Cross-modal Proxy Evolving for OOD Detection with Vision-Language Models

Hao Tang[1], **Yu Liu**[1], **Shuanglin Yan**[2*], **Fei Shen**[3], **Shengfeng He**[4], **Jing Qin**[1]
[1] Centre for Smart Health, The Hong Kong Polytechnic University
[2] College of Information Science and Technology, Nanjing Forestry University
[3] NExT++ Research Center, National University of Singapore
[4] School of Computing and Information Systems, Singapore Management University
{howard-hao.tang, leo-yu.liu, shuanglin.yan, harry.qin}@polyu.edu.hk, shenfei29@nus.edu.sg, shengfenghe@smu.edu.sg

Abstract

Reliable zero-shot detection of out-of-distribution (OOD) inputs is critical for deploying vision-language models in open-world settings. However, the lack of labeled negatives in zero-shot OOD detection necessitates proxy signals that remain effective under distribution shift. Existing negative-label methods rely on a fixed set of textual proxies, which (i) sparsely sample the semantic space beyond in-distribution (ID) classes and (ii) remain static while only visual features drift, leading to cross-modal misalignment and unstable predictions. In this paper, we propose **CoEvo**, a training- and annotation-free test-time framework that performs bidirectional, sample-conditioned adaptation of both textual and visual proxies. Specifically, CoEvo introduces a *proxy-aligned co-evolution* mechanism to maintain two evolving proxy caches, which dynamically mines contextual textual negatives guided by test images and iteratively refines visual proxies, progressively realigning cross-modal similarities and enlarging local OOD margins. Finally, we dynamically re-weight the contributions of dual-modal proxies to obtain a calibrated OOD score that is robust to distribution shift. Extensive experiments on standard benchmarks demonstrate that CoEvo achieves state-of-the-art performance, improving AUROC by 1.33% and reducing FPR95 by 45.98% on ImageNet-1K compared to strong negative-label baselines.

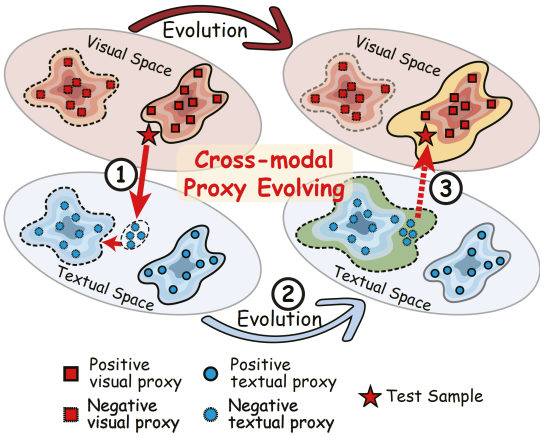

Figure 1: **Proxy-aligned co-evolution.** For each test sample, textual negatives are dynamically mined to expand the occupied space around its semantic context, while visual positive/negative proxies are updated online. This joint evolution maintains aligned cross-modal similarities under distribution shift, enabling robust zero-shot OOD decisions.

Introduction

Machine learning systems deployed in real-world environments (Wu et al. 2024a,b) frequently encounter inputs from previously unseen classes, commonly referred to as *out-of-distribution* (OOD) data. These inputs often differ significantly from the pre-defined *in-distribution* (ID) categories observed during training. When presented with such data, models tend to produce overconfident yet incorrect predictions (Scheirer et al. 2013; Nguyen, Yosinski, and Clune 2015), posing substantial safety and reliability concerns in high-stakes applications such as healthcare and autonomous driving. OOD detection aims to mitigate these risks by identifying and rejecting OOD inputs, thereby enhancing the robustness of downstream decision-making systems.

Traditional vision-based OOD detection methods primarily operate within the visual domain, relying solely on image features (Tang et al. 2020, 2022, 2023) while overlooking the rich semantic information embedded in class labels (Hendrycks and Gimpel 2017; Lee et al. 2018; Liang, Li, and Srikant 2018). Recent advances in vision-language pre-training, particularly CLIP (Radford et al. 2021), have enabled a multi-modal paradigm (Tang et al. 2025) that leverages both visual and textual information in zero-shot OOD detection (Esmaeilpour et al. 2022; Ming et al. 2022; Wang et al. 2023). Within this paradigm, a prominent line of work, known as *negative-label* methods, constructs a pool of textual labels that are semantically dissimilar from ID classes and employs them as textual proxies for "not-ID" concepts. For example, NegLabel (Jiang et al. 2024) selects negatives from WordNet (Fellbaum 1998), while CSP (Chen, Gao, and Xu 2024) augments labels with descriptive adjectives to synthesize semantically unrelated superclasses. A test image is then classified as OOD if its similarity to these negative labels surpasses that to ID labels.

While negative-label approaches have shown promise, their static design introduces two fundamental limitations:

*Corresponding author.

(i) Unmodeled negative space: A globally fixed negative set sparsely samples the vast semantic space beyond ID classes, leaving many informative, sample-specific negatives unrepresented during inference. **(ii) Modality misalignment:** Under test-time distribution shift, visual features shift to the new domain, whereas textual negatives remain fixed to presetting priors. This desynchronization distorts the cross-modal similarity geometry and destabilizes decision thresholds. Recent work such as AdaNeg (Zhang and Zhang 2024) partially addresses the first issue by constructing visual proxies using encountered test samples and jointly leveraging visual-textual evidence during scoring. However, it still employs *fixed textual negatives*: the negative words are preselected (e.g., far from ID labels in the text space) and kept fixed during inference. That is, adaptation is inherently *one-sided*: visual proxies adapt to test data, whereas textual negatives remain static. Consequently, cross-modal geometry is only partially realigned, and a substantial portion of the negative space remains unmodeled.

We argue that robust zero-shot OOD detection requires *bidirectional, sample-conditioned adaptation* of both modalities. Specifically, textual negatives should dynamically adapt to the current test sample and domain, while visual proxies should expose OOD structure as data arrive,without updating the backbone parameters or relying on labeled OOD samples. This motivates our proposed *proxy-aligned co-evolution* mechanism (Fig. 1), where visual cues guide the mining of contextual textual negatives, and the updated textual proxies, in turn, refine the visual decision boundary in a closed-loop manner.

To instantiate this principle, we introduce `CoEvo` (*Cross-modal Proxy Evolving*), a test-time zero-shot OOD detection framework illustrated in Fig. 2. `CoEvo` maintains two modality-specific proxy caches (textual and visual), each organized as positive/negative queues that are updated online through the proposed *proxy-aligned co-evolution* mechanism. These updates form a closed-loop process that iteratively aligns cross-modal similarities and enlarges local OOD margins. Additionally, we adaptively re-weight contributions from the dual-modal proxies to produce a calibrated OOD score. Extensive experiments on standard benchmarks validate the effectiveness of `CoEvo`. Notably, on the large-scale ImageNet dataset, `CoEvo` achieves a 1.33% improvement in AUROC and a 45.98% reduction in FPR95 over the best-performing negative label-based baselines.

We summarize our contribution as follows:

- We propose `CoEvo`, a zero-shot OOD detection framework that constructs semantically aligned ID/OOD proxy caches at test time by jointly leveraging visual and textual modalities.
- We introduce a proxy-aligned co-evolution mechanism that performs sample-conditioned, bidirectional adaptation of modality-specific proxies, mitigating cross-modal misalignment under distribution shift.
- Extensive experiments on widely used large-scale benchmarks show state-of-the-art performance; e.g., on ImageNet, `CoEvo` improves AUROC by 1.33% and reduces FPR95 by 45.98% over strong negative-label baselines.

Related Work

OOD Detection with Visual Modality. Existing visual OOD detection approaches can be broadly classified into three categories: score-based (Huang and Li 2021; Wang et al. 2022; Sun, Guo, and Li 2021), distance-based (Tack et al. 2020; Du et al. 2022; Ming et al. 2023), and generative-based methods (Ryu et al. 2018; Kong and Ramanan 2021). Among them, score-based methods are particularly prominent, as they introduce various scoring mechanisms to effectively discriminate between ID and OOD samples. Representative scoring strategies include confidence-based (Sun, Guo, and Li 2021), discriminator-based (Kong and Ramanan 2021), energy-based (Liu et al. 2020; Wang et al. 2021), and gradient-based scores (Huang, Geng, and Li 2021). In contrast, distance-based methods identify OOD samples by calculating distances between the test sample and the nearest ID sample (Tack et al. 2020), or distances to precomputed ID prototypes (Tao et al. 2023). Common metrics in this category include KNN (Sun et al. 2022; Ming et al. 2023) and RBF kernels (van Amersfoort et al. 2020). Despite the notable progress achieved, conventional single-modal visual OOD detection methods generally rely on manually labeled ID images and largely overlook the potential benefits derived from textual information integration.

OOD Detection with Dual Modalities. Recent zero-shot OOD detection methods increasingly leverage vision-language models (VLMs) to incorporate semantic cues (Li et al. 2025; Tang, He, and Qin 2025) from textual information. Early approaches such as ZOC (Esmaeilpour et al. 2022) and CLIPN (Wang et al. 2023) employ textual descriptions or auxiliary encoders to detect OOD samples but rely solely on positive in-distribution (ID) labels, often producing overly optimistic similarity scores for unknown classes. To alleviate this limitation, several post-hoc scoring mechanisms have been proposed to enhance OOD discrimination. For instance, MCM (Ming et al. 2022) computes the maximum softmax score over CLIP similarities, while NegLabel (Jiang et al. 2024) refines this approach by mining negative samples from external sources. CSP (Chen, Gao, and Xu 2024) expands the label space with a conjugated semantic pool, and AdaNeg (Zhang and Zhang 2024) dynamically adapts negative labels online using encountered OOD samples. However, these approaches construct textual negatives independently of the visual OOD features observed during inference, leading to a persistent *modality misalignment*: negative textual proxies fail to accurately capture the distribution of unseen images, thereby limiting detection robustness. Our method addresses this gap through a cross-modal proxy evolving framework that jointly refines textual and visual proxies, ensuring semantically consistent representations for improved zero-shot OOD detection.

Methodology

Problem Formulation. Let $\mathcal{Y}_{\text{ID}} = \{y_1, \dots, y_K\}$ denote the label set of K ID classes, and let $x \in \mathcal{X}$ represent an input image. Zero-shot OOD detection with vision–language models (Ming et al. 2022; Wang et al. 2023; Jiang et al. 2024), such as CLIP (Radford et al. 2021), aims to determine

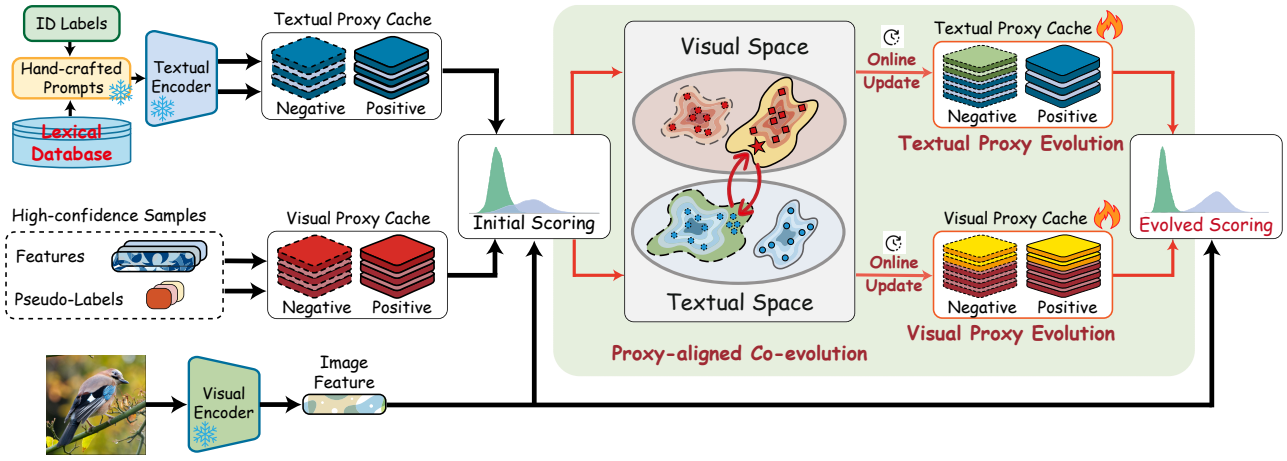

Figure 2: Pipeline of the proposed cross-modal proxy co-evolving framework (CoEvo). It dynamically updates both visual and textual proxy caches based on high-confidence samples, enabling bidirectional alignment and robust zero-shot OOD detection.

whether x belongs to any class in $\mathcal{Y}_{\text{ID}}$ or originates from an unseen class (OOD), without requiring any training samples or prompt tuning. The core objective is to design an ID confidence function $\mathcal{S} : \mathcal{X} \to \mathbb{R}$ that assigns higher values to ID samples than to OOD ones. Given a threshold $\theta \in \mathbb{R}$, the detector predicts *ID* if $S(\mathbf{x}) \geq \theta$ and *OOD* otherwise.

Overview

Existing negative label-based methods typically rely on *static textual proxies*, i.e., a fixed set of textual embeddings that serve as negative semantic anchors. However, these static proxies leave portions of the negative semantic region uncovered, i.e., an *unmodeled negative space*. Furthermore, due to these proxies remain unchanged at test time, they cannot adapt to image features that drift under distribution shift, which induces *modality misalignment* between image and text similarities. To address these limitations, we introduce *Cross-modal Proxy Evolving* (CoEvo), a test-time framework that constructs OOD proxies by jointly exploiting visual and textual modalities, as illustrated in Fig. 2. CoEvo maintains two online caches: a *textual proxy cache* that stores positive/negative textual proxies, and a *visual proxy cache* that stores positive/negative visual proxies. A *proxy-aligned co-evolution* mechanism then couples the two caches so that images teach text where the non-ID regions lie, and the updated text in turn regularizes visual decisions.

Textual Proxy Cache

Textual proxies offer a powerful means of encoding semantic priors for zero-shot OOD detection. Leveraging the expressive capability of pre-trained vision-language models (e.g., CLIP), we construct two complementary textual proxy queues: a *positive* proxy queue for ID classes, and a *negative* proxy queue for OOD concepts.

Positive Proxy Queue. Let the ID label set be defined as $\mathcal{Y}_{\text{ID}} = \{y_1, \ldots, y_K\}$, where each y_k denotes a semantic category (e.g., *cat*, *dog*, *bird*), and K denotes the number of ID classes. Each class name y_k is converted into a prompt-based textual embedding using a pre-trained CLIP text encoder as $\mathbf{f}_{\text{t},k} = \mathcal{E}_{\text{t}}(\mathcal{T}(y_k)) \in \mathbb{R}^D$, where $\mathcal{T}(\cdot)$ denotes the prompt template (e.g., "a photo of a <class>"), and D is the embedding dimension. We organize these embeddings into a fixed queue $\mathbf{T}_p \in \mathbb{R}^{K \times D}$, i.e., $\mathbf{T}_p[k] = \mathbf{f}_{\text{t},k}$. The static nature of $\mathbf{T}_p$ reflects the stable semantics of known categories.

Negative Proxy Queue. To model unknown OOD semantics, following NegLabel (Jiang et al. 2024), we initialize a negative textual proxy queue $\mathbf{T}_n \in \mathbb{R}^{M \times D}$ with M negative class embeddings sampled from a large-scale lexical corpus $\mathcal{D}$. Specifically, we define a disjoint set of negative labels $\mathcal{Y}_{\text{neg}} = \{\tilde{y}_1, \ldots, \tilde{y}_M\}$, where $\mathcal{Y}_{\text{neg}} \cap \mathcal{Y}_{\text{ID}} = \varnothing$, and obtain $\mathbf{T}_n = \mathcal{E}_{\text{t}}(\mathcal{T}(\mathcal{Y}_{\text{neg}})) \in \mathbb{R}^{M \times D}$. Unlike $\mathbf{T}_p$, the negative proxy cache $\mathbf{T}_n$ evolves during inference: new embeddings are dynamically incorporated based on test-time visual observations through our proposed proxy co-evolution mechanism. This adaptivity ensures that $\mathbf{T}_n$ remains responsive to emerging semantic patterns in unseen OOD inputs.

Textual OOD Scoring. Given a test image x, we obtain its visual embedding via the CLIP image encoder: $\mathbf{f}_{\text{v}} = \mathcal{E}_{\text{v}}(x) \in \mathbb{R}^D$. Assuming that ID inputs are more similar to positive textual proxies and dissimilar to negative ones, we define the textual OOD detection score as:

$$\mathcal{S}_{\text{T}}^{\text{pre}}(x) = \frac{\sum_{k=1}^{K} e^{\left(\text{sim}(\mathbf{f}_{\text{v}}, \mathbf{T}_p[k])/\tau\right)}}{\sum_{k=1}^{K} e^{\left(\text{sim}(\mathbf{f}_{\text{v}}, \mathbf{T}_p[k])/\tau\right)} + \sum_{m=1}^{M} e^{\left(\text{sim}(\mathbf{f}_{\text{v}}, \mathbf{T}_n[m])/\tau\right)}}, \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature scaling factor. Higher $\mathcal{S}_{\text{T}}^{\text{pre}}(x)$ indicates stronger alignment with known ID semantics, and lower values signal potential OOD samples.

Visual Proxy Cache

While textual proxies provide global semantic anchors, they are inherently limited in representing fine-grained visual variations of unseen data. In particular, zero-shot OOD inputs may exhibit appearance patterns that are poorly de-

scribed by label embeddings alone. To address this, we introduce *visual proxies* that directly encode instance-level image features and evolve online to complement textual proxies.

Positive Proxy Queue. We construct a positive visual proxy queue $\mathbf{V}_p \in \mathbb{R}^{K\times L\times D}$, where K is the number of ID classes and L denotes the number of stored visual instances per ID class. The queue is initially empty, except for the first slot of each class, which is initialized with the corresponding positive textual embedding from $\mathbf{T}_p \in \mathbb{R}^{K\times D}$, i.e., $\mathbf{V}_p[:,0,:] = \mathbf{T}_p$. This provides a semantically meaningful starting point before any labeled images are observed. As inference proceeds, high-confidence ID samples are inserted into the queue, progressively enriching class-specific visual proxies.

Negative Proxy Queue. Analogously, we maintain a negative visual proxy queue $\mathbf{V}_n \in \mathbb{R}^{M\times L\times D}$ aligned with the negative textual proxy queue $\mathbf{T}_n \in \mathbb{R}^{M\times D}$. Following AdaNeg (Zhang and Zhang 2024), test samples predicted as OOD with high confidence are enqueued, enabling $\mathbf{V}_n$ to capture the evolving OOD appearance space. A priority queue strategy discards low-similarity or outdated samples, ensuring proxies remain representative of the most informative shifts in the test distribution.

Visual Proxy Aggregation and Scoring. Given an input image x with embedding $\mathbf{f}_\text{v}$, we compute class-wise aggregated visual proxies via similarity-based attention over the L instances:

$$\mathbf{v}_k^p = \sum_{\ell=1}^{L} \frac{\exp\big(-\beta\big(1-\mathbf{f}_\text{v}\cdot(\mathbf{V}_p^{k,\ell})^\top\big)\big)}{\sum_{\ell'=1}^{L}\exp\big(-\beta\big(1-\mathbf{f}_\text{v}\cdot(\mathbf{V}_p^{k,\ell'})^\top\big)\big)}\,\mathbf{V}_p^{k,\ell}. \quad (2)$$

where β controls attention sharpness. Negative proxies $\mathbf{v}_m^n$ are obtained analogously from $\mathbf{V}_n$. The visual OOD score is then defined symmetrically to the textual score:

$$\mathcal{S}_\text{V}^\text{pre}(x) = \frac{\sum_{k=1}^{K} e^{(\text{sim}(\mathbf{f}_\text{v},\mathbf{v}_k^p)/\tau)}}{\sum_{k=1}^{K} e^{(\text{sim}(\mathbf{f}_\text{v},\mathbf{v}_k^p)/\tau)} + \sum_{m=1}^{M} e^{(\text{sim}(\mathbf{f}_\text{v},\mathbf{v}_m^n)/\tau)}}. \quad (3)$$

A higher score indicates stronger alignment with ID visual proxies, while lower scores suggest OOD samples.

Multi-modal OOD Score. To leverage the complementary strengths of textual and visual modalities, we combine their respective scores into a unified OOD detection score:

$$\mathcal{S}_\texttt{CoEvo}^\text{pre}(x) = \lambda\mathcal{S}_\text{T}^\text{pre}(x) + (1-\lambda)\mathcal{S}_\text{V}^\text{pre}(x), \quad (4)$$

where the hyperparameter $\lambda \in [0.5, 1)$ balances modality preference based on their reliability and effectiveness.

Proxy-Aligned Co-Evolution Mechanism

Static textual proxies provide a limited representation of the open-set negative space and are inherently sensitive to distributional shifts arising at test time. Adapting proxies with visual modality partially alleviates this issue but neglects complementary cross-modal cues that can enhance OOD discrimination. To overcome these limitations, we propose a *Proxy-Aligned Co-Evolution* mechanism, in which textual and visual proxy caches are dynamically refined through bidirectional interactions. This co-evolution process enables the proxies to remain mutually aligned and better capture the semantics of both in-distribution and previously unseen OOD instances.

Textual Proxy Evolution. Static textual proxies cannot faithfully track semantic shifts of test samples, so we evolve the textual negatives while guarding against error amplification. We first gate updates by a confidence margin around the adaptive threshold δ as in AdaND (Cao et al. 2025). Concretely, samples with $\mathcal{S}_\texttt{CoEvo}^\text{pre}(x) > \delta+\gamma(1-\delta)$ are treated as ID, whereas those with $\mathcal{S}_\texttt{CoEvo}^\text{pre}(x) < \delta-\gamma(1-\delta)$ are treated as OOD; samples within the margin are excluded from updates to avoid uncertain decisions.

We adapt only the negative proxy queue $\mathbf{T}_n$ and keep the positive proxy queue $\mathbf{T}_p$ fixed to preserve stable anchors for ID semantics and prevent drift toward spurious OOD cues. Let $\mathcal{D}$ denote a corpus of ℓ_2-normalized textual embeddings spanning a broad semantic vocabulary. Conditioned on the visual embedding $\mathbf{f}_\text{v}$, we retrieve two candidate sets:

$$\mathcal{N}_\text{near}(x) = \text{Top-N}_{e\in\mathcal{D}\setminus\mathbf{T}_n}\cos\big(\mathbf{f}_\text{v}, e\big), \quad (5)$$

$$\mathcal{N}_\text{far}(x) = \text{Top-N}_{e\in\mathcal{D}\setminus\mathbf{T}_n}\big(-\cos\big(\mathbf{f}_\text{v}, e\big)\big), \quad (6)$$

with a deduplication constraint that discards candidates whose textual labels already appear in $\mathbf{T}_n$. We then update $\mathbf{T}_n$ by

$$\mathbf{T}_n \leftarrow \begin{cases} \big[\mathbf{T}_n;\ \text{stack}(\mathcal{N}_\text{near}(x))\big], & \text{if OOD,} \\ \big[\mathbf{T}_n;\ \text{stack}(\mathcal{N}_\text{far}(x))\big], & \text{if ID.} \end{cases} \quad (7)$$

Intuitively, inserting semantically *near* negatives for predicted OOD samples tightens local open-set boundaries around the test sample, improving fine-grained separability from nearby ID proxies; inserting *far* negatives for predicted ID samples broadens the coverage of the negative space, strengthening global discrimination against unseen classes. After each update, we recompute the textual score $\mathcal{S}_\text{T}^\text{post}(x)$ via Eq. (1), enabling progressive refinement of the textual proxies without modifying backbone weights.

Visual Proxy Evolution. Updating textual proxies alters the shared semantic manifold across modalities. Without corresponding adjustments on the visual side, this shift yields misaligned decision boundaries and degraded OOD discrimination. To preserve cross-modal consistency while improving the representational quality of visual proxies, we adopt an instance-adaptive strategy that refines the visual proxy cache at test time.

Given the updated negative textual proxy queue $\mathbf{T}_n$, we expand the negative visual proxy queue from $\mathbb{R}^{M\times L\times D}$ to $\mathbb{R}^{(M+N)\times L\times D}$ to accommodate the newly exposed OOD semantics, where N matches the incremental textual negatives. Let $\{\mathbf{v}_k^p\}_{k=1}^K$ and $\{\mathbf{v}_m^n\}_{m=1}^{M+N}$ be the current visual proxies for ID and OOD (computed via Eq. (2)). For $\mathbf{f}_\text{v}$, we assign it to the most relevant proxy using soft similarity scores:

$$\mathbf{z}_k^p = \text{Softmax}\big(\cos(\mathbf{f}_\text{v}, \mathbf{v}_k^p)\big), \quad y_\text{id} = \arg\max_k \mathbf{z}_k^p, \quad (8)$$

$$\mathbf{z}_m^n = \text{Softmax}\big(\cos(\mathbf{f}_\text{v}, \mathbf{v}_m^n)\big), \quad y_\text{ood} = \arg\max_m \mathbf{z}_m^n. \quad (9)$$

Based on the preliminary multi-modal score from Eq. (4), the sample $\mathbf{f}_\text{v}$ is inserted into either the positive class-specific queue $\mathbf{V}_p^{y_\text{id}}$ (for ID samples) or the negative queue

$\mathbf{V}_n^{y_{\text{ood}}}$ (for OOD samples), enabling the proxy cache to adaptively track evolving data distributions. To ensure reliability and prevent proxy drift, we exploit the entropy $\mathcal{H}(\mathbf{z}) = -\sum_{i=1} \mathbf{z}_i \log \mathbf{z}_i$ as a confidence measure. A new sample is cached directly if space is available. Otherwise, it replaces the existing exemplar with the highest entropy (least confident) only if its own entropy is lower, ensuring that proxies are updated preferentially with high-confidence samples.

Following each update, we recompute the visual OOD detection score $\mathcal{S}_{\text{V}}^{\text{post}}(x)$ via Eq. (3), allowing the refined proxies to immediately influence detection outcomes. This instance-adaptive evolution maintains compact and representative visual proxies for ID classes, while preserving diverse and semantically rich proxies for OOD samples. Combined with textual proxy evolution, it ensures stable cross-modal alignment and enhances open-set discriminability under dynamic test-time conditions.

OOD Score Evolution. To ensure robust decision making during inference, we adapt the multi-modal OOD score to reflect the evolving proxy caches.

Pre-evolution scoring. Following Eq. (4), the initial score is computed as

$$\mathcal{S}_{\texttt{CoEvo}}^{\text{pre}}(x) = \lambda\, \mathcal{S}_{\text{T}}^{\text{pre}}(x) + (1-\lambda)\, \mathcal{S}_{\text{V}}^{\text{pre}}(x) \quad \lambda \in [0.5, 1). \tag{10}$$

The higher textual weight realises a *cold-start asymmetry*: stable semantic priors from pre-defined textual proxies are preferred while the visual proxies are still sparsely initialised.

Post-evolution scoring. After the preliminary ID/OOD assignment and subsequent proxy evolution, we recompute the unimodal scores $\mathcal{S}_{\text{T}}^{\text{post}}(x)$ and $\mathcal{S}_{\text{V}}^{\text{post}}(x)$ and fuse them in a symmetric manner:

$$\mathcal{S}_{\texttt{CoEvo}}^{\text{post}}(x) = (1-\lambda)\, \mathcal{S}_{\text{T}}^{\text{post}}(x) + \lambda\, \mathcal{S}_{\text{V}}^{\text{post}}(x). \tag{11}$$

The weight flip is driven by two observations: (i) *Cold-start asymmetry:* as discussed above, and (ii) *Post-adaptation reliability:* after evolution, the visual proxies, enriched with instance-specific samples, draw sharper local decision boundaries than their textual counterparts.

Final decision. Unless otherwise specified, $\mathcal{S}_{\texttt{CoEvo}}^{\text{pre}}(x)$ is used solely for proxy updates, while $\mathcal{S}_{\texttt{CoEvo}}^{\text{post}}(x)$ is employed for the final ID/OOD decision (see Algorithm 1).

Algorithm 1: Cross-modal Proxy Evolving (CoEvo)

Require: ID label set $\mathcal{Y}_{\text{ID}}$; text corpus $\mathcal{D}$ for negatives; test set $\mathcal{X}$; adaptive margin γ

Ensure: Final ID/OOD predictions $\hat{\mathcal{Y}}$ for all $x \in \mathcal{X}$

Initialize positive textual queue $\mathbf{T}_p$ from $\mathcal{Y}_{\text{ID}}$
Initialize negative textual proxy $\mathbf{T}_n$ from $\mathcal{D}$
Initialize visual proxy queues $\mathbf{V}_p, \mathbf{V}_n$ using $\mathbf{T}_p, \mathbf{T}_n$
for each sample $x \in \mathcal{X}$ **do**
  Compute textual score $\mathcal{S}_{\text{T}}^{\text{pre}}(x)$ (Eq. (1))
  Compute visual score $\mathcal{S}_{\text{V}}^{\text{pre}}(x)$ (Eq. (3))
  Fuse preliminary multi-modal score $\mathcal{S}_{\texttt{CoEvo}}^{\text{pre}}(x)$ (Eq. (4))
  Compute the adaptive threshold δ
  if $\mathcal{S}_{\texttt{CoEvo}}^{\text{pre}}(x) > \delta + \gamma(1-\delta)$ **then**
    Predict preliminary label $\hat{y}_{\text{id}}$
    Retrieve far textual negatives $\mathcal{N}_{\text{far}}(x)$ (Eq. (6))
    $\mathbf{T}_n \leftarrow \text{Enqueue}(\mathbf{T}_n, \mathcal{N}_{\text{far}}(x))$
    $\mathbf{V}_p^{\hat{y}_{\text{id}}} \leftarrow \text{Enqueue}(\mathbf{V}_p^{\hat{y}_{\text{id}}}, \mathbf{f}_{\text{v}})$
  else if $\mathcal{S}_{\texttt{CoEvo}}^{\text{pre}}(x) < \delta(1-\gamma)$ **then**
    Predict preliminary label $\hat{y}_{\text{ood}}$
    Retrieve near textual negatives $\mathcal{N}_{\text{near}}(x)$ (Eq. (5))
    $\mathbf{T}_n \leftarrow \text{Enqueue}(\mathbf{T}_n, \mathcal{N}_{\text{near}}(x))$
    $\mathbf{V}_n^{\hat{y}_{\text{ood}}} \leftarrow \text{Enqueue}(\mathbf{V}_n^{\hat{y}_{\text{ood}}}, \mathbf{f}_{\text{v}})$
  else
    Skip the proxy update for ambiguous sample
  end if
  Update post-evolution scores $\mathcal{S}_{\text{T}}^{\text{post}}(x), \mathcal{S}_{\text{V}}^{\text{post}}(x)$
  Fuse final score $\mathcal{S}_{\texttt{CoEvo}}^{\text{post}}(x)$ (Eq. (11))
  Final decision $\hat{y}$ from $\mathcal{S}_{\texttt{CoEvo}}^{\text{post}}(x)$
end for
return $\hat{\mathcal{Y}}$

Experiment

Experimental Setup

Datasets. Following prior work (Ming et al. 2022; Jiang et al. 2024), we conduct extensive experiments on the ImageNet-1K benchmark (Deng et al. 2009), where the large-scale ImageNet-1K dataset is used as ID source. Four commonly adopted datasets, iNaturalist (Horn et al. 2018), SUN (Xiao et al. 2010), Places (Zhou et al. 2017), and Textures (Cimpoi et al. 2014), are employed as OOD test sets. To further evaluate generalization under varying OOD difficulty, we adopt both Near-OOD (SSB-hard (Vaze et al. 2022), NINCO (Bitterwolf, Müller, and Hein 2023)) and Far-OOD (iNaturalist (Horn et al. 2018), Textures (Cimpoi et al. 2014), OpenImage-O (Wang et al. 2022)) settings, as defined by the OpenOOD benchmark (Zhang et al. 2023; Yang et al. 2022), where ImageNet serves as the shared ID dataset. In addition, we examine performance under imbalanced conditions between ID and OOD samples to assess the robustness of our approach in realistic deployment settings.

Evaluation Criteria. Following prior work (Ming et al. 2022), we adopt three standard metrics for evaluating OOD detection performance:(1) **FPR95:** the false positive rate of OOD samples when the true positive rate (TPR) of ID samples is at 95%; (2) **AUROC:** the area under the receiver operating characteristic curve, which measures the overall separability between ID and OOD samples; and (3) **ID ACC:** classification accuracy for ID samples.

Implementation Details. We employ the ViT-B/16 visual encoder pretrained by CLIP (Radford et al. 2021) as our backbone. The key hyperparameters are set as follows: the visual proxy queue length is $L = 10$; the temperature $\tau = 0.01$ in Eq. (1) and (3); the balancing weight $\lambda = 0.8$ in Eqs. (4) and (11); $N = 5$ in Eq. (5); and $\beta = 5.5$ in Eq. (2). The batch size is fixed at 128. The lexical database and corresponding negative mining strategy for the textual proxy cache are derived from two recent baselines: NegLabel (Jiang et al. 2024) and CSP (Chen, Gao, and Xu 2024), which initialize negative textual proxies with 10K and 9,493 OOD class names, respectively. Following NegLabel, we adopt the prompt template "The nice `cls`" to encode class names. All experiments are conducted on a

Methods	OOD datasets iNaturalist		Sun		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP (Hendrycks and Gimpel 2017)	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
Energy (Liu et al. 2020)	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm (Huang, Geng, and Li 2021)	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
NPOS (Tao et al. 2023)	96.19	16.58	90.44	43.77	89.44	45.27	88.80	46.12	91.22	37.93
ZOC (Esmaeilpour et al. 2022)	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
CLIPN (Wang et al. 2023)	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LoCoOp (Miyai et al. 2023)	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
LAPT (Zhang et al. 2024)	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
NegPrompt (Li et al. 2024)	98.73	6.32	95.55	22.89	93.34	27.60	91.60	35.21	94.81	23.01
Energy (Liu et al. 2020)	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
MCM (Ming et al. 2022)	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
NegLabel (Jiang et al. 2024)	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
CSP (Chen, Gao, and Xu 2024)	99.60	1.54	96.66	13.66	92.90	29.32	93.86	25.52	95.76	17.51
AdaNeg (Zhang and Zhang 2024)	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
CoEvo$_{\text{NegLabel}}$	99.81	0.53	**98.68**	**4.42**	**95.80**	**23.51**	**97.48**	**12.42**	**97.95**	**10.22**
CoEvo$_{\text{CSP}}$	**99.82**	**0.46**	98.61	4.68	95.58	25.83	97.38	12.78	97.85	10.94

Table 1: Comparison of OOD detection performance on ImageNet-1K. All methods utilize a CLIP ViT-B/16 encoder.

Method	FPR95 ↓		AUROC ↑		ACC ↑
	Near-OOD	Far-OOD	Near-OOD	Far-OOD	ID
GEN	–	–	78.97	90.98	81.59
AugMix$_{+\text{ReAct}}$	–	–	79.94	93.70	77.63
RMDS	–	–	80.09	92.60	81.14
SCALE	–	–	81.36	96.53	76.18
AugMix$_{+\text{ASH}}$	–	–	82.16	96.05	77.63
LAPT	58.94	24.86	82.63	94.26	67.86
MCM	79.02	68.54	60.11	84.77	66.28
NegLabel	69.45	23.73	75.18	94.85	66.82
CSP	73.14	21.52	74.88	95.87	67.35
AdaNeg	67.51	17.31	**76.70**	96.43	67.13
CoEvo$_{\text{NegLabel}}$	**64.64**	15.24	75.37	96.50	66.83
CoEvo$_{\text{CSP}}$	66.88	**14.47**	74.65	**96.70**	**67.36**

Table 2: Zero-shot OOD detection results on the OpenOOD benchmark, where ImageNet-1K is adopted as ID dataset.

single NVIDIA RTX 3090 GPU.

Main Results

Evaluation on ImageNet benchmark. As shown in Tab. 1, our method consistently outperforms existing methods, including both training-based methods (Hendrycks and Gimpel 2017; Liu et al. 2020; Huang, Geng, and Li 2021; Tao et al. 2023; Esmaeilpour et al. 2022; Wang et al. 2023; Miyai et al. 2023; Zhang et al. 2024; Li et al. 2024) and training-free methods (Liu et al. 2020; Ming et al. 2022; Jiang et al. 2024; Chen, Gao, and Xu 2024; Zhang and Zhang 2024). Specifically, CoEvo$_{\text{NegLabel}}$ achieves an average FPR95 of 10.22% and AUROC of 97.95%, outperforming the most competitive baseline by margins of 45.98% in FPR95 and 1.33% in AUROC, respectively.

Evaluation on OpenOOD benchmark. As illustrated in Tab. 2, our method achieves competitive performance under both Near-OOD and Far-OOD settings. Note that training-based baselines leverage the full ImageNet training set. Under Near-OOD conditions, **CoEvo**$_{\text{CSP}}$ obtains an average FPR95 of 66.88% and an AUROC of 74.65%, slightly underperforming AdaNeg (Zhang and Zhang 2024) in AUROC, indicating marginally reduced sensitivity to fine-grained OOD discrimination. Conversely, in Far-OOD scenarios, our method achieves a substantially lower average FPR95 of 14.47% and a high AUROC of 96.70%. Furthermore, our approach improves ID classification performance, achieving an average ID ACC of 67.36%, surpassing all competing training-free methods. These results collectively validate the effectiveness and robustness of our proposed framework across diverse OOD settings.

Proxy Evolution		Average	
Textual	Visual	FPR95 ↓	AUROC ↑
		24.97	94.56
✓		21.77	95.38
	✓	17.41	96.99
✓	✓	**10.22**	**97.95**

Table 3: Ablation study of proxy evolution mechanism on ImageNet-1K, evaluated across four standard OOD datasets.

Analyses and Discussions

Analysis of Proxy Evolution. We perform an ablation study to quantify the contribution of each component in the proposed proxy-aligned co-evolution mechanism (Tab.3). The baseline, NegLabel without any evolution step, achieves an average FPR95 of 24.97% and an AUROC of 94.56%. Activating textual evolution alone lowers the FPR95 to 21.77%, revealing that dynamically updating text proxies during test-time evolution enhances semantic alignment. Enabling visual evolution alone produces a greater gain, confirming that adapting visual proxies to the test distribution mitigates feature-space shifts. Combining both textual and visual evolution yields the best performance, an average FPR95 of 10.22% and an AUROC of 97.95%, highlighting the complementary strengths of cross-modal co-evolution for robust OOD detection.

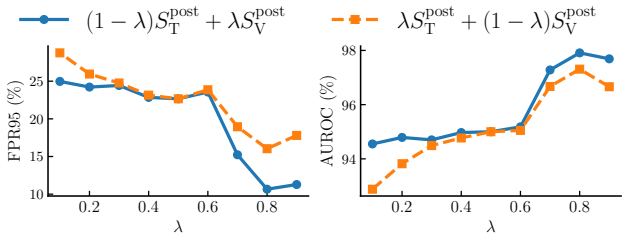

Figure 3: Sensitivity to the fusion weight λ. Results are reported on the ImageNet-1K benchmark.

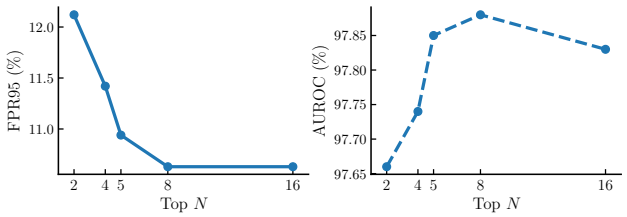

Figure 4: Sensitivity analysis of the hyperparameter Top-N, evaluated on the ImageNet-1K benchmark.

Analysis of Score Evolution. In Fig.3, we compare the proposed post-evolution fusion in Eq.(11) (solid line) against a no-flip variant that retains the pre-evolution weighting scheme, i.e., $\lambda \mathcal{S}_{\mathrm{T}}^{\mathrm{post}}(x) + (1-\lambda)\mathcal{S}_{\mathrm{V}}^{\mathrm{post}}(x)$ (dashed line). Low-λ regime (0.1–0.4): Eq. (4) allocates higher weight to visual scores; however, early in evolution the visual cache is sparsely populated, producing noisy estimates and unstable OOD separation. The no-flip strategy inherits this bias and overemphasizes unreliable visual evidence, while the flipped rule shifts preference toward the more stable textual cues, achieving better performance. High-λ regime (0.6–0.9): As evolution progresses, the visual proxies become richer and more discriminative due to the accumulation of diverse samples. The flipped rule adaptively assigns greater weight to these refined visual scores, surpassing the no-flip variant, with the performance gap peaking around $\lambda = 0.8$. These results demonstrate that fixed weight retention fails to adapt to the evolving reliability of modalities, whereas our score evolution mechanism dynamically aligns fusion weights with proxy quality, consistently improving OOD detection performance.

Analysis of Hyper-parameter λ. We conduct an ablation study on the fusion weight λ in Eq. (11). As shown in Fig. 3, performance on ImageNet-1K first increases with λ, peaking at $\lambda = 0.8$, and then slightly degrades as λ approaches 1.0. This trend indicates that a moderate emphasis on the visual score, combined with complementary textual cues, yields the best performance. When λ is too small, the model overrelies on textual proxies, which are generally coarse and less adaptive to instance-specific features. The optimal setting of $\lambda = 0.8$ thus achieves a trade-off between the adaptability of visual proxy and the semantic stability of textual proxy.

Impact of Top-N in Eq. (5). We investigate how varying the retrieval parameter N in Eq. (5) influences the evolution of textual proxies in $\texttt{CoEvo}_{\mathrm{CSP}}$ on ImageNet-1K (Fig. 4). With a small N, both $\mathcal{N}_{\mathrm{near}}$ and $\mathcal{N}_{\mathrm{far}}$ exhibit limited diversity, resulting in a sparsely populated negative queue $\mathbf{T}_n$ and weakened OOD discrimination. Increasing N initially improves performance by injecting semantically richer negatives, thereby refining textual adaptation. However, beyond a task-dependent threshold, the gains saturate and may even decline due to two factors despite the deduplication constraint: (i) the marginal similarity gap between successive candidates diminishes, introducing redundancy rather than novel information; and (ii) a larger N increases the likelihood of enqueuing weakly aligned or noisy candidates. Empirically, we set $N = 5$, which offers a favorable trade-off between semantic coverage, prediction stability, and computational efficiency.

ID:OOD Ratio	1:100	1:10	1:1	10:1	100:1
NegLabel	23.00	20.50	21.55	25.92	19.69
CSP	20.00	18.00	16.78	19.50	17.95
AdaNeg	28.00	14.00	8.01	10.08	17.40
CoEvo$_{\mathrm{NegLabel}}$	17.00	**6.70**	**5.27**	**5.76**	**14.77**
CoEvo$_{\mathrm{CSP}}$	**14.00**	7.50	5.58	6.15	15.38

Table 4: FPR95 (↓) under different ID:OOD mixture ratios on ImageNet-1K (ID) and SUN (OOD).

Robustness to Data Imbalance. We further evaluate the robustness of our CoEvo under varying ID-OOD data imbalance scenarios. Five experimental settings are constructed using ImageNet-1K as ID and SUN as OOD data. **(i) Low-ID regimes:** For ratios of 1:100, 1:10, and 1:1, we randomly sample 10 K SUN images as OOD data, paired with 100, 1 K, and 10 K ImageNet samples, respectively. **(ii) High-ID regimes:** For ratios of 10:1 and 100:1, we fix 10 K ImageNet samples as ID data, combined with 1 K and 100 SUN images as OOD data. As shown in Tab. 4, our method consistently outperforms all baselines across all imbalance ratios. Notably, it maintains strong performance under extreme imbalance (e.g., 100:1 with only 100 OOD samples), demonstrating robustness to real-world data distribution shifts.

Conclusion

In this work, we introduced CoEvo, a test-time zero-shot OOD detection framework that enables bidirectional, sample-conditioned adaptation across visual and textual modalities. CoEvo maintains modality-specific proxy caches and iteratively refines them through a proxy-aligned co-evolution mechanism, dynamically realigning cross-modal similarity under distribution shifts without updating the backbone parameters. Furthermore, a multi-modal score evolution strategy fuses dual-modal evidence to produce calibrated OOD scores. Extensive experiments on standard benchmarks validate the effectiveness of our approach, demonstrating consistent improvements over negative-label baselines. Beyond the proposed framework, this work highlights the importance of dynamic, cross-modal adaptation for robust open-world recognition, paving the way for future research on scalable, label-free OOD detection.

Acknowledgments

This work was supported by the Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project) under the Shenzhen Municipal Science and Technology Innovation Commission (Project No. SGDX20230821092359002) and a grant under the Collaborative Research with World-leading Research Groups scheme of The Hong Kong Polytechnic University (project no. G-SACF). This research was also supported by the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant No. 2023B1515020097), the National Research Foundation Singapore under its AI Singapore Programme (AISG Award Nos.: AISG3-GV-2023-011 and AISG4-TC-2025-018-SGKR), the Singapore Ministry of Education AcRF Tier 1 Grant (Grant No. MSS25C004), and the Lee Kong Chian Fellowships.

References

Bitterwolf, J.; Müller, M.; and Hein, M. 2023. In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In *ICML*, volume 202, 2471–2506.

Cao, C.; Zhong, Z.; Zhou, Z.; Liu, T.; Liu, Y.; Zhang, K.; and Han, B. 2025. Noisy Test-Time Adaptation in Vision-Language Models. In *ICLR*.

Chen, M.; Gao, J.; and Xu, C. 2024. Conjugated Semantic Pool Improves OOD Detection with Pre-trained Vision-Language Models. In *NeurIPS*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing Textures in the Wild. In *CVPR*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Du, X.; Gozum, G.; Ming, Y.; and Li, Y. 2022. SIREN: Shaping Representations for Detecting Out-of-Distribution Objects. In *NeurIPS*.

Esmaeilpour, S.; Liu, B.; Robertson, E.; and Shu, L. 2022. Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. In *AAAI*, 6568–6576.

Fellbaum, C. 1998. *WordNet: An electronic lexical database*. MIT press.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*.

Horn, G. V.; Aodha, O. M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. J. 2018. The INaturalist Species Classification and Detection Dataset. In *CVPR*, 8769–8778.

Huang, R.; Geng, A.; and Li, Y. 2021. On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In *NeurIPS*, 677–689.

Huang, R.; and Li, Y. 2021. MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In *CVPR*, 8710–8719.

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. In *ICLR*.

Kong, S.; and Ramanan, D. 2021. OpenGAN: Open-Set Recognition via Open Data Generation. In *ICCV*, 793–802.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*, 7167–7177.

Li, T.; Pang, G.; Bai, X.; Miao, W.; and Zheng, J. 2024. Learning Transferable Negative Prompts for Out-of-Distribution Detection. In *CVPR*, 17584–17594.

Li, Z.; Tang, H.; Peng, Z.; Qi, G.; and Tang, J. 2025. Knowledge-Guided Semantic Transfer Network for Few-Shot Image Recognition. *IEEE Trans. Neural Networks Learn. Syst.*, 36(11): 19474–19488.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR*.

Liu, W.; Wang, X.; Owens, J. D.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *NeurIPS*.

Ming, Y.; Cai, Z.; Gu, J.; Sun, Y.; Li, W.; and Li, Y. 2022. Delving into Out-of-Distribution Detection with Vision-Language Representations. In *NeurIPS*.

Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to Exploit Hyperspherical Embeddings for Out-of-Distribution Detection? In *ICLR*.

Miyai, A.; Yu, Q.; Irie, G.; and Aizawa, K. 2023. LoCoOp: Few-Shot Out-of-Distribution Detection via Prompt Learning. In *NeurIPS*.

Nguyen, A. M.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 427–436.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, 8748–8763.

Ryu, S.; Koo, S.; Yu, H.; and Lee, G. G. 2018. Out-of-domain Detection based on Generative Adversarial Network. In *EMNLP*, 714–718.

Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; and Boult, T. E. 2013. Toward Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7): 1757–1772.

Sun, Y.; Guo, C.; and Li, Y. 2021. ReAct: Out-of-distribution Detection With Rectified Activations. In *NeurIPS*, 144–157.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-Distribution Detection with Deep Nearest Neighbors. In *ICML*, volume 162, 20827–20840.

Tack, J.; Mo, S.; Jeong, J.; and Shin, J. 2020. CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. In *NeurIPS*.

Tang, H.; He, S.; and Qin, J. 2025. Connecting Giants: Synergistic Knowledge Transfer of Large Multimodal Models for Few-Shot Learning. In *IJCAI*, 6227–6235.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. BlockMix: Meta Regularization and Self-Calibrated Inference for Metric-Based Meta-Learning. In *ACM Multimedia*, 610–618.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2025. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3): 1958–1974.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3Net: Multi-view Encoding, Matching, and Fusion for Few-shot Fine-grained Action Recognition. In *ACM Multimedia*, 1719–1728.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognit.*, 130: 108792.

Tao, L.; Du, X.; Zhu, J.; and Li, Y. 2023. Non-parametric outlier synthesis. In *ICLR*.

van Amersfoort, J.; Smith, L.; Teh, Y. W.; and Gal, Y. 2020. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In *ICML*, volume 119, 9690–9700.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2022. Open-Set Recognition: A Good Closed-Set Classifier is All You Need. In *ICLR*.

Wang, H.; Li, Y.; Yao, H.; and Li, X. 2023. CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No. In *ICCV*, 1802–1812.

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022. ViM: Out-Of-Distribution with Virtual-logit Matching. In *CVPR*, 4911–4920.

Wang, Y.; Li, B.; Che, T.; Zhou, K.; Liu, Z.; and Li, D. 2021. Energy-Based Open-World Uncertainty Modeling for Confidence Calibration. In *ICCV*, 9282–9291.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *ACM Multimedia*, 8962–8971.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Trans. Dependable Secur. Comput.*, 21(6): 5293–5307.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 3485–3492.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Hendrycks, D.; Li, Y.; and Liu, Z. 2022. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection.

Zhang, J.; Yang, J.; Wang, P.; Wang, H.; Lin, Y.; Zhang, H.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Li, Y.; Liu, Z.; Chen, Y.; and Li, H. 2023. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *CoRR*, abs/2306.09301.

Zhang, Y.; and Zhang, L. 2024. AdaNeg: Adaptive Negative Proxy Guided OOD Detection with Vision-Language Models. In *NeurIPS*.

Zhang, Y.; Zhu, W.; He, C.; and Zhang, L. 2024. LAPT: Label-Driven Automated Prompt Tuning for OOD Detection with Vision-Language Models. In *ECCV*, volume 15130, 271–288.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6): 1452–1464.