding negative prompts. Specifically, we begin by optimizing the positive representations as follows:

Positive-Image Related Loss (PIR). Aligning positive prompts with their respective images allows prompts to capture rich intra-class features. Specifically, for a given image x, we extract its representation $\boldsymbol{h}^I(x)$ with the pre-trained image encoder $\mathcal{I}(\cdot)$, and project it through the learnable matrix $\mathbf{W}^I$ to obtain the transformed image representation $\hat{\boldsymbol{h}}^I(x)$. The prediction probability for an ID image x and the i-th positive prompts t^{i+} is then computed as follows:

$$p_i^+ = \frac{\sum_{n=1}^{N} e^{<\hat{\boldsymbol{h}}^I(x), \hat{\boldsymbol{h}}_n^T(t^{i+})>/\tau}}{\sum_{c=1}^{|\mathcal{C}|}\sum_{n=1}^{N} e^{<\hat{\boldsymbol{h}}^I(x), \hat{\boldsymbol{h}}_n^T(t^{c+})>/\tau}}, \tag{4}$$

where $|\mathcal{C}|$ denotes the total number of ID classes, and N represents the number of positive prompts per category. The positive-image related loss $\mathcal{L}_{pir}$ is defined as:

$$\mathcal{L}_{pir} = -\mathbb{E}_{(x,t^{i+})\sim\mathcal{D}^{tr}} \log p_i^+. \tag{5}$$

Positive-Positive Distant Loss (PPD). The positive-positive distant loss $\mathcal{L}_{ppd}$ is introduced to encourage diversity among the N class-specific positive representations and ensure that each one captures distinct semantic information.

$$\mathcal{L}_{ppd} = \sum_{c=1}^{|\mathcal{C}|}\sum_{i=1}^{N}\sum_{j=i+1}^{N} |<\hat{\boldsymbol{h}}_i^T(t^{c+}), \hat{\boldsymbol{h}}_j^T(t^{c+})>|. \tag{6}$$

The overall objective for optimizing positive representations is defined as $\mathcal{L}^+ = \mathcal{L}_{pir} + \lambda^+ \cdot \mathcal{L}_{ppd}$, with λ^+ serving as a hyperparameter to balance the two components. The optimization process for negative prompts is described below:

Negative-Image Related Loss (NIR). We denote class-specific negative template *"a photo of a* $\{y_a\}$*, which has* $\{$*no visual features of* $y_b\}$*"* as t^{a-b}, where $y_a, y_b \in \mathcal{C}^{super}_{[y_a,y_b]}$ and the negative features of y_b serve as the complementary descriptor for y_a. Accordingly, the representation $\hat{\boldsymbol{h}}^T(t^{a-b})$ should be aligned with the image representations of y_a, and diverge from those of y_b in the shared semantic space. This design facilitates the learning of well-defined category boundaries between y_a and y_b, and enables the model to capture more discriminative inter-class features for y_a. The matching probability between x and the i-th negative prompts t^{i-} is computed as follows:

$$p_i^- = \sum_{c=1, c\neq i}^{|\mathcal{C}^{super}_{[i,c]}|} \sum_{n=c\cdot N+1}^{(c+1)\cdot N+1} s_i^-(c,n), \tag{7}$$

where $s_i^-(c,n)$ is the matching score between image x and the n-th negative prompt of the c-th category within the co-existing super-class, formulated as follows:

$$s_i^-(c,n) = \frac{e^{<\hat{\boldsymbol{h}}^I(x), \hat{\boldsymbol{h}}_n^T(t^{i-c})>}}{e^{<\hat{\boldsymbol{h}}^I(x), \hat{\boldsymbol{h}}_n^T(t^{i-c})>} + e^{<\hat{\boldsymbol{h}}^I(x), \hat{\boldsymbol{h}}_n^T(t^{c-i})>}}, \tag{8}$$

where the negative representations for class y_c are indexed from $c\cdot N+1$ to $(c+1)\cdot N+1$. $\mathcal{L}_{nir}$ is computed as follows:

$$\mathcal{L}_{nir} = -\mathbb{E}_{(x,t^{i-})\sim\mathcal{D}^{tr}} \frac{1}{(|\mathcal{C}_i^{super}|-1)\cdot N} \log p_i^-. \tag{9}$$

Negative-Negative Distant Loss (NND). To ensure diversity and non-overlap among negative prompts, we enforce greater separation between distinct negative representations.

$$\mathcal{L}_{nnd} = \sum_{c=1}^{|\mathcal{C}|} \sum_{\substack{d=1 \\ d\neq c}}^{|\mathcal{C}^{super}_{[c,d]}|} \sum_{\substack{i=c\cdot \\ N+1}}^{(c+1)\cdot N} \sum_{\substack{j= \\ i+1}}^{(c+1)\cdot N} |<\hat{\boldsymbol{h}}_i^T(t^{c-d}), \hat{\boldsymbol{h}}_j^T(t^{c-d})>|. \tag{10}$$

The total loss for negative representations is formulated as $\mathcal{L}^- = \mathcal{L}_{nir} + \lambda^- \cdot \mathcal{L}_{nnd}$. In addition to the positive loss $\mathcal{L}^+$ and the negative loss $\mathcal{L}^-$, we introduce negative-positive distance loss $\mathcal{L}_{npd}$ to increase the separation between positive and their respective negative prompt representations.

Negative-Positive Distant Loss (NPD). Theoretically, the positive prompt t^{a+} and the negative prompt t^{b-a} are semantically opposite in the shared semantic space. To further enforce their distinction, we define $\mathcal{L}_{npd}$ as follows:

$$\mathcal{L}_{npd}=\sum_{c=1}^{|\mathcal{C}|}\sum_{\substack{d=1\\d\neq c}}^{|\mathcal{C}^{super}_{[c,d]}|}\sum_{n=1}^{N}|<\hat{\boldsymbol{h}}_n^T(t^{c+}),\hat{\boldsymbol{h}}_{d\cdot N+n}^T(t^{d-c})>|, \quad (11)$$

where for the positive representation $\hat{\boldsymbol{h}}_n^T(t^{c+})$, the corresponding negative index is $d\cdot N+n$ for $d\in\mathcal{C}^{super}_{[c,d]}$. Once optimized, the prompts are enriched with semantic-aware information. The next objective is to transfer this semantic supervision to the visual branch to enhance OOD detection.

Cross-Modal Graph Neural Networks

Multi-Modal Graph Construction. Graph-based architecture offers a natural solution for aggregating semantic supervision from the optimized prompt representations and propagating it to the visual branch. Heterogeneous graphs, in particular, facilitate bidirectional message passing and feature updates between different node types (Zhang et al. 2019). Within this structure, information flows via intra-modal and inter-modal edges within the multi-modal graph.

As class-specific prompts incorporate local detail descriptions, the model is guided to attend to semantically consistent image features during optimization. Consequently, the optimized prompt representations are more concentrated on these local features. To align with this locality, we employ the pre-trained ViT to extract patch features and link them to prompts via cross-modal edges. Specifically, for an image x and its corresponding label y, we obtain $|\mathcal{C}_y^{super}|\cdot N$ prompt representations and M patch representations. To facilitate this connection, we first construct an unordered node set $\mathcal{V}=\{\mathcal{V}^P,\mathcal{V}^T\}$, comprising patch and prompt nodes.

$$\mathcal{V}=\{v_1^P,\cdots,v_M^P,v_{M+1}^T,\cdots,v_{|\mathcal{C}_y^{super}|\cdot N+M}^T\}. \quad (12)$$

For each node $v_i\in\mathcal{V}^m$ of the modality $m\in\{P,T\}$, we identify its K^m nearest intra-modal neighbors as $\mathcal{N}(v_i^m)$.

$$\mathcal{N}(v_i^m)=\{v_j\in\mathcal{V}^m|i\neq j,v_j\in\text{Top}K^m[\text{sim}(v_i,v_j)]\}. \quad (13)$$

To lay the foundation for the construction of inter-modal edges, we also identify inter-modal neighbor set $\mathcal{N}(v_i^M)$.

$$\begin{aligned}\mathcal{N}(v_i^M)=&\{v_j\in\mathcal{V}^T|v_i\in\mathcal{V}^P,v_j\in\text{Top}K^M[\text{sim}(v_i,v_j)]\}\\&\cup\{v_j\in\mathcal{V}^P|v_i\in\mathcal{V}^T,v_j\in\text{Top}K^M[\text{sim}(v_i,v_j)]\},\end{aligned} \quad (14)$$

where we use the Euclidean distance as a measure of similarity. Based on the constructed neighbor sets, we then construct the intra-modal edge set $\mathcal{E}^{intra}=\{\mathcal{E}^P,\mathcal{E}^T\}$, and the inter-modal edge set $\mathcal{E}^{inter}$. Building upon these structures, we construct the multi-modal graph $G=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$, $\mathcal{E}=\{\mathcal{E}^{intra},\mathcal{E}^{inter}\}$ represent the set of nodes and edges.

Multi-Modal Graph Representation Learning. Vision GNN (ViG) (Han et al. 2022) is proposed to effectively capture the relationships between patches. In particular, graph-based architectures excel at modeling complex interactions among different types of nodes. Therefore, employing ViG to model intra-modal and inter-modal relationships holds significant potential. Given M patch representations and and $|\mathcal{C}_y^{super}|\cdot N$ optimized prompt representations, we employ ViG to obtain the aggregated node-level representations.

$$\mathbf{H}^{node}=\text{ViG}(G)=\{\boldsymbol{h}_1,\cdots,\boldsymbol{h}_M,\cdots,\boldsymbol{h}_{|\mathcal{C}_y^{super}|\cdot N+M}\}. \quad (15)$$

Semantic supervision, aggregated from optimized prompt representations, is propagated to the visual branch. To analyze the impact of visual representations on OOD detection, we slice the node-level output to get patch-related components $\{\boldsymbol{h}_i\}_{i=1}^M$. A global pooling operation is then applied to convert these components into graph-level representation.

$$\mathbf{H}^{global}=\text{Pooling}(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_M). \quad (16)$$

The graph-level representation is subsequently utilized to enhance energy-based OOD detection through cross-entropy loss $\mathcal{L}_{cls}$ and energy regularization loss $\mathcal{L}_{energy}$:

$$\min_{\theta\in\Theta}\mathbb{E}_{(G,y)\sim\mathcal{D}_{tr}}\big[\underbrace{-\log f_y(G)}_{\mathcal{L}_{cls}}+\lambda\cdot\underbrace{\text{ReLU}\big(E(G)-m_{in}\big)^2}_{\mathcal{L}_{energy}}\big], \quad (17)$$

where $E(G;f)=-T\cdot\log\sum_{i=1}^N e^{f_i(G)/T}$, and m_{in} represents the margins hyperparameter. During inference, the test image is first split into patches by the pre-trained ViT, and subsequently processed by the well-trained ViG to compute the final confidence score for OOD detection.

Experiments

Experimental Details

Datasets. We conduct experiments on two benchmarks:

- The small-scale `CIFAR-100` (Krizhevsky, Hinton et al. 2009). Following common practice (Huang and Li 2021), the OOD datasets used are `CIFAR-10` (Krizhevsky, Hinton et al. 2009), `SVHN` (Netzer et al. 2011), `Texture` (Cimpoi et al. 2014) and `Places365` (Zhou et al. 2017).
- The large-scale `ImageNet-1K` (Deng et al. 2009). The OOD datasets used are `iNaturalist` (Van Horn et al. 2018), `SUN` (Xiao et al. 2010), `Places` (Zhou et al. 2017), and `Texture` (Cimpoi et al. 2014).

Baselines. We compare our method with two types of baselines: zero-shot methods and methods that require additional training. For zero-shot methods, we select Energy (Liu et al. 2020), MCM (Ming et al. 2022), CLIPN (Wang et al. 2023), Neglabel (Jiang et al. 2024), among others. For methods that require training, we compare with MaxLogit (Hendrycks et al. 2019), CoOp (Zhou et al. 2022), NegPrompt (Li et al. 2024b), etc. Note that global pooling and subsequent training are applied exclusively to the aggregated image representations, with no textual features involved. More details are in Baselines section in the appendix.

Experimental Settings. All experiments are conducted using the pre-trained CLIP (ViT-B/16) as the backbone, with encoder parameters frozen during training. For category feature generation, we employ several commonly used LLMs: GPT-4o, DeepSeek-R1, Gemini 2.0 Pro, OpenAI o1,

	CIFAR-10			SVHN			Texture			Places365			ID-Acc↑
	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	
Energy	84.04±1.12	82.72±1.36	59.16±2.07	88.20±2.60	89.72±1.82	72.54±2.68	80.43±1.94	85.12±2.73	65.55±0.45	83.47±1.45	80.45±1.68	59.86±2.53	78.30±0.87
MCM	83.08±2.31	73.41±1.74	78.36±2.14	89.96±2.12	88.72±0.20	64.45±2.77	73.61±0.59	82.10±2.01	90.30±1.34	61.37±0.92	60.91±1.35	98.42±2.28	76.34±1.08
CLIPN	88.66±1.08	89.12±0.14	50.67±2.31	88.20±0.17	49.82±0.14	71.72±2.82	90.92±1.65	93.38±1.89	37.74±0.91	87.25±1.71	86.16±0.94	51.06±1.87	79.62±0.75
Neglabel	77.60±1.34	78.34±1.39	72.09±2.17	93.15±0.91	86.81±1.32	22.78±2.49	90.40±1.45	91.31±2.47	56.11±1.08	89.74±1.79	79.35±0.97	40.86±1.52	76.20±0.93
MSP	78.31±1.21	79.52±1.32	81.82±2.03	76.04±1.12	60.76±1.10	83.69±0.94	76.93±1.23	85.24±0.76	83.83±1.42	79.44±1.38	62.39±1.41	81.24±1.79	77.13±0.83
ODIN	78.18±1.13	79.12±1.41	83.16±1.92	71.08±2.85	52.36±1.37	89.76±1.56	79.39±2.01	86.67±1.35	78.37±2.58	79.83±2.01	60.85±1.93	81.27±0.74	76.92±2.91
GradNorm	71.33±0.97	67.28±2.14	82.32±2.52	71.32±2.20	50.77±2.36	79.72±1.64	64.75±2.34	70.58±1.97	82.15±1.83	69.64±1.67	36.36±0.82	81.98±2.58	77.99±1.04
ReAct	76.62±1.35	78.97±0.23	70.81±2.25	83.73±1.80	76.43±0.60	77.41±0.55	81.73±1.45	89.01±2.29	76.76±0.67	79.63±2.44	59.44±1.79	79.18±0.96	75.80±0.94
VIM	71.50±1.42	71.39±1.65	88.00±1.84	81.20±0.47	72.82±1.54	82.79±2.91	87.41±0.88	92.15±1.14	55.90±2.36	75.76±0.85	56.24±2.07	83.85±1.35	72.31±1.15
KNN	76.52±2.26	73.01±1.82	82.11±2.16	82.21±0.59	71.46±1.78	74.27±2.33	83.81±2.73	89.44±0.63	66.40±1.87	79.10±1.52	57.47±1.74	78.74±2.36	69.92±1.24
MaxLogit	85.07±1.29	78.13±1.46	57.58±1.93	91.01±1.80	90.14±0.51	59.05±1.79	82.08±1.31	87.93±1.72	62.82±2.06	80.88±2.08	69.48±0.79	65.58±1.62	79.48±0.96
EOE	85.67±1.24	80.32±0.37	59.17±2.06	88.78±0.06	86.10±2.85	68.47±0.27	82.64±2.58	90.31±1.39	66.89±0.79	78.06±2.13	61.14±1.59	77.60±0.81	78.82±0.85
VOS	79.23±1.16	80.03±1.28	58.87±2.12	85.01±2.91	74.34±2.90	47.44±0.59	78.26±0.84	85.56±1.57	61.36±2.41	79.71±1.26	61.41±2.41	57.17±1.64	77.04±2.91
CoOp	76.01±1.47	74.12±1.52	67.30±2.43	87.97±2.50	84.79±2.43	39.89±0.14	69.95±1.67	86.34±0.29	86.36±1.75	56.78±0.69	56.34±1.97	91.63±2.51	72.83±1.12
CoCoOp	76.85±1.33	74.98±2.56	65.73±2.35	90.50±0.64	88.96±0.91	31.80±0.98	71.55±1.28	84.19±1.49	85.80±2.68	58.45±1.94	54.19±2.58	91.37±1.72	71.97±1.05
LoCoOp	77.67±2.28	76.24±1.49	83.61±1.19	93.89±0.55	92.39±0.29	29.27±1.17	78.61±0.42	93.52±2.06	71.17±1.57	48.08±1.35	48.20±0.89	97.59±1.47	72.34±1.11
NegPrompt	88.53±1.09	89.01±1.18	49.96±2.22	94.19±0.55	92.18±2.05	30.99±0.81	91.14±0.92	93.11±1.79	24.13±0.84	88.34±2.02	90.16±1.54	44.31±2.65	78.64±0.89
PNPS-DS	89.06±1.07	90.16±1.11	60.86±2.31	94.88±0.87	94.72±0.10	30.37±1.63	95.18±0.95	**98.98**±2.48	22.39±1.64	90.23±0.94	89.99±1.62	**28.62**±2.55	80.96±1.88
PNPS-Gem	86.65±1.14	87.70±1.22	60.15±2.14	94.86±1.84	94.36±2.73	29.77±0.42	95.32±1.28	98.91±1.56	20.58±0.71	92.84±2.08	92.69±1.35	33.42±1.46	82.03±0.81
PNPS-OAI	**89.72**±0.97	90.33±1.06	**49.08**±1.88	95.12±0.42	94.65±0.78	27.84±2.41	93.64±2.48	98.48±0.66	28.19±1.89	92.42±1.78	92.62±2.09	35.70±0.93	80.93±0.86
PNPS-Clau	87.69±1.10	86.70±0.96	55.34±2.25	94.30±0.88	93.49±1.99	30.99±0.22	94.44±1.38	98.71±2.14	29.78±0.92	91.80±1.26	91.84±2.54	35.81±1.39	82.03±0.81
PNPS-GPT	89.63±0.98	**90.34**±1.05	49.75±2.31	**96.29**±1.30	**95.97**±1.49	**19.92**±0.84	**95.55**±1.74	98.94±1.09	**18.83**±1.56	**93.71**±1.47	**93.56**±2.41	30.23±1.73	78.53±1.34

Note: Due to space constraints, we abbreviate LLMs as follows: DeepSeek (DS), Gemini (Gem), OpenAI o1 (OAI), Claude (Clau), and GPT-4o (GPT).

Table 1: OOD detection results on the small-scale CIFAR-100 dataset, in terms of AUROC, AUPR, and FPR95 (mean ± std). The best results are highlighted in **bold**, while the second-best results are underlined. ↑ (↓) indicates that the larger (smaller) values are better. The first section are zero-shot methods, while the second section are training-required methods.

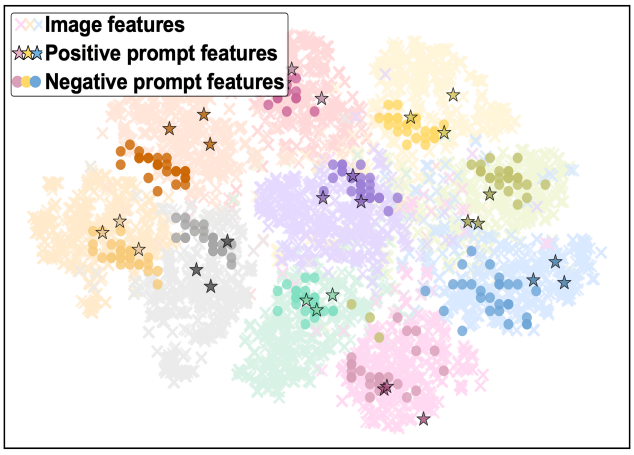


Figure 3: T-SNE visualization of optimized image features, along with positive and negative prompt features on the CIFAR-10 dataset in the shared semantic space.

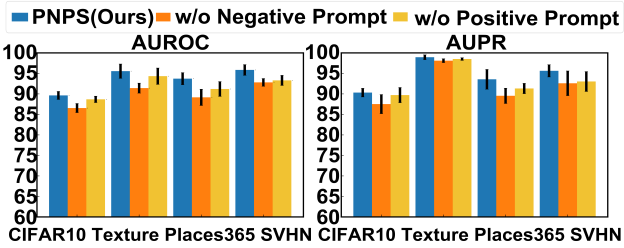


Figure 4: Performance in terms of AUROC, AUPR, and FPR95 under different settings on CIFAR-100 dataset.

and Claude 3.7 Sonnet. The hyperparameters for the positive and negative losses, λ^+ and λ^-, are set to $1e-5$ and $1e-3$, respectively. The number of super-classes is set to 20 for CIFAR-100 and 50 for ImageNet-1K. For the ViG model, we use the isotropic architecture with 4 and 5 interactive GCN layers for the two datasets, respectively. The Top-K values $\{K^T, K^P, K^M\}$ are set to $\{2, 10, 8\}$ for CIFAR-100 and $\{2, 20, 18\}$ for ImageNet-1K. The margin m_{in} is set to 10 and 12 for two datasets. All experiments are performed on two NVIDIA A800 GPUs. Following common practice (Huang and Li 2021), we evaluate our method using AUROC, AUPR, FPR95, and ID-Acc.

Experimental Results

We report the OOD detection results on the CIFAR-100 dataset in Table 1. The average AUROC scores for five different LLMs are 92.34%, 92.42%, 92.73%, 92.06%, and 93.80%, respectively. These results are consistent, with a minor variance of $3e-5$, suggesting that the choice of LLMs has little impact on overall performance, and that subsequent optimization and training steps are more critical. Notably, GPT-4o achieves the best overall performance among the five LLMs, underscoring its relative advantage in text understanding and the generation of discriminative content. Among zero-shot baselines, while CLIPN achieves competitive results by leveraging negative prompts to empower the logic of saying "no" within CLIP, it may introduce overlapping and noisy OOD features. In contrast, our method guides negative prompts to focus on ID features near category boundaries, resulting in higher AUROC, with improvements of 0.97% on CIFAR-10, 8.09% on SVHN, 4.63% on Texture, and 6.46% on Places365 over CLIPN.

Among the training-based baselines, NegPrompt, which captures negative semantics relative to ID classes, achieves the best overall performance. On the CIFAR-10, SVHN, Texture, and Places365 datasets, our PNPS outperforms it in terms of AUROC by 1.19%, 1.65%, 4.41%, and 5.37%, respectively. While our model achieves comparable performance to NegPrompt on the more challenging datasets (CIFAR-10 and SVHN), it demonstrates a substantially greater advantage on the simpler datasets (Texture and Places365). These results indicate that our method not only maintains robustness in difficult scenarios but also excels in less complex tasks. The results for ImageNet-1K

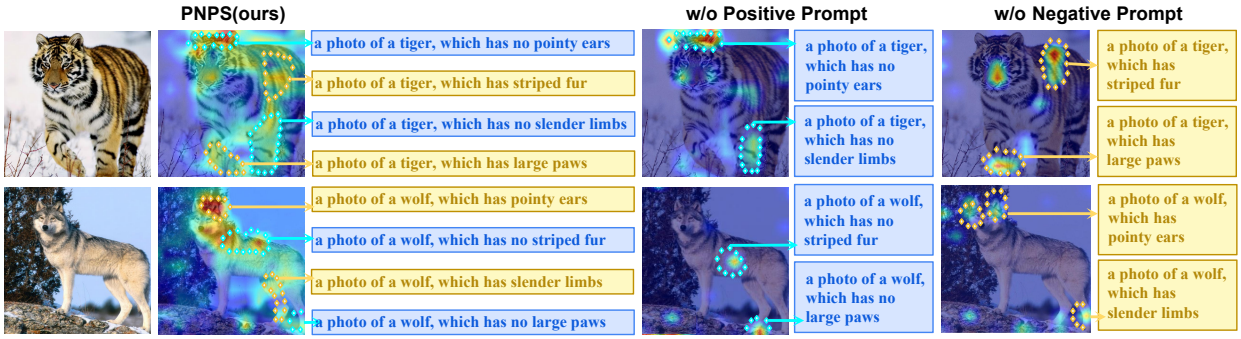
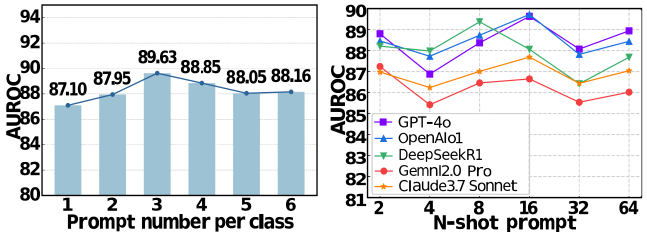


Figure 5: Grad-CAM maps visualization of features captured by full, without positive, and without negative prompts.



(a) AUROC for feature count. (b) AUROC for N-shot learning.

Figure 6: Left: AUROC for LLMs-generated feature count. Right: AUROC for prompt learning under N-shot settings.

are provided in Experimental Results in the appendix.

Experimental Analysis

Following the above visualization in the shared semantic space, we further analyze the specific visual features learned by the optimized positive and negative prompts for a given image. As shown in Figure 5, we present Grad-CAM visualization of the image regions highlighted by both prompts, as well as by negative prompts alone and positive prompts alone. It is evident that combining both prompts enables the model to focus on a broader range of feature regions. Moreover, the visual regions highlighted by the prompts align well with their textual semantics, demonstrating the effectiveness of our optimization. Specifically, positive prompts tend to focus on distinctive category features, such as the striped fur of a tiger, whereas negative prompts emphasize the absence of these distinctive features in other categories, thereby highlighting inter-class differences.

Ablation Study

The Effectiveness of Negative and Positive Prompts. We conduct ablation experiments to evaluate the contribution of positive and negative prompt representations by removing each component individually. The conditions "w/o positive prompt" and "w/o negative prompt" illustrate the impact of excluding positive and negative representations, as shown in Figure 4. Our results demonstrate that retaining only negative representations yields better performance than retaining only positive representations. This finding, together with the t-SNE visualization in Figure 3, further validates that negative representations are more effective than positive representations in assisting the classifier to define decision boundaries between categories for OOD detection.

The Effectiveness Graph-Based Connection. Graph-based architectures inherently support information propagation between heterogeneous nodes, making them a more direct and efficient solution for modeling complex intra-modal and inter-modal relationships. To validate the effectiveness of graph-based aggregation and propagation, we directly employ the optimized positive and negative representations, together with the MCM score (Ming et al. 2022), for VLM-based OOD detection. We conduct experiments on the `CIFAR-100` and `ImageNet-1K` datasets, and report the relevant results in the appendix. Compared to the zero-shot CLIPN method, our approach achieves slightly lower performance when the graph-based structure is omitted. However, after incorporating graph connections, its performance improves significantly. These findings demonstrate that aggregating and propagating semantic supervision via graph-based structures can effectively enhance OOD detection.

Conclusion

In this work, we propose a three-phase PNPS framework. Initially, LLMs are leveraged to construct class-specific positive and negative prompts. Subsequently, these prompts are optimized via learnable textual matrics to capture intra-class and inter-class features, which facilitate comprehensive learning of ID features and clearer category boundaries. Additionally, a graph-based model aggregates semantic supervision from the optimized prompt representations and transfers it to image features, thus enhancing OOD detection performance. Experiments on two ID datasets and eight OOD datasets using five different LLMs demonstrate that our approach outperforms state-of-the-art baselines.

Acknowledgments

This work done by Zhixia He, Qin Tian, and Minglai Shao is supported by the National Natural Science Foundation of China (No. 62272338) and the Research Fund of the Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education (EBME25-F-06). Chen Zhao, Xintao Wu, Xujiang Zhao, Dong Li, and Linlin Yu did not receive any financial support for this work and contributed only by developing the research ideas, participating in discussions, and providing feedback on the manuscript.

References

Bogdoll, D.; Nitsche, M.; and Zöllner, J. M. 2022. Anomaly detection in autonomous driving: A survey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4488–4499.

Cao, C.; Zhong, Z.; Zhou, Z.; Liu, Y.; Liu, T.; and Han, B. 2024. Envisioning outlier exposure by large language models for out-of-distribution detection. *arXiv preprint arXiv:2406.00806*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *arXiv preprint arXiv:2110.04544*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Han, K.; Wang, Y.; Guo, J.; Tang, Y.; and Wu, E. 2022. Vision gnn: An image is worth graph of nodes. *Advances in neural information processing systems*, 35: 8291–8303.

Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2019. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-distribution Detection for Large Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jiang, X.; Liu, F.; Fang, Z.; Chen, H.; Liu, T.; Zheng, F.; and Han, B. 2024. Negative label guided ood detection with pretrained vision-language models. *arXiv preprint arXiv:2403.20078*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, D.; Wan, G.; Wu, X.; Wu, X.; Chen, X.; He, Y.; Lian, C. G.; Sorger, P. K.; Semenov, Y. R.; and Zhao, C. 2025a. Multi-Modal Foundation Models for Computational Pathology: A Survey. *arXiv preprint arXiv:2503.09091*.

Li, D.; Wan, G.; Wu, X.; Wu, X.; Nirmal, A. J.; Lian, C. G.; Sorger, P. K.; Semenov, Y. R.; and Zhao, C. 2025b. A Survey on Computational Pathology Foundation Models: Datasets, Adaptation Strategies, and Evaluation Tasks. *arXiv preprint arXiv:2501.15724*.

Li, D.; Zhao, C.; Shao, M.; and Wang, W. 2024a. Learning fair invariant representations under covariate and correlation shifts simultaneously. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1174–1183.

Li, T.; Pang, G.; Bai, X.; Miao, W.; and Zheng, J. 2024b. Learning transferable negative prompts for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17584–17594.

Lin, Y.; Li, D.; Shao, M.; Wan, G.; and Zhao, C. 2025a. FADE: Towards Fairness-aware Generation for Domain Generalization via Classifier-Guided Score-based Diffusion Models. *In Proceedings of the International Joint Conference on Artificial Intelligence*.

Lin, Y.; Li, D.; Wu, X.; Shao, M.; Zhao, X.; Chen, Z.; and Zhao, C. 2025b. Face4FairShifts: A Large Image Benchmark for Fairness and Robust Learning across Visual Domains. *arXiv preprint arXiv:2509.00658*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.

Menon, S.; and Vondrick, C. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35: 35087–35102.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 4. Granada.

Nie, J.; Zhang, Y.; Fang, Z.; Liu, T.; Han, B.; and Tian, X. 2024. Out-of-distribution detection with negative prompts. In *The twelfth international conference on learning representations*.

Parmar, J.; Chouhan, S.; Raychoudhury, V.; and Rathore, S. 2023. Open-world machine learning: applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10): 1–37.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15691–15701.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Shao, M.; Li, D.; Zhao, C.; Wu, X.; Lin, Y.; and Tian, Q. 2024. Supervised algorithmic fairness in distribution shifts: A survey. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 8225–8233.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34: 144–157.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

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

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4921–4930.

Wu, X.; Chen, X.; Wu, X.; Li, D.; Chen, Z.; He, Y.; and Zhao, C. 2025. Explainable Image-Centric Forgery Detection: A Survey. *Authorea Preprints*.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 3485–3492. IEEE.

Zhang, C.; Song, D.; Huang, C.; Swami, A.; and Chawla, N. V. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 793–803.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-Language Models for Vision Tasks: A Survey. arXiv:2304.00685.

Zhao, C.; Chen, F.; and Thuraisingham, B. 2021. Fairness-aware online meta-learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2294–2304.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision (IJCV)*.

Zimmerer, D.; Full, P. M.; Isensee, F.; Jäger, P.; Adler, T.; Petersen, J.; Köhler, G.; Ross, T.; Reinke, A.; Kascenas, A.; et al. 2022. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE transactions on medical imaging*, 41(10): 2728–2738.