Generalized Out-of-Distribution Detection and Beyond in Vision Language Model Era: A Survey

Anonymous authors
Paper under double-blind review

Abstract

Detecting out-of-distribution (OOD) samples is crucial for ensuring the safety of machine learning systems and has shaped the field of OOD detection. Meanwhile, several other problems are closely related to OOD detection, including anomaly detection (AD), novelty detection (ND), open set recognition (OSR), and outlier detection (OD). To unify these problems, a generalized OOD detection framework was proposed, taxonomically categorizing these five problems. However, Vision Language Models (VLMs) such as CLIP have significantly changed the paradigm and blurred the boundaries between these fields, again confusing researchers. In this survey, we first present a generalized OOD detection v2, encapsulating the evolution of these fields in the VLM era. Our framework reveals that, with some field inactivity and integration, the demanding challenges have become OOD detection and AD. Then, we highlight the significant shift in the definition, problem settings, and benchmarks; we thus feature a comprehensive review of the methodology for OOD detection and related tasks to clarify their relationship to OOD detection. Finally, we explore the advancements in the emerging Large Vision Language Model (LVLM) era, such as GPT-4V. We conclude with open challenges and future directions.

1 Introduction

A reliable visual recognition system should not only accurately predict known contexts, but also identify and reject unknown examples (Amodei et al., 2016; Mohseni et al., 2021; Hendrycks et al., 2021b; Hendrycks & Mazeika, 2022). In critical applications such as autonomous driving, the system must alert and cede control to the driver upon encountering unfamiliar scenes or objects not seen during training. However, most existing machine learning models are trained based on the closed-world assumption (Krizhevsky et al., 2012; He et al., 2015), where the test data is assumed to be drawn *i.i.d.* from the same distribution as the training data, known as in-distribution (ID). Therefore, the development of classifiers capable of detecting out-of-distribution (OOD) samples is a crucial challenge for real-world applications. This challenge is precisely the focus of research in the field of OOD detection.

While OOD detection primarily focuses on semantic distribution shift, several other tasks share similar goals and motivations, including outlier detection (OD) (Aggarwal & Yu, 2001; Hodge & Austin, 2004; Ben-Gal, 2005; Wang et al., 2019a), anomaly detection (AD) (Ruff et al., 2021; Pang et al., 2021; Bulusu et al., 2020; Chalapathy & Chawla, 2019), novelty detection (ND) (Pimentel et al., 2014; Miljković, 2010; Markou & Singh, 2003a;b), and open set recognition (OSR) (Boult et al., 2019; Geng et al., 2020; Mahdavi & Carvalho, 2021). Subtle differences in the specific definitions among these sub-topics have caused confusion in the field, leading to similar approaches being proposed across them.

To address this issue, the generalized OOD detection framework was introduced (Yang et al., 2024). The taxonomy of the generalized OOD detection framework is shown in Fig. 1. The generalized OOD detection framework introduces a taxonomy built upon four criteria: distribution shift type (covariate/semantic shift), the type of ID data (single/multi-class), necessity of ID classification, and learning setting (transductive/inductive). According to the above taxonomy, these five problems can be clearly categorized as shown in Fig. 1: AD is categorized into sensory AD, which deals with covariate shift, and semantic AD, which deals

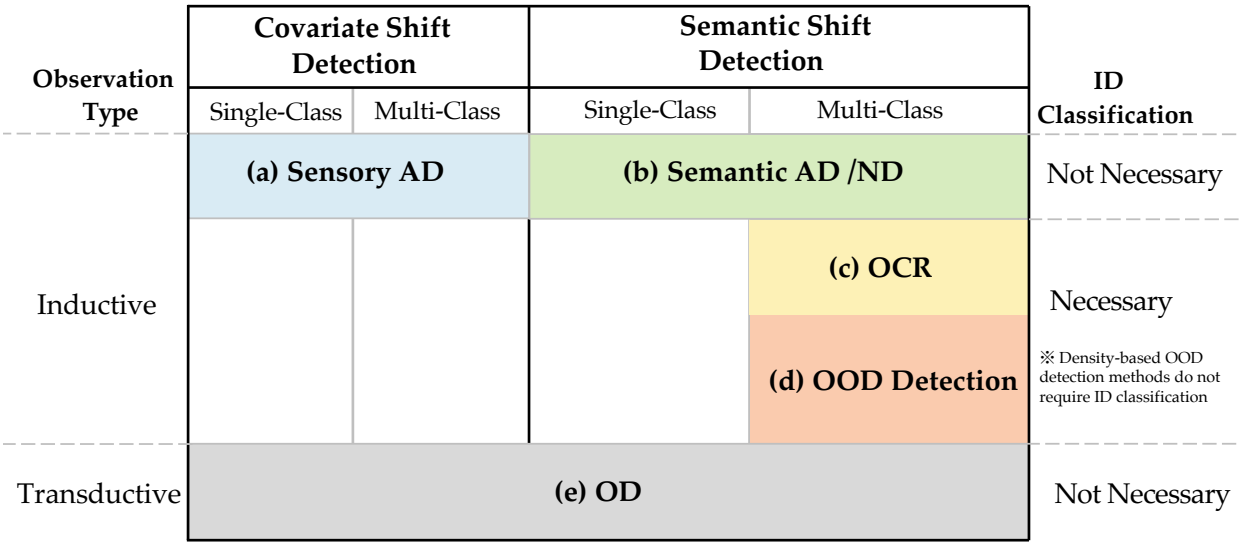

Figure 1: Taxonomy of generalized OOD detection framework (Yang et al., 2024), illustrated by classification tasks. (a) Sensory anomaly detection is categorized as covariate shift detection. (b) Semantic anomaly detection and novelty detection fall under semantic shift detection. (c) Open set recognition and (d) out-of-distribution detection are classified as multi-class semantic shift detection tasks that require ID classification. (e) Outlier detection is characterized by having a transductive observation type.

with semantic shift. ND falls under the same category as semantic AD. In multi-class settings requiring ID classification, both OSR and OOD detection are included. OD belongs to a transductive category (*i.e.*, it has access to all observations). This framework provides clear definitions and fosters a deeper understanding of each field.

In recent years, the emergence of Vision Language Models (VLMs), represented by CLIP (Radford et al., 2021), has rapidly accelerated research in the field of computer vision. This has changed the paradigm of the recognition field, allowing for zero-shot (Radford et al., 2021) or few-shot learning (Zhou et al., 2022c;b) in various domains. VLMs have significantly influenced the aforementioned five problems (OD, AD, ND, OSR, and OOD detection), and the application of VLMs, particularly CLIP, has become a highly notable research field (Ming et al., 2022a; Jeong et al., 2023; Miyai et al., 2023b; Zhou et al., 2024a). However, alongside this remarkable progress, the paradigm shift with the advent of the VLMs has blurred the boundaries between the five problems. Due to the difficulty of a clear understanding of the distinctions and interrelations between these tasks, each community within the fields is facing significant challenges in identifying the optimal direction to pursue in this VLM era.

In this survey, we introduce a novel unified framework termed *generalized OOD detection v2*, which extends the previous generalized OOD detection framework and summarizes the evolution of these five problems in the VLM era. To create it, we systematically review the use of VLMs across these five problem areas, tracing their development from the start to the present, and summarize the evolutionary trajectory of each problem. Importantly, our framework reveals that a paradigm shift has caused some fields to become inactive or integrate with others, and the demanding challenges in the VLM era become AD and OOD detection, which is a remarkable finding for each community. In addition to the inter-field evolution, we elaborate on the important shifts in the definition of OOD detection as well as the problem settings and benchmarks, with the contrast of those for related tasks. Then, we conduct a thorough review of the methodology for OOD detection and related tasks in the VLM era, intending to clarify their similarities and differences and inspire future research in OOD detection.

Table 1: Number of VLM-based papers in the Top Venues from 2021 to April 2025.

Task	Top Venue
(a) Sensory AD	CVPR2023×1 (Jeong et al., 2023), ICLR2024×1 (Zhou et al., 2024a), CVPR2024×4 (Li et al., 2024c; Ho et al., 2024; Zhu & Pang, 2024; Huang et al., 2024), ACMMM×2 (Gu et al., 2024a; Zhu et al., 2024a), IJCAI2024×1 (Zuo et al., 2024), ECCV2024×2 (Cao et al., 2024c; Qu et al., 2024), NeurIPS2024×3 (Li et al., 2024f; Arodi et al., 2024), ICLR2025×2 (Jiang et al., 2025; Yun et al., 2025), CVPR2025×6 (Sun et al., 2025; Ma et al., 2025; Qu et al., 2025; Zhang et al., 2025b; Fan et al., 2025; Gu et al., 2025)
(b) Semantic AD/ND	TMLR2022×1 (Liznerski et al., 2022), CVPR2024×1 (Zhu & Pang, 2024), ICLR2025×1 (Yun et al., 2025)
(c) OSR	ECCV2024×1 (Miller et al., 2024)
(d) OOD Detection	NeurIPS2021×1 (Fort et al., 2021), AAAI2022×1 (Esmaeilpour et al., 2022), NeurIPS2022×1 (Ming et al., 2022a), ICCV2023×1 (Wang et al., 2023a), IJCV2023×1 (Ming & Li, 2024a), NeurIPS2023×4 (Miyai et al., 2023b; Tu et al., 2023; Park et al., 2023; Liu et al., 2023b), ICLR2024×2 (Nie et al., 2023; Jiang et al., 2024), CVPR2024×2 (Bai et al., 2024a; Li et al., 2024b), ICML2024×1 (Cao et al., 2024a), ECCV2024×4 (Zhang et al., 2024d; Liu & Christopher, 2024; Zhang et al., 2024f; Lafon et al., 2024), NeurIPS2024×5 (Li et al., 2024e; Yu et al., 2024; Chen et al., 2024a; Zhang et al., 2024a; Zhang & Zhang, 2024), TMLR2024×1 (Adaloglou et al., 2024), IJCV2025×1 (Miyai et al., 2025), ICLR2025×1 (Zeng et al., 2025)
(e) OD	ICML2024×1 (Liang et al., 2024)

Finally, we introduce the early evolution of these problems in the emerging Large Vision Language Model (LVLM) era, such as GPT-4V (OpenAI et al., 2023) or LLaVA (Liu et al., 2023a; 2024b). We summarize the definition of each evolving problem, the findings so far, and future challenges.

To summarize, this survey presents four contributions to the research field:

1. **Proposing a Generalized OOD Detection v2**: We analyze the progression of five related topics (AD, ND, OSR, OOD detection, and OD) in the VLM era and propose *generalized OOD detection v2*. Our framework reveals that the paradigm shift has led to some field inactivity or integration, and the demanding challenges are AD and OOD detection. We hope that these observations highlight the demanding challenges in the VLM era and foster collaborative efforts among each community.

2. **An Extensive Survey for VLM-based OOD Detection**: We provide a comprehensive survey of VLM-based OOD detection and AD methods. In particular, we examine recent advances in zero-shot and few-shot settings by categorizing methods according to their training strategies and the use of additional prompts. Although various surveys on OD, AD, ND, OSR, and OOD detection have been conducted (Ruff et al., 2021; Pang et al., 2021; Bulusu et al., 2020; Chalapathy & Chawla, 2019; Geng et al., 2020; Yang et al., 2024; Cao et al., 2024b; Liu et al., 2024c; Xu & Ding, 2025), this work is the first to comprehensively review VLM-based OOD detection methods. By examining the connections among related tasks, we aim to provide insights that enhance the understanding of OOD detection.

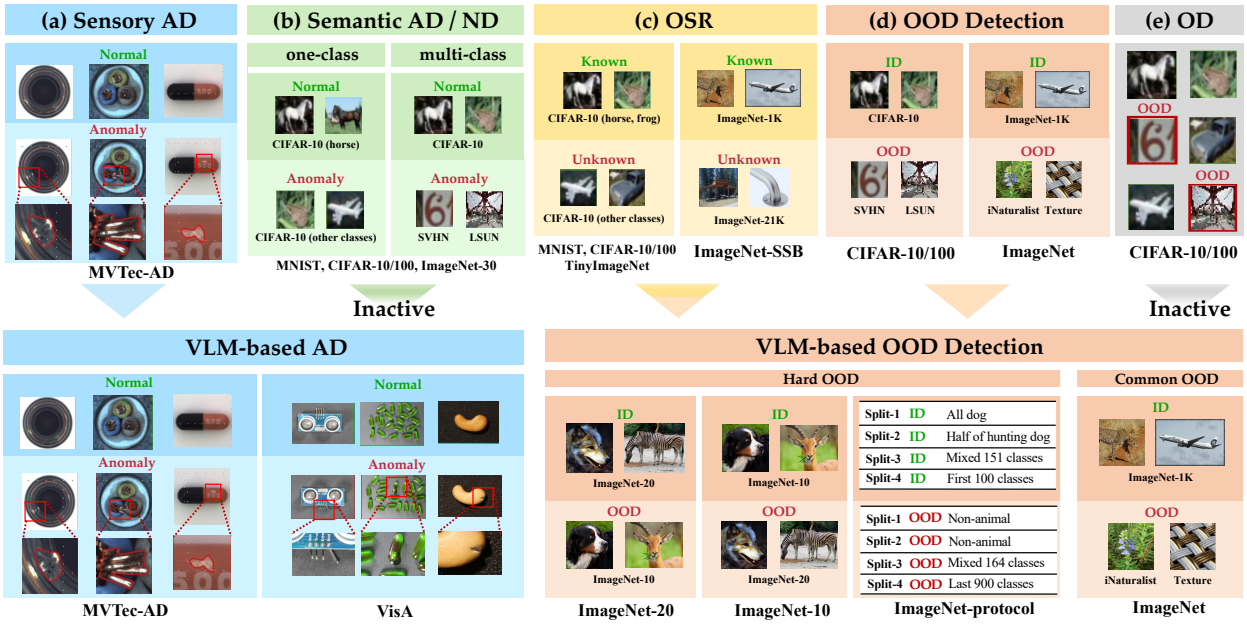

Figure 2: Generalized OOD detection framework v2, reflecting the evolution of each problem in the VLM era. (a) Sensory AD has become a highly active and noteworthy field in the VLM era. In terms of benchmarks, in addition to the commonly used MVTec-AD (Bergmann et al., 2019), VisA (Zou et al., 2022), the largest industrial anomaly detection dataset, has also become a standard benchmark in the field. (b) Semantic AD/ND has become inactive in the VLM era. (c) OSR has been integrated into hard OOD detection. VLM-based hard OOD detection incorporates the benchmark setup of OSR and creates new benchmarks such as ImageNet-10/ImageNet-20 (Ming et al., 2022a) and ImageNet-protocol (Palechor et al., 2023; Li et al., 2024b). (d) OOD detection is a highly active research area in the VLM era. (e) OD has become inactive in the VLM era.

3. **An Introduction to the Evolution in the LVLM Era**: We further introduce the evolution of each problem in the LVLM era. Despite the infant stage of these fields, this survey offers an in-depth introduction to each problem, aiming to facilitate future advancements in this area.

4. **Open Challenges and Future Directions**: Finally, we discuss open challenges and future research directions. In particular, by conducting a comparative analysis between the fields of AD and OOD, we identify key areas that are especially important for advancing VLM-based OOD detection. We hope that this survey will serve as a valuable reference for future research on OOD detection and related tasks in the VLM era.

In the VLM era, research on open-vocabulary segmentation (Gu et al., 2022) and referring segmentation (Wang et al., 2022b) has been related to the discovery of unseen classes. However, these fields mainly focus on tasks that aim to generalize to examples from unseen classes during training when the class name is already known. In contrast, OOD detection and related fields deal with entirely new data, where even the class name is unknown. The goal of OOD detection and related fields is to detect them instead of generalizing to them. Therefore, due to this fundamental difference in motivation and objective, we put the tasks such as open-vocabulary segmentation outside the scope of this survey.

The paper content is organized as follows. In Sec. 2, we introduce the new version of generalized OOD detection by summarizing the evolution of the five related fields in the VLM era. We then overview the two key problems (OOD detection and AD) that have evolved and remain active in Sec. 3, with a detailed breakdown of existing methodologies being presented in Sec. 4 (VLM-based OOD detection) and Sec. 5

(VLM-based AD). In Sec. 7, we introduce early advancements of OOD detection and AD in the LVLM era. Sec. 8 features future directions. Finally, we conclude with Sec. 9.

2 Generalized OOD Detection V2

In this section, we introduce a novel unified framework termed *generalized OOD detection v2*, which summarizes the evolution of the five related fields in the VLM era. We first revisit the previous generalized OOD detection framework in Sec. 2.1. Next, in Sec. 2.2, we show the results of the investigation of the research activity in each field. In Sec. 2.3, we introduce the evolution of each problem. Finally, in Sec. 2.4, we have a discussion about the future directions.

2.1 Background: Generalized OOD Detection V1

We first briefly revisit a previous *generalized OOD detection*, which encapsulates five related sub-topics: anomaly detection (AD), novelty detection (ND), open set recognition (OSR), out-of-distribution (OOD) detection, and outlier detection (OD). These sub-topics can be similar in the sense that they all define a certain *in-distribution*, with the common goal of detecting *out-of-distribution* samples under the open-world assumption. Previously, subtle differences existed among the sub-topics in terms of the specific definition and properties of in-distribution (ID) and OOD data.

To provide a clear definition, a generalized OOD detection framework was proposed (Yang et al., 2024). The taxonomy for generalized OOD detection is shown in Fig. 1. It is based on the following four bases: (1) Distribution shift to detect: The task focuses on detecting either covariate shift (*e.g.*, OOD samples from a different domain) or semantic shift (*e.g.*, OOD samples from a different semantic). (2) ID data type: The in-distribution (ID) data contains either a single class or multiple classes. (3) Whether the task requires ID classification: Some tasks require classification of the ID data, while others do not. (4) Transductive vs. inductive learning: Transductive tasks require all observations (both ID and OOD), while inductive tasks follow the common train-test scheme. According to the above taxonomy, these five problems can be clearly categorized as shown in Fig. 1: Anomaly detection is categorized into sensory anomaly detection, which deals with covariate shift, and semantic anomaly detection, which deals with semantic shift. Novelty detection falls under the same category as semantic anomaly detection. When addressing a multi-class scenario that necessitates ID classification, both open-set recognition and out-of-distribution detection are encompassed within this category. The main difference between OSR and OOD detection was the benchmark setup (Yang et al., 2024; Salehi et al., 2022) (Sec. 2.3 (c)). Outlier detection belongs to a different category from the other tasks, as this problem is transductive (*i.e.*, it has access to all observations).

The more precise definitions of these tasks are as follows:

Anomaly Detection Anomaly detection (AD) focuses on identifying anomalous instances during inference that deviate from a predefined notion of normality (Chandola et al., 2009). These anomalies can arise from two distinct types of distributional changes: covariate shift and semantic shift (Ruff et al., 2021). Accordingly, AD can be categorized into two primary sub-problems: sensory AD and semantic AD. Sensory AD targets anomalies caused by covariate shift, under the assumption that all normal instances are drawn from the same distribution. In this setting, no semantic shift is considered. In contrast, semantic AD addresses label shift, where normalities are assumed to come from a fixed semantic category. The objective in this setting is to detect test samples that belong to novel, previously unseen classes.

Novelty Detection Novelty detection (ND) aims to identify test samples that do not belong to any of the categories seen during training. Depending on the number of training classes, ND can be divided into two settings: One-class novelty detection, where the training set contains only one class, and multi-class novelty detection, where multiple classes are present in the training data. It is important to note that, even when multiple ID classes are available, the goal of multi-class novelty detection is solely to distinguish novel samples from known ones, without requiring classification among the ID classes. Both one-class and multi-class ND are binary classification tasks.

Open Set Recognition Open set recognition (OSR) requires a multi-class classifier to perform two tasks at the same time: (1) correctly classify test samples that belong to known classes, and (2) detect test samples that come from unknown classes. Here, it is common practice to refer to OOD classes as unknown classes and ID classes as known classes. However, there is no difference in meaning between these terms.

Out-of-Distribution Detection Out-of-distribution (OOD) detection aims to identify test samples that come from a distribution different from the training distribution. In most machine learning tasks, it typically refers to the label distribution, meaning that OOD samples should have labels not seen in the training data. It is also important to note that the training set usually includes multiple classes, and OOD detection should not degrade the model's ability to classify ID samples correctly.

Outlier Detection Outlier detection (OD) aims to identify samples that significantly differ from the rest of the given observation set. These differences may arise from either covariate shift or semantic shift. OD is often used as a pre-processing step for downstream tasks such as learning from open-set noisy labels (Wang et al., 2018) and open-set semi-supervised learning (Yu et al., 2020).

2.2 Investigation of Research Activity in Each Field

To investigate the research activity in each field, we comprehensively investigated papers that utilize VLMs from top venues and summarized them in Table 1. We have listed the papers in top venues by 2024.

Our survey reveals that CLIP (Radford et al., 2021) is by far the most widely used VLM for OOD detection, whereas other models such as Grounding DINO (Liu et al., 2024e) and SAM (Kirillov et al., 2023) have seen limited adoption in this context. While many existing methods are developed with CLIP, the concept of OOD detection with CLIP is not limited to this particular model. Therefore, throughout this survey, we use the term VLM-based OOD detection to refer broadly to OOD detection methods that leverage VLMs, and we similarly use the prefix VLM-based for other tasks (*e.g.*, VLM-based AD). However, it is important to note that not all VLMs employ softmax-based scoring mechanisms—for example, SigLIP (Zhai et al., 2023) uses sigmoid-based classification. As a result, some methods (Ming et al., 2022a; Miyai et al., 2025), which rely on softmax outputs, cannot be directly applied to such models. For clarity and consistency, this survey focuses primarily on CLIP as the representative VLM and investigates VLM-based methods for OOD detection and related tasks within this scope.

As OOD detection research is primarily focused on the image domain, we conduct a survey of other tasks within the image domain that are common and have strong connections to OOD detection research. For instance, our survey does not cover video domain tasks (Du et al., 2024; Zara et al., 2023; Wu et al., 2024b) due to their limited connection to OOD detection. From the results in Table 1, it is clear that Sensory AD and OOD detection have a very high number of papers, while the other fields are not as active. In the next section, we will examine these findings in greater depth and discuss the evolution of each field.

2.3 Generalized OOD Detection v2: Evolution in VLM Era

We propose a generalized OOD detection v2, encapsulating the evolution of each field in the VLM era. The evolution trajectory of the Generalized OOD detection framework v2 is shown in Fig. 2. The evolution of each field is as follows:

(a) Sensory AD → VLM-based AD Sensory AD has continued to develop as a common problem setting for VLM-based AD, inheriting the challenges of traditional sensory AD (Jeong et al., 2023; Deng et al., 2023; Chen et al., 2023c;b; Tamura, 2023; Chen et al., 2023c; Zhou et al., 2024a; Gu et al., 2024a; Li et al., 2024c; Zhu & Pang, 2024). As shown in Table 1, the first appearance in a top venue was at CVPR 2023, and since then, numerous papers have been published in top venues. Therefore, it is evident that sensory AD has become a highly active and noteworthy field in the VLM era.

(b) Semantic AD/ND → Inactive Research on semantic AD/ND appears to become inactive in the VLM era. As shown in Table 1, there are three two papers, TMLR 2022 (Liznerski et al., 2022), CVPR 2024 (Zhu & Pang, 2024), and ICLR 2025 (Yun et al., 2025). However, the CVPR 2024 (Zhu & Pang, 2024) and ICLR 2025 (Yun et al., 2025) work aims to build a generalist anomaly detector that solves many AD

tasks, including sensory AD and semantic AD, and is not primarily focused on semantic AD. The reasons for the inactivity include saturation of performance for one-class semantic AD/ND, and incompatibility of CLIP and multi-class semantic AD/ND settings. As for one-class semantic AD/ND, TMLR (Liznerski et al., 2022) exists, but the performances with common CIFAR and ImageNet-30 datasets have already achieved around 99%. As for multi-class semantic AD/ND, a common approach is to treat ID classes as a single class, but treating ID classes as a single class is less compatible with CLIP's class-wise discriminative capability.

(c) OSR → VLM-based OOD Detection We consider that OSR has been integrated into VLM-based hard OOD detection. According to Table 1, there is only one top venue publication (Miller et al., 2024) on OSR research in the VLM era. Originally, the main difference between OSR and OOD detection was the benchmark setup (Yang et al., 2024; Salehi et al., 2022). OSR typically divides the classes in the one dataset into some known (ID) classes and unknown (OOD) classes, as seen in MNIST-4/6 (Deng, 2012) CIFAR-4/6 (Krizhevsky et al., 2009a), CIFAR-50/50 (Krizhevsky et al., 2009b), and TinyImageNet-20/180 (Torralba et al., 2008). However, in recent years, some works on VLM-based OOD detection incorporate the benchmark setup of OSR and create new benchmarks such as ImageNet-10/ImageNet-20 (Ming et al., 2022a) and ImageNet-protocol (Palechor et al., 2023; Li et al., 2024b) for hard OOD detection. The only work (Miller et al., 2024) discusses the OSR problem in the VLM era, but they do not mention the differences from VLM-based OOD detection, and the definitions of VLM-based OSR are identical to those of VLM-based OOD detection. Considering the number of papers on VLM-based OOD detection and its identical definitions with OSR, we consider that the boundary between OOD detection and OSR has effectively disappeared, and all research in the VLM era has been integrated into OOD detection.

Nevertheless, while pure OSR research is declining, some studies have used the term "open-set" in the context of domain generalization (Shu et al., 2023). These studies deviate from the original scope of OSR research and are rather closely aligned with the field of domain generalization (Zhou et al., 2022a). Therefore, within our generalized OOD detection v2, we do not classify these studies as falling under OSR research.

(d) OOD Detection → VLM-based OOD Detection OOD detection is a highly active research area in the VLM era. As shown in Table 1, there are many papers in top venues, indicating a high interest from the community. Additionally, as mentioned above, OSR has been integrated with OOD detection as a field of hard OOD detection (Ming et al., 2022a; Li et al., 2024b). Therefore, it is expected that OOD detection will continue to grow and develop further.

(e) OD → Inactive OD has become less active in the VLM era. Previously, OD was used for open-set semi-supervised learning (Yu et al., 2020; Saito et al., 2021; Cao et al., 2022), learning with open-set noisy labels (Wang et al., 2018), and novelty discovery (Han et al., 2019; Zhao & Han, 2021; Jia et al., 2021; Vaze et al., 2022b; Joseph et al., 2022). The reason for the inactivity is that the use of CLIP led to a reduction in training costs and only a small amount of data needs to be collected, eliminating the need for large amounts of unlabeled data and reducing the need to consider noisy data. However, recently, Liang et al. (2024) proposed Unsupervised Universal Fine-Tuning, a new problem setting for VLM-based OD in ICML2024. Unsupervised Universal Fine-Tuning assumes a more realistic problem setting for unsupervised tuning of the downstream task with CLIP where some OOD samples are included in the unlabeled samples. With this new problem setting, there is still a possibility that OD will become active in the future. However, as OD is not currently an active area, we exclude OD from the main discussion of this survey. Unsupervised Universal Fine-Tuning is deeply related to OOD detection and will be discussed in detail in Sec. 4.3.

2.4 Discussion

Through Sec. 2.3, we found that previously mixed fields have been correctly organized in the VLM era, and that the focus becomes OOD detection and sensory AD. These fields are still developing, with an increasing number of methodologies and benchmarks, and are expected to become more active in the future. Note here that this does not mean that other fields have come to an end. For example, one reason why one-class semantic AD/ND has not been studied is the saturation of performance (Liznerski et al., 2022). If more fine-grained and challenging datasets could be constructed, the field could be reactive. We put this in out-of-scope for this survey paper, but this is an important future challenge.

3 Overview of Each Problem in VLM Era

In addition to the above inter-field evolution, we emphasize that the advent of VLMs has significantly changed the field of OOD detection itself. In this section, we present an overview of VLM-based OOD detection, highlighting the key changes in the problem definition, the problem setting, and benchmarks. In addition, we also present an overview of VLM-based AD in the hope that the understanding of each field will lead to a deeper understanding of VLM-based OOD detection. We describe the changes in problem definition, problem setting, and benchmarks in the VLM era. As for the background, applications, and evaluation in each field, we refer the readers to the original generalized OOD detection paper (Yang et al., 2024).

3.1 VLM-based Out-of-Distribution Detection

Definition The definition of VLM-based OOD detection differs significantly from that of conventional OOD detection. Conventional OOD detection aims to detect test samples drawn from a distribution that is different from the training distribution. As another definition, OOD detection is defined as a task to detect test samples that the model cannot or does not want to generalize (Yang et al., 2024). However, for VLM-based OOD detection, CLIP has a vast amount of knowledge, so the OOD sample is completely unrelated to the distribution of the CLIP's pretraining data or the CLIP's own generalization ability. Therefore, traditional definitions cannot adequately describe the definition of VLM-based OOD detection.

Unlike the previous definition, VLM-based OOD detection is defined as follows (Ming et al., 2022a; Esmaeilpour et al., 2022): VLM-based OOD detection aims to detect samples that do not belong to any ID class text provided by the user. Given a pre-trained model, a classification task of interest is defined by a set of class labels $\mathcal{Y}_{\text{ID}}$, which we refer to as the ID classes. The semantic distribution is represented by the distribution $P(\mathcal{Y}_{\text{ID}})$. VLM-based OOD detection aims to detect test samples that come from the distribution with the semantic shift from the ID classes, *i.e.*, $P(\mathcal{Y}_{\text{ID}}) \neq P(\mathcal{Y}_{\text{OOD}})$. Following the definition of the generalized OOD detection framework (Yang et al., 2024), ideal OOD detectors should keep the classification performance on test samples from ID class space $\mathcal{Y}_{\text{ID}}$, and reject OOD test samples with semantics outside the support of $\mathcal{Y}_{\text{ID}}$.

Problem Setting VLM-based OOD detection focuses on solving the image classification task in a computationally efficient way. Unlike traditional OOD detection settings, which primarily involve training an ID classifier with whole ID data, VLM-based OOD detection primarily focuses on a zero-shot (Ming et al., 2022a) (*i.e.*, without utilizing ID images) or few-shot (Miyai et al., 2023b) (*i.e.*, utilizing only a few ID images) setting. Each detailed definition of both settings is described later in Sec. 4. The field is advancing towards greater computational efficiency, requiring minimal or no training data.

Benchmark The benchmark has also changed between VLM-based methods and previous approaches. Earlier works before CLIP often utilized small-scale datasets such as CIFAR (Krizhevsky et al., 2009a;b) and MNIST (Deng, 2012). In contrast, most recent works in VLM-based OOD detection use high-resolution and large-scale datasets such as ImageNet (Ming et al., 2022a; Miyai et al., 2023b; Bai et al., 2024a; Li et al., 2024b; Cao et al., 2024a). The common ImageNet OOD benchmark uses ImageNet as ID and other datasets (Van Horn et al., 2018; Zhou et al., 2017; Xiao et al., 2010; Cimpoi et al., 2014) as OOD. However, in this common benchmark, the semantics between ID and OOD are far, which may allow easy distinction between the ID and OOD. Therefore, recent works use more challenging OOD benchmarks where they split ImageNet classes into ID and OOD categories for hard OOD detection (Ming et al., 2022a; Li et al., 2024b; Jung et al., 2024). The representative datasets are ImageNet-20 (Ming et al., 2022a), ImageNet-10 (Ming et al., 2022a), and ImageNet-protocol (Palechor et al., 2023) created by dividing into multiple variations of ID/OOD pairs from ImageNet-1K. This creation strategy initially focused on OSR but has recently been repurposed for OOD detection. More recently, Noda et al. (2025) proposed ImageNet-X for hard OOD detection and ImageNet-FS-X and Wilds-FS-X for hard full-spectrum OOD detection, which include more challenging conditions reflecting real-world scenarios. These changes in the datasets shift OOD detection closer to the real world and make it a more challenging and practical task.

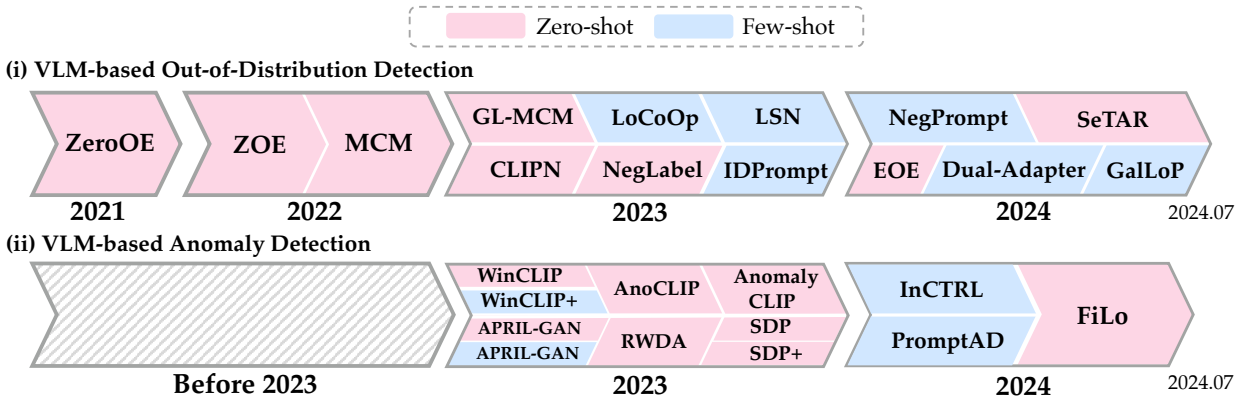

Figure 3: Timeline for representative methodologies for VLM-based out-of-distribution detection and VLM-based anomaly detection. We observe that an increasing number of methods have recently been proposed for both tasks, indicating the growing activity in these fields.

3.2 VLM-based Anomaly Detection

Definition Unlike OOD detection, the definition of anomaly detection (AD) has not changed between conventional AD and VLM-based AD. AD is intended for use in specific circumstances (industrial inspection), where samples that deviate from predefined normality are considered an anomaly (Yang et al., 2024; Ruff et al., 2021). Whether a model can generalize is irrelevant to the definition of "Anomaly". Therefore, even with the emergence of CLIP, the definition has not changed.

Problem Setting VLM-based AD focuses on solving anomaly classification and segmentation in a computationally efficient way. Anomaly classification is a binary classification task that distinguishes between normality and abnormality. Anomaly segmentation involves segmenting the location of anomalies. Like VLM-based OOD detection, VLM-based AD also primarily focuses on a zero-shot (Jeong et al., 2023) (*i.e.*, without utilizing images in the target dataset) or few-shot (Jeong et al., 2023) (*i.e.*, using only a few images in the target dataset) setting. Each detailed definition of zero-shot and few-shot settings is described later in Sec. 5. As another shift, conventional AD created separate models for each category (Bergmann et al., 2020; Defard et al., 2021; Li et al., 2021; Liznerski et al., 2021; Yi & Yoon, 2020; Zavrtanik et al., 2021; Thulasidasan et al., 2021), while VLM-based AD creates a single unified model across multiple categories (Jeong et al., 2023; Deng et al., 2023; Zhou et al., 2024a; Chen et al., 2023b; Zhu & Pang, 2024), which leads to a more computationally efficient approach.

One key difference from VLM-based OOD detection is that VLM-based OOD detection does not involve localization tasks, while these are mainstream in VLM-based AD. This will be discussed in detail in Sec. 5.4.

Benchmark Most works on VLM-based AD tackle industrial inspection (Bergmann et al., 2019; Chu & Kitani, 2020; Atha & Jahanshahi, 2018). As for the benchmarks, MVTec-AD dataset (Bergmann et al., 2019) and VisA dataset (Zou et al., 2022) are commonly used. MVTEC-AD includes 15 classes including 5 different texture categories and 10 different object categories with totally 5,354 high-resolution images. VisA covers 12 objects such as in 3 domain consisting of 10,821 high-resolution color images. The VisA benchmark includes objects with complex structures such as printed circuit boards and multiple instances with different locations within a single view, making it one of the most challenging datasets currently available in the open datasets. As for other datasets, KSDD (Tabernik et al., 2020), BTAD (Mishra et al., 2021), and MPDD (Jezek et al., 2021) are often used. However, apart from MVTec-AD and VisA, the datasets used vary depending on the papers (Yang et al., 2025; Cao et al., 2024c; Qu et al., 2024; Li et al., 2024f).

Table 2: Representative paper list for VLM-based out-of-distribution detection and anomaly detection.

Task	ID Image Availability	Training Type	OOD Prompts	Methods
§ 4 VLM-based OOD Detection	§ 4.1 Zero-shot	§ 4.1.1 Training-free	✓	ZeroOE (Fort et al., 2021), ZOC (Esmaeilpour et al., 2022), NegLabel (Jiang et al., 2024), EOE (Cao et al., 2024a), AdaNeg (Zhang & Zhang, 2024), CSP (Chen et al., 2024a)
			✗	MCM (Ming et al., 2022a), GL-MCM (Miyai et al., 2025), SeTAR (Li et al., 2024e), TAG (Liu & Christopher, 2024), CoVer (Zhang et al., 2024a)
		§ 4.1.2 Auxiliary Training	✓	CLIPN (Wang et al., 2023a)
	§ 4.2 Few-shot	§ 4.2.1 ID Training	✗	PEFT-MCM (Ming & Li, 2024a), LoCoOp (Miyai et al., 2023b), GalLoP (Lafon et al., 2024), SCT (Yu et al., 2024)
			✓	LSN (Nie et al., 2023), NegPrompt (Li et al., 2024b), ID-like-Prompt (Bai et al., 2024a), Local-Prompt (Zeng et al., 2025)
		§ 4.2.2 Training-free	✗	Dual-Adapter (Chen et al., 2024b), DPM (Zhang et al., 2024f)
§ 5 VLM-based Anomaly Detection	§ 5.1 Zero-shot	§ 5.1.1 Training-free	✓	WinCLIP (Jeong et al., 2023), AnoCLIP (Deng et al., 2023), SDP (Chen et al., 2023c), ALFA (Zhu et al., 2024a)
		§ 5.1.2 Auxiliary Training	✓	APRIL-GAN (zero-shot) (Chen et al., 2023b), RWDA (Tamura, 2023), SDP+ (Chen et al., 2023c), AnomalyCLIP (Zhou et al., 2024a), FiLo (Gu et al., 2024a), AdaCLIP (Cao et al., 2024c), VCP-CLIP (Qu et al., 2024), AA-CLIP (Ma et al., 2025), Bayes-PFL (Qu et al., 2025),
	§ 5.2 Few-shot	§ 5.2.1 Training-free	✓	WinCLIP+ (Jeong et al., 2023), UniVAD (Gu et al., 2025)
		§ 5.2.2 ID Training	✓	CLIP-FSAC (Zuo et al., 2024), PromptAD (one-class) (Li et al., 2024c), One-to-Normal (Li et al., 2024f)
		§ 5.2.3 Auxiliary Training + ref.	✓	APRIL-GAN (few-shot) (Chen et al., 2023b), InCTRL (Zhu & Pang, 2024)

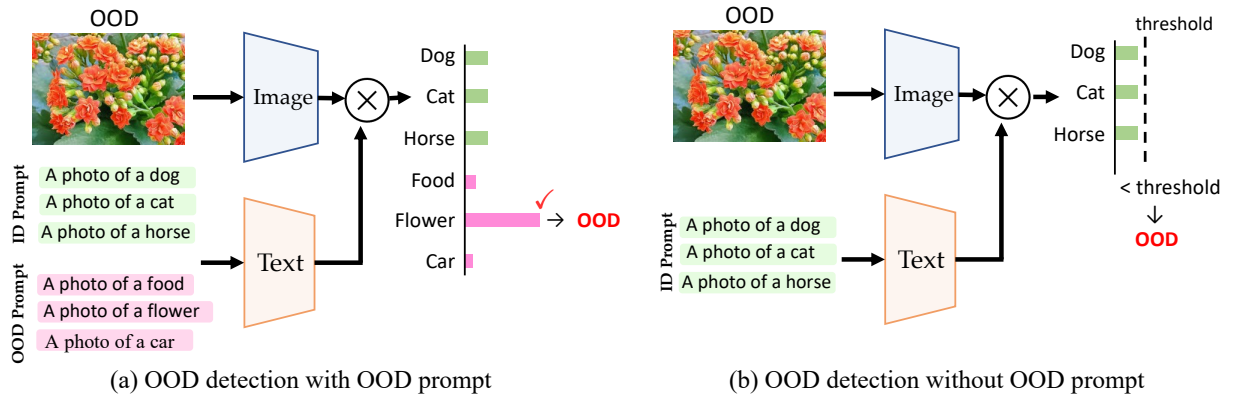

Figure 4: Illustration of the OOD detection methods with OOD prompt and without OOD prompt.

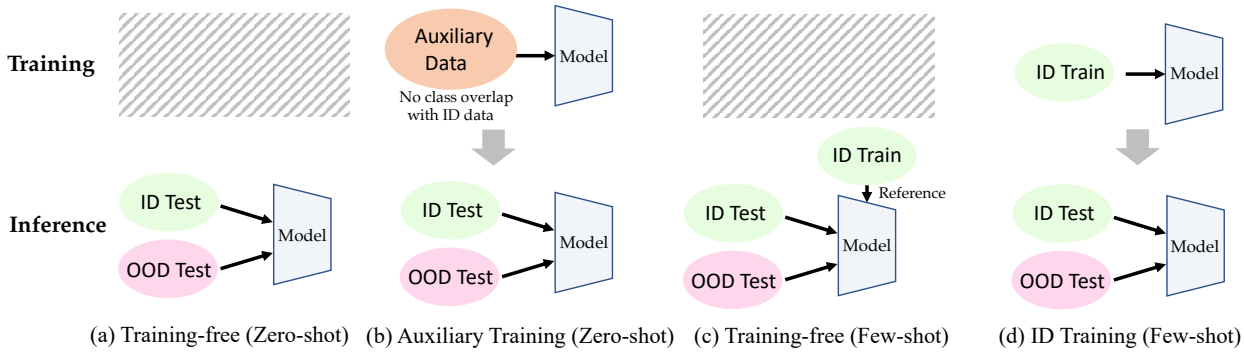

Figure 5: Illustration of the OOD detection methods for each training type in both zero-shot and few-shot settings.

4 VLM-based OOD Detection: Methodology

In this section, we introduce the methodologies for VLM-based out-of-distribution (OOD) detection. Fig. 3 presents the timeline for representative methodologies for VLM-based OOD detection. Table 2 presents representative methods. We introduce methods for zero-shot OOD detection in Sec. 4.1, few-shot OOD detection in Sec. 4.2, and other research directions in Sec. 4.3. We categorize each methodology by the training type and the use of OOD prompts. As for the OOD prompts, the methods that employ additional OOD prompts and do not use them are shown in Fig. 4 (a) and (b), respectively. As for the training type, we categorize the methods into training-free, ID training, and auxiliary training (training with other data than ID training data). The illustration of each training scenario is shown in Fig. 5.

4.1 Zero-shot Out-of-Distribution Detection

Zero-shot OOD detection was proposed in 2021 by Fort et al. (2021). Since then, a growing number of methods have been proposed year by year.

Definition of Zero-shot OOD Detection In zero-shot OOD detection, the term “Zero-shot" refers to the non-use of ID images during both training and inference phases. For instance, the method with additional training with auxiliary datasets (non-use of ID images) can be regarded as a zero-shot method (Wang et al., 2023a). The method with the pre-processing of the ID class texts can also be regarded as a zero-shot method (Fort et al., 2021; Esmaeilpour et al., 2022; Jiang et al., 2024; Cao et al., 2024a).

4.1.1 Training-free Methods

a. With OOD Prompts VLM-based OOD detection started in this setting. The earliest work is ZeroOE (Fort et al., 2021). ZeroOE feeds the potential OOD labels to the textual encoder of CLIP. However, the method of using known OOD labels is infeasible for real-world applications. To solve this problem, ZOC (Esmaeilpour et al., 2022) proposed to train an OOD label generator based on the visual encoder of CLIP and use the generated pseudo-OOD labels for OOD detection. However, when dealing with large-scale datasets encompassing a multitude of ID classes, the label generator may not generate effective candidate OOD labels, resulting in poor performance. Building on these early works (Fort et al., 2021; Esmaeilpour et al., 2022), recent works focus on how to obtain high-quality OOD labels through either (i) OOD label retrieval (Jiang et al., 2024; Ding & Pang, 2024) or (2) OOD label generation (Cao et al., 2024a). (i) A representative retrieval-based method is NegLabel (Jiang et al., 2024). NegLabel selects high-quality OOD labels from extensive corpus databases by calculating the distance between an extracted OOD label and ID label. AdaNeg (Zhang & Zhang, 2024) extends NegLabel by incorporating adaptive negative proxies, which are dynamically generated during inference by leveraging actual OOD images. This adaptation ensures a closer alignment with the test-time OOD label space, thereby enhancing the effectiveness of negative proxy guidance. Chen et al. (2024a) proposed Conjugated Semantic Pool (CSP), which enhances OOD detection by expanding the semantic space with modified superclass names. (ii) A representative generation-based method is EOE (Cao et al., 2024a). EOE utilizes Large Language Models (LLMs) to produce high-quality OOD labels. By modifying the prompts given to the LLM, EOE can be generalized to a variety of tasks, including far and near OOD detection.

b. Without OOD Prompts In zero-shot OOD detection, many methods utilize OOD labels, but the difficulty and cost of creating these labels pose challenges. To address these issues, Ming et al. (2022a) proposed MCM, which uses only ID labels to detect OOD. MCM is a simple approach that devises softmax scaling to align visual features with textual concepts for OOD detection. Despite its simplicity, MCM has high effectiveness and scalability, and it serves as a crucial baseline in VLM-based OOD detection. Building on the concept of MCM, Miyai et al. (2025) proposed GL-MCM, which extends MCM by just adding a local MCM score to enhance the fine-grained detection capability in local regions. Based on MCM and GL-MCM SeTAR (Li et al., 2024e) enhances them by changing the model's weight matrices using a simple greedy search algorithm. As for another direction, some works adopt data-centric approaches (Liu & Christopher, 2024; Zhang et al., 2024a). Liu & Christopher (2024) proposed TAG, which applies simple augmentations to the ID text prompts. Zhang et al. (2024a) proposed CoVer, which enhances separability by averaging confidence scores across corrupted inputs, leveraging the greater confidence drop in OOD samples under corruptions.

We consider these methods to be post-hoc methods for VLM-based OOD detection in that they directly employ an ID classifier for OOD detection. Due to their simplicity and high scalability, these post-hoc methods can bring fundamental performance improvements for many subsequent methods (Miyai et al., 2023b; Nie et al., 2023; Li et al., 2024b). Therefore, we expect that this field should be developed further in the future, reflecting the trajectory of the field before CLIP emerged (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Liu et al., 2020; Sastry & Oore, 2020; Wang et al., 2021; Zhang et al., 2023; Sun & Li, 2022; Sun et al., 2022; Lin et al., 2021; Sastry & Oore, 2019).

4.1.2 Auxiliary Training-based Methods

CLIPN (Wang et al., 2023a) is the only auxiliary training-based method for zero-shot OOD detection. CLIPN aims to empower the logic of saying "no" within CLIP, and it designs a novel learnable "no" prompt and an additional "no" text encoder to capture negation semantics within images. To create an additional text encoder, CLIPN needs to be pre-trained on the CC-3M dataset (Sharma et al., 2018). While the extensive pre-training of CLIPN may indeed lead to intensive computations and lower scalability, once pre-trained, it performs zero-shot OOD detection across a wide range of domains with comparable performance to few-shot OOD detection methods (Miyai et al., 2023b; Nie et al., 2023).

4.2 Few-shot Out-of-Distribution Detection

Few-shot OOD detection was concurrently proposed by Miyai et al. (2023b) and Ming & Li (2024a) in June 2023. Since then, it has become an active research area in VLM-based OOD detection.

Definition of Few-shot OOD Detection VLM-based few-shot OOD detection aims to detect OOD images using only a few labeled ID images. In few-shot OOD detection, the term "Few-shot" refers to the use of a few ID images during training or inference phases. For instance, the method with additional training with a few ID images can be regarded as a few-shot method (Li et al., 2024c). Even without training, if a method uses a few ID images as a reference, it is treated as a few-shot method (Jeong et al., 2023). Regarding the number of shots, it is common to experiment with 1-shot to 16-shot (Miyai et al., 2023b; Chen et al., 2024b), following the closed-set setting (Zhou et al., 2022c).

4.2.1 ID Training-based Methods

a. Without OOD Prompts Few-shot OOD detection began in this setting. Ming & Li (2024a) proposed PEFT-MCM for VLM-based OOD detection, which demonstrates the effectiveness of combining parameter-efficient tuning methods (*e.g.*, prompt learning (Zhou et al., 2022c) or adapter (Zhang et al., 2022)) and MCM (Ming et al., 2022a). Concurrently, Miyai et al. (2023b) proposed LoCoOp, a pioneer prompt learning approach for few-shot OOD detection. LoCoOp enhances CoOp's (Zhou et al., 2022c) OOD detection capabilities by performing OOD regularization with local OOD features. LoCoOp is the simplest prompt learning method and serves as a crucial baseline in few-shot OOD detection. To improve the performance of LoCoOp, Self-Calibrated Tuning (SCT) (Yu et al., 2024) dynamically adjusts the weight of OOD regularization based on prediction uncertainty to prevent unreliable OOD features from degrading the model's performance. Unlike LoCoOp, which utilizes non-ID local regions for OOD regularization, GalLoP (Lafon et al., 2024) proposes an approach that utilizes local ID regions to enable a more fine-grained distinction between ID and OOD samples. GalLoP learns a diverse set of prompts by utilizing both global and local visual representations, thereby enhancing detection capabilities.

b. With OOD Prompts Similar to zero-shot OOD detection, recent works in few-shot OOD detection utilize additional OOD prompts (Nie et al., 2023; Bai et al., 2024a; Li et al., 2024b). As representative methods, LSN (Nie et al., 2023) and NegPrompt (Li et al., 2024b) were proposed concurrently. They state that the simple negative prompts added "not" (*e.g.*, "not a photo of a [cls]") fail to capture the dissimilarity for identifying OOD samples. Therefore, by preparing negative prompts and training with them, LSN and NegPrompt can learn suitable negative prompts, enabling more accurate detection of OOD samples. The difference between LSN and NegPrompt lies in their approach to the use of negative prompts. LSN prepares unique negative prompts for each class and learns suitable negative prompts for each class. In contrast, NegPrompt prepares multiple negative prompts common to all ID classes and trains them to learn generic templates representing the negative semantics of any given class labels. Additionally, NegPrompt tested the performances in the hard OOD detection setting with ImageNet-protocol (Palechor et al., 2023), outperforming LoCoOp and CoOp. Alternatively, ID-like-Prompt (Bai et al., 2024a) takes a different approach by introducing ID-like prompts, which are designed for capturing OOD features that are close to the ID features. It extracts ID-like OOD regions from ID training images and trains ID-like prompts using these extracted OOD data, enabling the model to capture more subtle differences within images. In a unique direction, LAPT (Zhang et al., 2024d) proposes an automatic sample collection strategy that retrieves or generates training ID images only with ID class names, which achieves high performance without image collection and annotation costs. LAPT then performs distribution-aware prompt learning, which distinguishes between ID class and OOD class tokens. LAPT is positioned within the context of more efficient few-shot OOD detection in this survey paper since it requires generating or retrieving "ID images" for the data collection.

In the context of few-shot OOD detection, recently, Li et al. (2024b) proposed a new problem setting called open-vocabulary OOD (OV-OOD) detection. While common few-shot OOD detection involves training on images from all ID classes during training, OV-OOD detection involves training on images from just a small subset of ID classes and performing OOD detection using all ID classes at inference time. Formally, we define

a subset of semantic labels $\mathcal{Y}_{\text{ID,sub}} \subset \mathcal{Y}_{\text{ID}}$, where $\mathcal{Y}_{\text{ID}}$ represents all ID labels. Based on this subset of labels, we define a corresponding subset dataset $\mathcal{D}_{\text{ID,sub}}^{\texttt{train}} \subset \mathcal{D}_{\text{ID}}^{\texttt{train}}$. During training, only $\mathcal{D}_{\text{ID,sub}}^{\texttt{train}}$ is used. Then, at inference time, all ID classes $\mathcal{Y}_{\text{ID}}$ are used, and the goal is to detect OOD from a combination of all ID test data $\mathcal{D}_{\text{ID}}^{\texttt{test}}$ with $\mathcal{Y}_{\text{ID}}$ and OOD test data $\mathcal{D}_{\text{OOD}}^{\texttt{test}}$ with $\mathcal{Y}_{\text{OOD}}$. For this setting, existing few-shot OOD detection methods (Miyai et al., 2023b; Li et al., 2024b) can be easily applied by simply combining the rest of the ID classes. In particular, NegPrompt (Li et al., 2024b) learns general negative prompts that are not specific to the training ID classes, so it achieves high performance in OV-OOD detection.

4.2.2 Training-free Methods

Recently, training-free few-shot OOD detection methods, which use ID images solely as references during inference, have gained attention. Dual-Adapter (Chen et al., 2024b) adopts a prior-based method Tip-Adapter (Zhang et al., 2022), which leverages both textual and visual features with a cache model and enhances performance without training. To adapt this to few-shot OOD detection, Dual-Adapter employs the concept of dual cache modeling and constructs Positive-Adapter and Negative-Adapter, and identifies OOD samples with the prediction difference with both adapters. Dual-Pattern Matching (DPM) (Zhang et al., 2024f) enhances CLIP for OOD detection by utilizing both textual and visual ID patterns. DPM stores ID class-wise text features as the textual pattern and the aggregated ID visual information as the visual pattern. During inference, it evaluates the similarity of inputs to both patterns to identify OOD samples.

4.3 Other Directions

4.3.1 VLM-based Full-spectrum OOD Detection

VLM-based full-spectrum OOD (FS-OOD) detection is a crucial challenge (Lu et al., 2023). FS-OOD detection was proposed by Yang et al. (2023) as an important setting that considers both the detection of semantic shift and the ability to recognize covariate-shifted data (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a). Unlike standard OOD detection, which only focuses on semantic shifts between training and test distributions, FS-OOD detection further considers non-semantic covariate shift by including covariate-shifted ID images. As for the benchmarks, OpenOOD v1.5 (Zhang et al., 2024b) provides two large-scale benchmarks based on ImageNet-200 and ImageNet-1K, incorporating ImageNet-C (Hendrycks & Dietterich, 2019) with image corruptions, ImageNet-R (Hendrycks et al., 2021a) with style changes, and ImageNet-V2 (Recht et al., 2019) with resampling bias as ID. As for VLM-based methods, LSA (Lu et al., 2023) uses a bidirectional prompt customization mechanism, which adjusts discriminative ID and OOD boundary. The latest work, Noda et al. (2025) introduces ImageNet-FS-X and Wilds-FS-X as new benchmarks. ImageNet-FS-X utilizes the hierarchical structure of ImageNet labels to define a more challenging ID/OOD split while also evaluating robustness to covariate shifts in ImageNet variants. To further align these benchmarks with real-world testbeds, Wilds-FS-X extends Wilds (Koh et al., 2021) for FS-OOD evaluation. These benchmarks present substantial room for further improvements.

4.3.2 Other Tasks with VLM-based OOD Detection

Unsupervised Universal Fine-Tuning VLM-based OOD detection is useful for a new task called Unsupervised Universal Fine-Tuning (UUFT) (Liang et al., 2024). UUFT is a problem of unsupervised learning for outlier detection (OD). Existing studies for unsupervised learning assumed that all unlabeled images belong to one of the ID classes (Huang et al., 2022; Shu et al., 2022; Tanwisuth et al., 2023), but they require prior knowledge of exact class names linked to ground truth labels, which restricts their usefulness in various real-world situations. For a more realistic setting, UUFT assumes that OOD images are included in the unlabeled images. To detect OOD images during training, they developed MCM (Ming et al., 2022a) and proposed UEO, which leverages sample-level confidence to approximately minimize the conditional entropy of confident instances and maximize the marginal entropy of less confident instances.

Open-world Prompt Tuning VLM-based OOD detection is useful for a new task called Open-world Prompt Tuning (Zhou et al., 2024b). Open-world Prompt Tuning is a task that evaluates the classification accuracy on a mix of known and novel ID classes while training the model with known classes. To solve

this problem, Zhou et al. (2024b) proposed DeCoOp which incorporates OOD detection into the inference pipelines and improves the base-to-new separability, preventing performance degradation on new classes.

5 VLM-based AD: Methodology

In this section, we introduce methodologies for VLM-based anomaly detection (AD) in the hope that the contrast with OOD detection clarifies the similarities and differences between each task and facilitates a deeper understanding of VLM-based OOD detection.

Common Findings in All Settings In VLM-based AD, all representative methods utilize anomaly prompts (*e.g.*, "`anomalous [class]`") to detect anomalies. This hypothesis is supported by several observations from existing work (Jeong et al., 2023). Firstly, the concepts of normality and anomalies are context-dependent states (Isola et al., 2015) of an object, with language playing a crucial role in defining these states. Secondly, language provides additional insights that help differentiate defects from acceptable variations in normality.

5.1 Zero-shot Anomaly Detection

VLM-based zero-shot AD was proposed in 2023 by Jeong et al. (2023). Although it started about two years later than OOD detection, many methods have been proposed up to the present.

Definition of Zero-shot AD The meaning of the term "Zero-shot" for zero-shot AD is similar to that for zero-shot OOD detection. In zero-shot AD, the term "Zero-shot" refers to the non-use of the images in the target domain during both training and inference phases. For instance, the method with additional training with auxiliary datasets can be regarded as a zero-shot method (Chen et al., 2023b; Tamura, 2023; Chen et al., 2023c; Zhou et al., 2024a; Gu et al., 2024a). The method with the pre-processing of the target class texts can also be regarded as a zero-shot method (Jeong et al., 2023; Deng et al., 2023; Chen et al., 2023c).

5.1.1 Training-free Methods

The simplest zero-shot AD methods are (i) to perform anomaly classification with CLIP using text prompts for normality and anomalies as classes (*i.e.*, "`normal [class]`" *vs.* "`anomalous [class]`") and (ii) to calculate the similarity to the normal prompt (*i.e.*, "`normal [class]`") as the score. These methods are called CLIP-AC (Jeong et al., 2023). Jeong et al. (2023) reported that CLIP-AC with both normal and anomaly prompts outperforms that with only normal text prompts, which indicates the importance of the use of anomaly prompts. However, the performances for this naive method are not yet satisfactory due to the wide range of variations of anomalies. To solve this issue, Jeong et al. (2023) proposed WinCLIP. WinCLIP performs a compositional ensemble on a large number of pre-defined normal and anomaly templates and efficient extraction and aggregation of window/patch/image-level features aligned with the text. WinCLIP outperforms CLIP-AC by a large margin. Because of its simplicity and pioneering work, WinCLIP has become an important baseline for VLM-based AD. AnoCLIP (Deng et al., 2023) follows WinCLIP's approach of using a large number of pre-defined normal and anomaly templates but modifies the templates to be domain-aware (*e.g.*, industrial photo) and contrastive state for normal and anomaly (*e.g.*, perfect and imperfect). However, it is noteworthy that the performance of the ensemble strategies of previous methods heavily depends on the text descriptions (Jeong et al., 2023; Deng et al., 2023). Also, it is observed that more descriptions are not always better (Chen et al., 2023c), which makes the previous approaches (Jeong et al., 2023; Deng et al., 2023) using a naive ensemble of large templates somewhat uncontrollable and random in their applications. Therefore, SDP (Chen et al., 2023c) proposes RVS, a representative vector selection paradigm, which makes the mechanism of extracting representative vectors from large templates controllable, allowing for a more diverse selection of representative vectors. As a more recent work, ALFA is a zero-shot anomaly detection framework that leverages LLMs to generate adaptive anomaly descriptions at runtime and optimizes the image-text alignment of CLIP at both global and local levels, enhancing detection accuracy and interpretability (Zhu et al., 2024a).

5.1.2 Auxiliary Training-based Methods

Existing methods with auxiliary training perform training on the test set of one dataset and perform zero-shot testing on the other dataset (Chen et al., 2023b; Zhou et al., 2024a; Gu et al., 2024a) (*e.g.*, training with MVTec-AD and evaluation with VisA.) In recent years, the development of auxiliary training-based methods has received more attention than training-free zero-shot methods. There are two main reasons why training is necessary for AD: (i) The first is the domain gap between semantics and anomalies. CLIP is pre-trained to understand the semantics of images, so when applied in a zero-shot manner, it captures the semantics of the image. However, actual anomalies are not semantics, but rather represent the state of an object and appear only in local areas of the image. Therefore, without training, this domain gap between semantics and anomalies cannot be bridged. (ii) The second reason is that there are limitations to relying on a large set of manually crafted anomaly prompts. This incurs prompt creation costs and also makes it difficult to respond to unknown anomalies.

To address the above issue (i), APRIL-GAN (also known as VAND) (Chen et al., 2023b) was proposed. APRIL-GAN tackles the domain gap between semantics and anomaly by adding additional linear layers in vision encoders. These linear layers project image features at each scale into the text space, creating and aggregating anomaly maps at each stage. Similarly, SDP+ (Chen et al., 2024b) also incorporates additional linear layers into SDP (Chen et al., 2024b) to effectively project image features into the text feature space, addressing the misalignment between image and text. To solve both the issue (i) and (ii), AnomalyCLIP (Zhou et al., 2024a) was proposed. AnomalyCLIP is a textual prompt learning-based method similar to CoOp. By replacing anomaly prompts with learnable parameters, it eliminates the need to prepare a large number of manually pre-defined prompts such as those in WinCLIP (Jeong et al., 2023). Furthermore, unlike CoOp (Zhou et al., 2022c) which learns object semantics, AnomalyCLIP learns object-agnostic text prompts that capture generic normality and abnormality in an image regardless of its semantics. To achieve this, AnomalyCLIP introduces object-agnostic text prompt templates for both normal and anomaly and performs global and local context optimization. FiLo (Gu et al., 2024a) leverages Large Language Models (LLMs) to generate fine-grained anomaly descriptions for each object category. This method replaces generic abnormal descriptions with LLM-generated specific anomaly content for each sample. By adding learnable prompts before the generated anomaly prompts, FiLo performs global and local context optimization, enhancing the ability to detect anomalies. As a more recent method, AdaCLIP (Cao et al., 2024c) employs static and dynamic learnable prompts in both the image and text encoders. Static prompts provide a general adaptation for zero-shot detection, while dynamic prompts are generated per test image for adaptive refinement. Their combination, termed hybrid prompts, enhances zero-shot performance.

As a unique direction from these methods, RWDA (Tamura, 2023) proposes a data augmentation approach by utilizing CLIP's text embeddings as training data. RWDA adds randomly generated words into normal and anomaly prompts to generate a diverse set of normal and anomaly training samples and trains a regular feed-forward neural network with diverse text embeddings. VCP-CLIP (Qu et al., 2024) proposes to leverage visual context prompting (VCP) to enhance CLIP's ability to detect and segment anomalies without requiring product-specific text prompts. VCP eliminates reliance on predefined textual descriptions and improves generalization across unseen product categories.

5.2 Few-shot Anomaly Detection

VLM-based few-shot AD was proposed in 2023 by Jeong et al. (2023), concurrently with the development of zero-shot AD (Jeong et al., 2023).

Definition of Few-shot AD VLM-based few-shot AD aims to detect anomaly images using only a few images in the target domain. The meaning of the term "Few-shot" is similar to that of few-shot OOD detection. In few-shot AD, the term "Few-shot" refers to the use of a few images in the target domain during training or inference phases. For instance, the method with additional training with a few images in the target domain can be regarded as a few-shot method (Li et al., 2024c). Even without training, if a method uses a few images in the target domain as a reference, it is regarded as a few-shot method (Jeong et al., 2023).

5.2.1 Training-free Methods

The earliest approach in VLM-based few-shot AD is WinCLIP+ (Jeong et al., 2023), an improved method of WinCLIP. WinCLIP, a base zero-shot AD method, cannot identify certain defects that can only be defined visually rather than textually. For example, the "Metal-nut" category in MVTecAD has an anomaly type labeled "flipped upside down," which can only be identified relative to a normal image. To address this, WinCLIP+ incorporates a few normal reference images into a memory bank (Roth et al., 2022) and calculates the anomaly score with the cosine similarity between the query image and its most similar image in the memory bank.

5.2.2 ID Training-based Methods

Training-based methods for VLM-based AD are generally categorized as auxiliary training (in Sec. 5.1.2), but some approaches use data from the same domain as the test data for training. CLIP-FSAC (Zuo et al., 2024) enhances few-shot anomaly classification by introducing a two-stage fine-tuning framework with modality-specific adapters, where the image adapter is trained first using text features, followed by the text adapter while freezing the image adapter. To address data scarcity, it employs synthetic anomaly generation using random perturbation and Poisson-based editing to create diverse training samples. PromptAD (Li et al., 2024c) proposes a prompt learning method for one-class AD (where the normal class consists of one class). In one-class AD, traditional prompt learning methods for multi-class classification (*e.g.*, CoOp (Zhou et al., 2022c)) do not work well. To address this, PromptAD creates a large number of anomaly prompts by adding a learnable anomaly suffix to the normal prompt. It then learns to bring the visual features closer to the normal prompt and further away from the anomaly prompts. As a different direction, Kwak et al. (2024) proposed ADP, which learns anomaly-aware prompts using the personalization capabilities of diffusion models. Also, Li et al. (2024f) has proposed a novel anomaly personalization approach, which applies a one-to-normal transformation to query images using a customized anomaly-free generation model, ensuring alignment with the normal data manifold. This method leverages a triplet contrastive anomaly inference strategy, incorporating comparisons between the query image, a generated anomaly-free data pool, and textual prompts to enhance prediction stability and robustness.

5.2.3 Auxiliary Training- and Reference-based Methods

We explore the methods trained on auxiliary datasets and utilize the normal images in the target domain as references during inference. An early work in this category is APRIL-GAN (few-shot) (Chen et al., 2023b), which uses a linear layer trained with auxiliary datasets. Similar to WinCLIP+ (Jeong et al., 2023), APRIL-GAN (few-shot) utilizes a few ID reference images with a memory bank-based approach (Roth et al., 2022). More recently, Zhu & Pang (2024) proposed an in-context-learning-based method called InCTRL. InCTRL trains a model to discriminate anomalies from normal samples by learning to identify residuals or discrepancies between query images and a set of few-shot normal images (in-context sample prompts) from auxiliary data. During inference, InCTRL identifies anomalies by measuring the discrepancy between the features of the query image and a few in-context normal samples from the target dataset.

5.3 Other Research Direction

5.3.1 Anomaly Detection with Localization Models

Some works (Cao et al., 2023a; Li et al., 2024a) tackle AD using foundation models for localization, such as SAM (Kirillov et al., 2023) or GroundingDINO (Liu et al., 2024e). SAA (Cao et al., 2023a) is a pioneering approach that integrates SAM into anomaly segmentation tasks. It first employs prompt-driven object detection techniques such as GroundingDINO (Liu et al., 2024e) to generate box-level regions conditioned on prompts, highlighting the target anomaly areas. These generated boxes serve as prompts for SAM, which then produces the final anomaly segmentation results. Building on this, SAA+ (Cao et al., 2023a) introduces hybrid prompts that blend domain-specific knowledge with contextual image information, helping to reduce ambiguities inherent in language-based prompts. In contrast, CLIPSAM (Li et al., 2025) replaces GroundingDINO with CLIP, leveraging its superior localization capability to provide more precise prompts

for SAM. To bridge the domain gap from natural images, Yang et al. (2025) propose a promptable anomaly segmentation model with SAM, incorporating a novel Self-Perception Tuning (SPT) method. They first design a Self-Draft Tuning strategy, in which SAM initially generates a coarse draft of the anomaly mask, followed by a mask refinement process. Furthermore, they introduce a Visual-Relation-Aware Adapter to enhance the internal perception knowledge of discriminative relation information during decoding.

Given that AD necessitates localization, it is expected that the number of works employing localization foundation models such as SAM will continue to increase.

5.3.2 Medical Anomaly Detection

While most works on VLM-based AD focus on industrial AD, recent studies have begun to challenge medical anomaly detection (medical AD) (Chen et al., 2023c; Zhu & Pang, 2024; Huang et al., 2024; Hua et al., 2024; Zhang et al., 2024c; Cao et al., 2024c). VLM-based medical AD is a challenging area as well as industrial AD due to the larger gap between different data modalities. A representative work on medical AD is MVFA (Huang et al., 2024). MVFA is a method specifically tailored for medical AD. It incorporates multiple residual adapters into the CLIP's visual encoder to reduce the domain gap, enabling a stepwise enhancement of visual features across different levels. The future progression of medical AD and industrial AD offers an intriguing perspective, exploring whether these fields will develop independently or influence each other. However, when considering practical applications, it should be noted that medical AD faces the challenge that anomalies are not always describable in the language. Therefore, the development of medical AD methods that do not use CLIP is also important.

5.3.3 Video Anomaly Detection

Compared to OOD detection, one important point is that anomaly detection has been actively studied not only in image domains but also in video domains (Zhu et al., 2024b). Video Anomaly Detection (VAD) aims to automatically detect unusual events in videos, and it has many applications such as surveillance and monitoring (Wang et al., 2019b). VAD has several challenges. First, there is no clear and unified definition of anomalies, because what is normal or abnormal (for example, walking on a sidewalk vs. on a highway) depends on the context. Second, anomalies happen rarely and unpredictably, so it is difficult to collect high-quality datasets. This makes it hard to learn such patterns effectively. Because of these challenges and interests, VAD has become an active field, and many survey papers have been published (Suarez & Naval Jr, 2020; Abbas & Al-Ani, 2022; Jiao et al., 2023; Zhu et al., 2024b; Abdalla et al., 2024).

Before the VLM era, most VAD research focused on building task-specific models using unsupervised learning, one-class learning, or weakly supervised learning, because there was a lack of labeled anomaly data (Wu et al., 2024a). After the introduction of VLMs, the research direction changed a lot. Using powerful features from pre-trained VLMs, a lot of work tried zero-shot and few-shot VAD with much less task-specific data (Zanella et al., 2024b). Furthermore, the language understanding capabilities of VLMs have enabled new task settings, such as the integration of textual information (Zanella et al., 2024a;b) and the generation of natural language explanations for detected anomalies (Ye et al., 2025; Zhang et al., 2025a).

These recent trends show that the research direction of VAD has changed a lot with VLMs. On the other hand, OOD detection in video domains is still not widely studied. In this survey, our main goal is to better understand OOD detection itself, while referring to related areas such as AD only as supplementary information. Therefore, we do not give a full review of VAD in this survey, and instead refer readers to existing VAD survey papers (Suarez & Naval Jr, 2020; Abbas & Al-Ani, 2022; Jiao et al., 2023; Zhu et al., 2024b; Abdalla et al., 2024).

5.4 Discussion

We discuss the similarities and differences between VLM-based OOD detection and VLM-based AD to deepen our understanding of VLM-based OOD detection.

5.4.1 Difference between Each Methodology

Differing Scopes of OOD OOD detection and AD differ significantly in the scope of OOD (anomaly) they cover, which leads to differences in methodologies, particularly in the use of OOD prompts. As explained in Sec. 3, sensory AD is intended for specific use cases like industrial inspection, where samples deviating from predefined normality (*e.g.*, defective products) are considered anomalies (Yang et al., 2024; Ruff et al., 2021). In other words, in sensory AD, the anomaly space is limited to damaged objects with shared semantics, and anomalies like images of dogs are not expected. This limited anomaly space allows even simple prompts to achieve decent performance. Therefore, as shown in Table 2, all AD methods utilize anomaly prompts. Conversely, in OOD detection, as explained in Sec. 3.1, anything semantically different from the ID class is considered OOD. Thus, utilizing naive manual OOD prompts is forbidden (even if it improves benchmark performance). This vastness of the OOD space is the key factor differentiating the methodologies between the two fields.

Difficulty of Localization Task VLM-based OOD detection and VLM-based AD differ significantly in the difficulty of OOD (anomaly) localization tasks. In VLM-based AD, anomaly segmentation is a mainstream task, performed alongside classification in many papers. However, in VLM-based OOD detection, there has been no research on object-level OOD detection/segmentation. Object-level OOD detection aims to detect OOD objects (Du et al., 2022c;b;a). The inactivity is related to the size of the OOD space, and the too-vast space of OOD makes it difficult to identify OOD objects with prompts effectively. To pave the way for future development, the foundation models for localization such as SAM (Kirillov et al., 2023), which can segment individual objects, have the potential to address object-level OOD detection/segmentation. Object-level OOD detection/segmentation using SAM is a promising future research direction.

5.4.2 Similarity between Each Methodology

Each Problem Setting The existing problem settings for VLM-based AD and VLM-based OOD detection are similar. Both primarily focus on zero-shot and few-shot settings and can be categorized into training-free, auxiliary training-based, and ID training-based methods. Examining each problem more closely provides valuable insights into future directions for both fields.

History of Approaches The history of the progress of methods for VLM-based AD and VLM-based OOD detection is similar. For instance, both problems initially started with naive methods with manual OOD prompts (ZeroOE (Fort et al., 2021) for OOD detection, WinCLIP (Jeong et al., 2023) for AD). To address the issues with these initial approaches, subsequent methods emerged that replaced OOD prompts with learnable parameters (LSN (Nie et al., 2023) and NegPrompt (Li et al., 2024b) for OOD detection, and AnomalyCLIP (Zhou et al., 2024a) for AD). Therefore, by carefully examining each other's fields, there is potential for mutual enhancement and interaction in the future

6 Benchmarks and Experiments

In this section, we compare several representative VLM-based OOD detection methods.

6.1 Benchmarks and Evaluation Metrics

In OOD detection, it is common to consider an entire dataset as ID and to use several other datasets that are semantically disjoint from any ID classes as OOD datasets. However, as described in Sec.3.1, recent advances in hard OOD detection focus on constructing benchmarks where both ID and OOD samples are derived from the same large-scale dataset such as ImageNet. This setup is inspired by the benchmark design of OSR.

Following prior work (Zhang et al., 2024b; Ming et al., 2022a), we adopt the following terminology: among OOD samples drawn from datasets different from the ID dataset, those that are semantically similar to the ID classes are referred to as near OOD, while those that are semantically distant are referred to as far OOD.

When both ID and OOD samples are derived from the same dataset, as in OSR, we refer to this setting as hard OOD.

In this section, we report results on the widely used ImageNet OOD benchmark (Huang & Li, 2021) and two ImageNet-based hard OOD benchmark.

ImageNet Far OOD Benchmark In the ImageNet OOD benchmark, ImageNet is used as the ID dataset, while datasets such as iNaturalist (Van Horn et al., 2018) serve as OOD datasets. Among these, the most commonly used benchmark involves using the ImageNet validation set as the ID data and the following four datasets as OOD data: iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), Places (Zhou et al., 2017), and TEXTURE (Cimpoi et al., 2014). iNaturalist contains approximately 859,000 images of plants and animals across over 5,000 species. For OOD detection evaluation, 10,000 images are randomly sampled from 110 classes that are disjoint from ImageNet-1K. SUN consists of over 130,000 scene images spanning 397 categories. For OOD evaluation, 10,000 images are randomly sampled from 50 categories that do not overlap with ImageNet. Places is another scene-centric dataset with a similar concept space to SUN. For OOD detection, 10,000 images are randomly sampled from 50 non-overlapping classes with ImageNet-1K. TEXTURE comprises 5,640 real-world texture images from 47 categories. The entire dataset is used for OOD evaluation. When training is required, the training set of ImageNet is used for model training.

ImageNet Near OOD Benchmark In previous studies, two datasets, SSB-Hard (Vaze et al., 2022a) and NINCO (Bitterwolf et al., 2023), have been proposed as OOD datasets that are more semantically similar to ImageNet. SSB-Hard contains 49,000 images from 980 categories, which are selected from ImageNet-21K (Ridnik et al., 2021). NINCO is a dataset with 5,879 manually collected images (Bitterwolf et al., 2023), which is constructed to be semantically close to ImageNet-1K, but without any overlapping classes.

ImageNet-20 OOD Benchmark The ImageNet-20 OOD Benchmark is one of the earliest benchmarks proposed for hard OOD detection (Ming et al., 2022a). In this setting, ImageNet-20 is used as the ID dataset, and ImageNet-10, which has no overlapping categories, is used as the OOD dataset. The direction can also be reversed. These subsets are selected with reference to CIFAR-10 (Krizhevsky et al., 2009a) classes, making the semantic distance between ImageNet-20 and ImageNet-10 relatively small. ImageNet-20 contains 1,000 images, while ImageNet-10 contains 500 images.

ImageNet-X OOD Benchmark ImageNet-X is a recently proposed benchmark for hard OOD detection (Noda et al., 2025). It leverages the semantic hierarchy of ImageNet and creates ID and OOD splits within the same superclass. Both ID and OOD sets consist of 500 classes each. The ID and OOD subsets of ImageNet-X contain 25,000 images respectively.

6.2 Metrics

We use the ID accuracy and AUROC scores. AUROC scores measure the area under the Receiver Operating Characteristic (ROC) curve.

6.3 Experimental Setup

We compare commonly used eight OOD detection methods in zero-shot and few-shot settings.

Zero-shot Methods For the zero-shot setting, we use five methods: MCM (Ming et al., 2022a), GL-MCM (Miyai et al., 2025), NegLabel (Jiang et al., 2024), EOE (Cao et al., 2024a), and CLIPN (Wang et al., 2023a). As shown in Table 2, MCM and GL-MCM are categorized as methods without training and without OOD prompts. NegLabel and EOE are categorized as methods without training but with OOD prompts. CLIPN is categorized as a method with training and with OOD prompts.

Few-shot Methods For the few-shot settings, we use CoOp (Zhou et al., 2022c; Ming & Li, 2024a), LoCoOp (Miyai et al., 2023b), and ID-like-Prompt (IDPrompt) (Bai et al., 2024a). CoOp and LoCoOp use MCM as the detection method during inference. For CoOp and LoCoOp, we follow the hyperparameter

Method	Train Type	OOD Prompts	ImageNet			ImageNet-20		ImageNet-X	
			Far-OOD	Near-OOD	ID Acc.	Hard-OOD	ID Acc.	Hard-OOD	ID Acc.
Zero-shot									
MCM	Free	✗	90.66	68.91	66.73	97.38	91.30	74.30	75.97
GL-MCM	Free	✗	91.47	71.21	66.73	98.49	91.30	75.53	75.97
NegLabel	Free	✓	**94.15**	73.33	66.73	97.47	91.30	72.15	75.97
EOE	Free	✓	92.95	70.22	66.73	98.16	91.30	77.07	75.97
CLIPN-A	Auxiliary	✓	94.11	**80.94**	66.73	97.16	91.30	**77.08**	75.97
Few-shot									
CoOp	ID	✗	91.16	76.46	**71.77**	97.78	**94.80**	75.23	80.38
LoCoOp	ID	✗	93.02	69.11	71.61	**99.04**	94.67	75.17	**80.42**
ID-like-Prompt	ID	✓	92.46	67.18	68.43	92.90	89.80	66.01	77.18

Table 3: Comparison of OOD detection methods across ImageNet, ImageNet-20, and ImageNet-X. We use AUROC for the evaluation of OOD detection.

settings from previous studies and train with 16 shots. On the other hand, since IDPrompt requires higher training costs, we follow the original implementation (Bai et al., 2024a) and conduct training with only 1 shot.

6.4 Experimental Results

The experimental results are summarized in Table 3. Below, we highlight several key findings.

Performance Rankings Vary Across Datasets From Table 3, we observe that the ranking of methods differs significantly depending on the dataset. For example, LoCoOp outperforms CoOp on the standard ImageNet far OOD benchmark, but underperforms on the near OOD benchmarks. This suggests that LoCoOp's OOD regularization strategy, which uses background OOD images during training, is especially effective for detecting semantically distant classes such as in far OOD settings, but not effective for near OOD settings. These results indicate the importance of evaluating OOD detection methods across multiple benchmarks, rather than relying on a single one.

ImageNet-20 May Already Be Saturated On ImageNet-20, most methods achieve over 97% accuracy, making it difficult to distinguish performance differences among methods. Therefore, future research on OOD detection is likely to shift toward benchmarks with more diverse and larger numbers of classes.

Room for Improvement Remains in Near and Hard-OOD Detection The results on ImageNet near OOD and ImageNet-X show that the methods with the highest score reach only around 80% accuracy. This indicates significant room for further improvement. We believe that future research in OOD detection will increasingly focus on these challenging benchmarks.

7 Evolution in LVLM Era

In this section, we introduce the early advances in OOD detection and AD in the Large Vision Language Models (LVLM) era. While previous sections focused on VLMs such as CLIP, this section shifts our focus to the more emerging topic of "Large" VLMs. Recent advancements in computer vision have led to the emergence of LVLMs such as GPT-4V (OpenAI et al., 2023) and LLaVA (Liu et al., 2024b). Although these fields are still in their early stages with limited papers, this survey provides a deep introduction to each problem in the hope that our detailed review can help foster further advancements in this area.

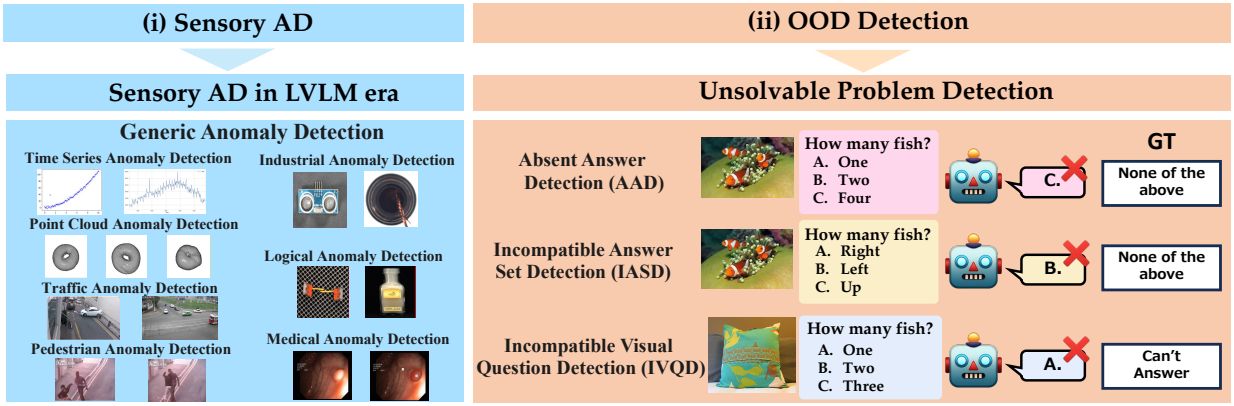

Figure 6: Overview of the evolution of each problem in the Large Vision Language Model (LVLM) era. (i) Sensory AD has become a highly active and noteworthy field in the LVLM era (Cao et al., 2023b). (ii) Notably, OOD detection is evolving into a new task called Unsolvable Problem Detection (Miyai et al., 2024) in the LVLM era. Figure adapted partially from (Cao et al., 2023b) and from (Miyai et al., 2024).

7.1 Change of Each Problem

i. Sensory AD → Sensory AD Sensory AD has continued to develop in the LVLM era (Gu et al., 2024b; Cao et al., 2023b; Li et al., 2023b). The use of LVLMs has made AD applicable in many domains and modalities (Cao et al., 2023b).

ii. OOD Detection → Unsolvable Problem Detection In the LVLM era, OOD detection has evolved into a new task termed Unsolvable Problem Detection (UPD) (Miyai et al., 2024). UPD evaluates the LVLMs' ability to recognize and abstain from answering unexpected or unsolvable input questions, effectively expanding the scope of OOD detection into the context of Visual Question Answering (VQA) tasks. This shift to the VQA task has significantly broadened the concept of OOD detection to a wider range of AI tasks involving LVLMs.

7.2 Unsolvable Problem Detection

7.2.1 Summary of Problem

Background Following the recent revolutionary development of LLMs (Chiang et al., 2023; Touvron et al., 2023; Wei et al., 2023; Zhao et al., 2023), LVLMs (Awadalla et al., 2023; Dai et al., 2023; Liu et al., 2024b; Wang et al., 2023b; Ye et al., 2023; Li et al., 2023a; Lin et al., 2024) have demonstrated remarkable capabilities in diverse applications (Liu et al., 2024a;f; Yue et al., 2024). However, a significant concern has arisen regarding the reliability of these models, specifically their ability to generate accurate and trustworthy information. These models frequently produce incorrect or misleading information, a phenomenon referred to as "hallucination" (Bai et al., 2024b). Among the various hallucination issues (Bai et al., 2024b), the challenge of identifying out-of-place questions is crucial for deploying LVLMs in safety-critical applications. This challenge extends the concept of OOD detection to the VQA tasks for LVLMs and represents a specific aspect of LVLMs' trustworthiness.

Definition Unsolvable Problem Detection (UPD) is a task to measure the trustworthiness of LVLMs, which is designed to evaluate models' capacity to withhold answers when faced with unsolvable problems. The UPD task can be categorized into three distinct problem types: Absent Answer Detection (AAD), Incompatible Answer Set Detection (IASD), and Incompatible Visual Question Detection (IVQD). Examples in each setting are shown in the right of Fig. 6. AAD evaluates the model's capacity to determine when the correct answer is absent from the provided options. IASD assesses the model's ability to discern answer choices that are completely irrelevant to the given question and image. IVQD evaluates the model's capacity to discern whether a question and image are unrelated or mismatched.

Benchmark Miyai et al. (2024) created MM-UPD Bench for the UPD challenge. MM-UPD encompasses MM-AAD, MM-IASD, and MM-IVQD benchmarks for each UPD problem. Each benchmarks are created on the top of MMBench (dev) (Liu et al., 2024f), which is a systematically designed objective benchmark for evaluating various abilities of LVLMs. Following the definition of each ability in MMBench (*e.g.*, "Coarse Perception: Image Scene" and "Logic Reasoning: Future Prediction"), MM-UPD evaluates the trustworthiness of LVLMs from various abilities.

Although MM-UPD is the main benchmark, the adaptation cost of creating UPD problems is not high, making it highly applicable to other benchmarks. For instance, the recently proposed MuirBench (Wang et al., 2025), a benchmark for multi-image understanding, has incorporated the concept of UPD by adding unsolvable problems.

Application UPD has a wide range of applications, from everyday use of LVLMs to robot manipulation. Especially when incorporating LVLMs into safety-critical domains such as robot manipulation (Liu et al., 2024d) and autonomous driving (Li et al., 2024d), there is a risk of significant problems if the LVLM fails to identify erroneous user questions and makes incorrect predictions. UPD serves as a task to ensure safety in such safety-critical scenarios.

Evaluation UPD employs three evaluation metrics: Standard Accuracy, which measures performance on solvable problems; UPD Accuracy, which assesses accuracy on unsolvable problems; and Dual Accuracy, which considers both aspects. Dual Accuracy considers a response correct only if the model successfully answers both the solvable and unsolvable problems within each paired set. The rationale is that ideal LVLMs should not only give correct answers for the solvable problems but need to withhold answering for unsolvable problems.

7.2.2 Findings

In the following, we briefly summarized the findings of the UPD challenge (Miyai et al., 2024).

1. Most LVLMs Hardly Hesitate to Answer. Most LVLMs, especially open-source LVLMs, have significantly low UPD accuracies, which indicates the difficulty of the UPD challenge. For example, LLaVA-1.5 (Liu et al., 2024b) and CogVLM (Wang et al., 2023b), which are state-of-the-art LVLMs, completely fail to withhold answering. GPT-4V achieves higher performances than other LVLMs due to its safety training process (OpenAI, 2023). However, there is still a performance gap from the upper bound scores.

2. Performance Tendency Differs a lot by Each Ability in the Benchmark. The performance of LVLMs differs in each ability in the MM-UPD Bench. For instance, GPT-4V has its limitation in attribute comparison and LLaVA-NeXT-34B has its limitation in object localization.

3. Effective Prompt Strategies Vary Across Different LVLMs Effective prompt strategies vary across different LVLMs. In the original paper, they experimented with an option-based prompt approach that adds an option of "None of the above" and an instruction-based approach that adds an instruction "If all options are incorrect, answer None of the above". As a result, the effectiveness of each approach differs significantly depending on the type of LVLMs. This highlights the difficulty of finding an effective prompt strategy for all LVLMs.

7.3 Anomaly Detection in LVLM Era

7.3.1 Summary of Problem

Background Anomaly detection is a crucial task in a variety of domains and data types. However, existing anomaly detection models are often designed for specific domains or modalities (Cao et al., 2023b). Also, current AD methods only provide an anomaly score for the test sample and require a manual threshold to distinguish between normal and anomalous instances for each sample (Gu et al., 2024b). To facilitate real-world applications, developing a system capable of expressing anomalies in natural language across various modalities and domains is crucial for ensuring accessibility to a wider range of users.

Definition The definition of AD remains consistent with conventional and VLM-based AD, aiming to identify samples that deviate from predefined normality. The key difference lies in the output format. Traditional methods generate an anomaly score as a numerical value, requiring a manual threshold to determine whether a sample is anomalous. In contrast, AD with LVLMs aims to recognize and describe anomalies using text, removing the need for manual thresholds and improving human interpretability.

Benchmark Since the field of AD with LVLMs is in its infant stage, there are still no unified benchmarks. AnomalyGPT (Gu et al., 2024b) focuses on industrial image anomaly detection/localization and uses the standard benchmarks MVTec-AD (Bergmann et al., 2019) and VisA (Zou et al., 2022). More recently, Cao et al. (2023b) extend the domain and modality and demonstrate the applications in industrial image anomaly detection/localization (*e.g.*, MVTec-AD (Bergmann et al., 2019)), point cloud anomaly detection (MVTec 3D (Bergmann et al., 2022b)), medical image anomaly detection/localization (*e.g.*, Chest X-ray (Kermany et al., 2018), Head CT (Felipe, 2018)), logical anomaly detection (*e.g.*, MVTec LOCO (Bergmann et al., 2022a)), pedestrian anomaly detection (*e.g.*, UCF-Crime Dataset (Sultani et al., 2018)), traffic anomaly detection (*e.g.*, Kaggle Accident Detection (Kay, 2018)), and time series anomaly detection (*e.g.*, Outlier Detection Dataset (User, 2021)).

Evaluation Evaluation in anomaly detection with LVLMs is an open challenge. AnomalyGPT (Gu et al., 2024b) asks LVLMs the question "Is there an anomaly in this image?" and determines anomaly or normal based on the simple rule-based approach of whether the response contains a "yes" or "no". However, this rule-based approach is not robust, as a response is considered correct even if the explanation following "yes" is completely incorrect. On the other hand, Cao et al. (2023b) conducted only qualitative evaluations and left quantitative evaluations as an open challenge. Therefore, the evaluation of anomaly detection by LVLM is a future challenge.

7.3.2 Findings

Cao et al. (2023b) described the observations of GPT-4V in the paper, so we briefly summarized them here.

1. GPT-4V Excels in Zero/One-shot Settings across Various Modalities and Fields. GPT-4V shows proficiency in identifying anomalies in multi-modality (*e.g.*, images, point clouds, X-rays) and multi-field (*e.g.*, industrial, medical, pedestrian, traffic, and time series anomaly detection). In addition, GPT-4V demonstrates strong performance in both zero-shot and one-shot settings.

2. GPT-4V can Understand Both Global and Fine-grained Anomalies. GPT-4V can recognize both global and local abnormal patterns or behaviors, which indicates the ability to understand global and fine-grained semantics.

3. GPT-4V can be Enhanced with Increasing Prompts. By giving more context and information, the model significantly improves its ability to detect anomalies accurately.

8 Future Directions

In this section, we discuss the future directions of OOD detection. We explore not only OOD detection for VLMs but also single-modal OOD detection, with a specific focus on emerging challenges as VLMs evolve. For a discussion of the long-standing challenges in OOD detection, we can refer the readers to the previous generalized OOD detection paper (Yang et al., 2024).

8.1 OOD Detection for Vision Language Models

a. Hard OOD Detection Hard OOD detection will become increasingly important in the future due to its high practicality and the challenging nature of the problem. Hard OOD detection utilizes the OSR benchmark setup, where some classes within a single dataset are designated as ID and others as OOD. In this field, not only small datasets such as ImageNet-10 and ImageNet-20 (Ming et al., 2022a) but also datasets with a larger number of classes such as ImageNet-protocol (Palechor et al., 2023) and ImageNet-X (Noda

et al., 2025) have been recently proposed. Many existing studies, such as LoCoOp (Miyai et al., 2023b) and LSN (Nie et al., 2023), primarily use the common ImageNet OOD benchmark, so hard OOD detection has not yet been well studied. This field will develop further in the future.

b. Post-hoc Methods To propose post-hoc methods is important for the fundamental performance improvement of VLM-based OOD detection. The methods of directly employing an ID classifier such as MCM (Ming et al., 2022a) are called post-hoc methods. Prior to CLIP, various approaches were proposed (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Liu et al., 2020; Sastry & Oore, 2020; Sun & Li, 2022; Sun et al., 2022; Sastry & Oore, 2019). Nevertheless, VLM-based post-hoc methods often underperform methods with additional OOD prompts, so they are not extensively researched in zero-shot OOD detection. However, we should focus on the scalability of post-hoc methods. The post-hoc methods (Ming et al., 2022a; Miyai et al., 2025) can be easily applied to many subsequent methods (Miyai et al., 2023b; Ming et al., 2022b; Nie et al., 2023; Li et al., 2024b), bringing fundamental performance improvements even for the few-shot setting. Furthermore, recently, post-hoc methods specifically tailored for prompt learning methods have also emerged (Jung et al., 2024). Therefore, proposing post-hoc methods and demonstrating the improvements not only in zero-shot but also in subsequent few-shot settings (Miyai et al., 2023b; Ming et al., 2022b) is crucial. This field should continue evolving, mirroring its growth before the advent of CLIP.

c. Bridging the Gap with Closed-set Classifiers. OOD detection ensures the safety of ID classifiers, so it is crucial to bridge the gap between the advancements in existing closed-set classifiers and OOD detection. Currently, the representative method for few-shot OOD detection is LoCoOp (Miyai et al., 2023b), a text prompt learning method based on CoOp. However, in closed-set settings, VLM-based few-shot learning methods have been proposed other than CoOp (Chen et al., 2023a; Bulat & Tzimiropoulos, 2023; Lu et al., 2022). Therefore, adopting recent methods for OOD detection is essential for bridging the gap with closed-set ID classifiers.

d. Training-free Few-shot OOD Detection The research direction of training-free few-shot OOD detection is still in its infant stage (Chen et al., 2024b; Zhang et al., 2024f). Considering the advancements of training-free methods in VLM-based AD, we anticipate a similar trajectory for VLM-based OOD detection. Future directions include refining adapter-based methods or leveraging external knowledge such as retrieval augmentation (Udandarao et al., 2023; Ming & Li, 2024b). Addressing training-free few-shot OOD detection is a pivotal step towards realizing more computationally efficient OOD detection in the future.

e. Full-spectrum OOD Detection VLM-based full-spectrum OOD (FS-OOD) detection is a promising research area (Lu et al., 2023; Yang et al., 2023). In practical applications, there is a strong motivation to create models that can not only detect semantically shifted OOD inputs but also generalize to covariate-shifted data (Yang et al., 2023; Bai et al., 2023). Within VLM-based methods, OOD detection and generalization are often discussed in separate contexts (Miyai et al., 2023b; Khattak et al., 2023), resulting in a trade-off between detection and generalization performance (Lafon et al., 2024). Therefore, we need to discuss both aspects together to mitigate the trade-off.

f. Open-vocabulary OOD Detection Open-vocabulary OOD (OV-OOD) detection has a high practical potential, but it is still in its infant stages (Li et al., 2024b). The open-vocabulary setting has been actively explored in AD (5.1) and has the potential to become increasingly important in the field of OOD detection as well.

g. Object-level OOD Detection Object-level OOD detection remains an unexplored area in VLM-based OOD detection. As discussed in Sec. 5.4, this is due to the vastness of the OOD space, which makes it difficult to identify OOD objects using texts. To pave the way for future advancements in object-level OOD detection/segmentation, foundation models for localization, such as SAM (Kirillov et al., 2023), offer a promising solution. By integrating these models with methods such as MCM (Ming et al., 2022a), we can potentially achieve object-level OOD detection and segmentation, opening up a new frontier in OOD detection research.

8.2 Single-modal OOD Detection

a. Leveraging Large Pre-trained Models Leveraging large pre-trained models is important for single-modal OOD detection. Numerous methods for OOD detection conduct experiments using backbones trained from scratch and do not utilize pre-trained models (Yang et al., 2024; Zhang et al., 2024b; Hendrycks & Gimpel, 2017; Leys et al., 2018; Liu et al., 2020; Wang et al., 2022a; Sun & Li, 2022; Sun et al., 2022). In a recent study, Miyai et al. (2023a) systematically investigated the impact of pre-training on OOD detection from both the perspectives of the types of OOD data and pre-training algorithms (Chen et al., 2020; Caron et al., 2021). Dong et al. (2023) explored parameter-efficient learning for single-modal OOD detection and proposed DSGF, which leverages both fine-tuned features and original pre-trained features. While leveraging large pre-trained models (Dosovitskiy et al., 2021) with lightweight tuning is an active area of research in single-modal closed-set classification (Hu et al., 2022; Zhang et al., 2024e), there have been limited studies for single-modal OOD detection, which presents a promising avenue for future research.

b. Real-world Benchmarks and Evaluations Considering the future development of VLM-based OOD detection, there should be increasing focus on expanding the scope of benchmarks to encompass real-world scenarios where CLIP is less applicable. For instance, recently, Baek et al. (2024) introduced ImageNet-ES, consisting of variations in environmental and camera sensor factors. Besides, utilizing datasets such as WILDS (Koh et al., 2021; Cultrera et al., 2023), which consider real-world data shifts, or datasets for medical OOD detection (Hong et al., 2024), can provide valuable insights, especially in safety-critical applications such as autonomous driving and medical image analysis.

9 Conclusion

In this survey, we comprehensively review the evolution of the five problems including AD, ND, OSR, OOD detection, and OD in the VLM era, and propose a framework of *generalized OOD detection v2*. Our framework identifies OOD detection and AD as the primary challenges in the VLM era. By articulating the shifts in the definitions, problem settings, benchmarks and methodologies, we encourage subsequent works to accurately understand their evolving target problems in the VLM era. By shedding light on recent studies in the LVLM era, we hope that researchers within each community can identify promising research directions in this emerging era. By providing future directions, we hope that our survey will clarify the tasks to be tackled by future works in the VLM era, facilitating future advances in the right direction.

References

Zainab K. Abbas and Ayad A. Al-Ani. A comprehensive review for video anomaly detection on videos. In *CSASE*, 2022.

Moshira Abdalla, Sajid Javed, Muaz Al Radi, Anwaar Ulhaq, and Naoufel Werghi. Video anomaly detection in 10 years: A survey and outlook. *arXiv preprint arXiv:2405.19387*, 2024.

Nikolas Adaloglou, Felix Michels, Tim Kaiser, and Markus Kollmann. Adapting contrastive language-image pretrained (clip) models for out-of-distribution detection. *TMLR*, 2024.

Charu C Aggarwal and Philip S Yu. Outlier detection for high dimensional data. In *ACM SIGMOD*, 2001.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Akshatha Arodi, Margaux Luck, Jean-Luc Bedwani, Aldo Zaimi, Ge Li, Nicolas Pouliot, Julien Beaudry, and Gaétan Marceau Caron. Cableinspect-AD: An expert-annotated anomaly detection dataset. In *NeurIPS, Datasets and Benchmarks Track*, 2024.

Deegan J Atha and Mohammad R Jahanshahi. Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Structural Health Monitoring*, 2018.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *CVPR*, 2024.

Haoyue Bai, Gregory Canal, Xuefeng Du, Jeongyeol Kwon, Robert D Nowak, and Yixuan Li. Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection. In *ICML*, 2023.

Yichen Bai, Zongbo Han, Changqing Zhang, Bing Cao, Xiaoheng Jiang, and Qinghua Hu. Id-like prompt learning for few-shot out-of-distribution detection. In *CVPR*, 2024a.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024b.

Irad Ben-Gal. Outlier detection. In *Data Mining and Knowledge Discovery Handbook*. 2005.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *IJCV*, 130(4):947–969, 2022a.

Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The mvtec 3d-ad dataset for unsupervised 3d anomaly detection and localization. In *VISAPP*, 2022b.

Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023.

Terrance E Boult, Steve Cruz, Akshay Raj Dhamija, M Gunther, James Henrydoss, and Walter J Scheirer. Learning and the unknown: Surveying steps toward open world recognition. In *AAAI*, 2019.

Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *CVPR*, 2023.

Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 8:132330–132347, 2020.

Chentao Cao, Zhun Zhong, Zhanke Zhou, Yang Liu, Tongliang Liu, and Bo Han. Envisioning outlier exposure by large language models for out-of-distribution detection. In *ICML*, 2024a.

Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. In *ICLR*, 2022.

Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023a.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023b.

Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024b.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *ECCV*, 2024c.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys*, 41(3):1–58, 2009.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023a.

Mengyuan Chen, Junyu Gao, and Changsheng Xu. Conjugated semantic pool improves ood detection with pre-trained vision-language models. In *NeurIPS*, 2024a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Xinyi Chen, Yaohui Li, and Haoxing Chen. Dual-adapter: Training-free dual adaptation for few-shot out-of-distribution detection. *arXiv preprint arXiv:2405.16146*, 2024b.

Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023b.

Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, Yunsheng Wu, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. *arXiv preprint arXiv:2311.00453*, 2023c.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Wen-Hsuan Chu and Kris M Kitani. Neural batch sampling with reinforcement learning for semi-supervised anomaly detection. In *ECCV*, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

Luca Cultrera, Lorenzo Seidenari, and Alberto Del Bimbo. Leveraging visual attention for out-of-distribution detection. In *ICCV*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*, 2021.

Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2023.

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Choubo Ding and Guansong Pang. Zero-shot out-of-distribution detection with outlier label exposure. In *IJCNN*, 2024.

Jiuqing Dong, Yongbin Gao, Heng Zhou, Jun Cen, Yifan Yao, Sook Yoon, and Park Dong Sun. Towards few-shot out-of-distribution detection. *arXiv preprint arXiv:2311.12076*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *CVPR*, 2024.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. Siren: Shaping representations for detecting out-of-distribution objects. In *NeurIPS*, 2022a.

Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don' t know from videos in the wild. In *CVPR*, 2022b.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don' t know by virtual outlier synthesis. In *ICLR*, 2022c.

Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI*, 2022.

Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. Manta: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *CVPR*, 2025.

Kitamura Felipe. Head ct - hemorrhage. https://www.kaggle.com/datasets/felipekitamura/headct-hemorrhage, 2018.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *NeurIPS*, 2021.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE TPAMI*, 43(10):3614–3631, 2020.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *ACMMM*, 2024a.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*, 2024b.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Univad: A training-free unified model for few-shot visual anomaly detection. In *CVPR*, 2025.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *CVPR*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for AI research. *arXiv preprint arXiv:2206.05862*, 2022.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021a.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021b.

Chih-Hui Ho, Kuan-Chuan Peng, and Nuno Vasconcelos. Long-tailed anomaly detection with learnable class names. In *CVPR*, 2024.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 2004.

Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Liujie Hua, Yueyi Luo, Qianqian Qi, and Jun Long. Medicalclip: Anomaly-detection domain generalization with asymmetric constraints. *Biomolecules*, 14(5):590, 2024.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *CVPR*, 2024.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021.

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *ICUMT*, pp. 66–71, 2021.

Xuhui Jia, Kai Han, Yukun Zhu, and Bradley Green. Joint representation learning and novel category discovery on single-and multi-modal data. In *ICCV*, 2021.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: The first-ever comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *ICLR*, 2025.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models. In *ICLR*, 2024.

Runyu Jiao, Yi Wan, Fabio Poiesi, and Yiming Wang. Survey on video anomaly detection in dynamic scenes with moving cameras. *Artificial Intelligence Review*, 56(Suppl 3):3515–3570, 2023.

KJ Joseph, Sujoy Paul, Gaurav Aggarwal, Soma Biswas, Piyush Rai, Kai Han, and Vineeth N Balasubramanian. Novel class discovery without forgetting. In *ECCV*, 2022.

Myong Chol Jung, He Zhao, Joanna Dipnall, Belinda Gabbe, and Lan Du. Enhancing near ood detection in prompt learning: Maximum gains, minimal costs. *arXiv preprint arXiv:2405.16091*, 2024.

C. Kay. Accident detection from cctv footage. https://www.kaggle.com/datasets/ckay16/accident-detection-fromcctv-footage, 2018. Kaggle dataset.

Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009a.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

Sangkyung Kwak, Jongheon Jeong, Hankook Lee, Woohyuck Kim, Dongho Seo, Woojin Yun, Wonjin Lee, and Jinwoo Shin. Few-shot anomaly detection via personalization. *IEEE Access*, 2024.

Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *ECCV*, 2024.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

Christophe Leys, Olivier Klein, Yves Dominicy, and Christophe Ley. Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of Experimental Social Psychology*, 74:150–156, 2018.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023a.

Chenghao Li, Lei Qi, and Xin Geng. A sam-guided two-stream lightweight model for anomaly detection. *ACM TOMM*, 2024a.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *CVPR*, 2021.

Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *Neurocomputing*, 618:129122, 2025.

Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *CVPR*, 2024b.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *CVPR*, 2024c.

Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024d.

Yixia Li, Boya Xiong, Guanhua Chen, and Yun Chen. Setar: Out-of-distribution detection with selective low-rank approximation. In *NeurIPS*, 2024e.

Yiyue Li, Shaoting Zhang, Kang Li, and Qicheng Lao. One-to-normal: Anomaly personalization for few-shot anomaly detection. In *NeurIPS*, 2024f.

Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070*, 2023b.

Jian Liang, Lijun Sheng, Zhengbo Wang, Ran He, and Tieniu Tan. Realistic unsupervised CLIP fine-tuning with universal entropy optimization. In *ICML*, 2024.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024.

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *CVPR*, 2021.

Fuxiao Liu, Hao Tan, and Chris Tensmeyer. Documentclip: Linking figures and main body text in reflowed documents. In *ICPRAI*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024b.

Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024c.

Jinyi Liu, Yifu Yuan, Jianye Hao, Fei Ni, Lingzhi Fu, Yibin Chen, and Yan Zheng. Enhancing robotic manipulation with ai feedback from multimodal large language models. *arXiv preprint arXiv:2402.14245*, 2024d.

Kai Liu, Zhihang Fu, Chao Chen, Sheng Jin, Ze Chen, Mingyuan Tao, Rongxin Jiang, and Jieping Ye. Category-extensible out-of-distribution detection via hierarchical context descriptions. In *NeurIPS*, 2023b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024e.

Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

Xixi Liu and Zach Christopher. Tag: Text prompt augmentation for zero-shot out-of-distribution detection. In *ECCV*, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024f.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. In *ICLR*, 2021.

Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Exposing outlier exposure: What can be learned from few, one, and zero outlier images. *TMLRh*, 2022.

Fan Lu, Kai Zhu, Kecheng Zheng, Wei Zhai, and Yang Cao. Likelihood-aware semantic alignment for full-spectrum out-of-distribution detection. *arXiv preprint arXiv:2312.01732*, 2023.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *CVPR*, 2025.

Atefeh Mahdavi and Marco Carvalho. A survey on open set recognition. *arXiv preprint arXiv:2109.00893*, 2021.

Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003a.

Markos Markou and Sameer Singh. Novelty detection: a review—part 2:: neural network based approaches. *Signal processing*, 83(12):2499–2521, 2003b.

Dubravko Miljković. Review of novelty detection methods. In *MIPRO*, 2010.

Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *ECCV*, 2024.

Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *IJCV*, 132(2):596–609, 2024a.

Yifei Ming and Yixuan Li. Understanding retrieval-augmented task adaptation for vision-language models. In *ICML*, 2024b.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022a.

Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *AAAI*, 2022b.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *ISIE*, pp. 01–06, 2021.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Can pre-trained networks detect familiar out-of-distribution data? *arXiv preprint arXiv:2310.00847*, 2023a.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *NeurIPS*, 2023b.

Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Li, Ziwei Liu, and Kiyoharu Aizawa. Unsolvable problem detection: Evaluating trustworthiness of vision language models. *arXiv preprint arXiv:2403.20331*, 2024.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection. *IJCV*, pp. 1–11, 2025.

Sina Mohseni, Haotao Wang, Zhiding Yu, Chaowei Xiao, Zhangyang Wang, and Jay Yadawa. Practical machine learning safety: A survey and primer. *arXiv preprint arXiv:2106.04823*, 2021.

Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *ICLR*, 2023.

Shiho Noda, Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. A benchmark and evaluation for real-world out-of-distribution detection using vision-language models. *arXiv preprint arXiv:2501.18463*, 2025.

OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Andres Palechor, Annesha Bhoumik, and Manuel Günther. Large-scale open-set classification protocols for imagenet. In *WACV*, 2023.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2), 2021.

Sangha Park, Jisoo Mok, Dahuin Jung, Saehyung Lee, and Sungroh Yoon. On the powerfulness of textual outlier exposure for visual ood detection. In *NeurIPS*, 2023.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal processing*, 99:215–249, 2014.

Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *ECCV*, 2024.

Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *CVPR*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS, Datasets and Benchmarks Track*, 2021.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In *NeurIPS*, 2021.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *TMLR*, 2022.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. In *NeurIPS-W*, 2019.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *ICML*, 2020.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *ICML*, 2023.

Jessie James P Suarez and Prospero C Naval Jr. A survey on deep learning techniques for video anomaly detection. *arXiv preprint arXiv:2009.14146*, 2020.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018.

Han Sun, Yunkang Cao, and Olga Fink. Cut: A controllable, universal, and training-free visual anomaly generation framework. In *CVPR*, 2025.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020.

Masato Tamura. Random word data augmentation with clip for zero-shot anomaly detection. In *BMVC*, 2023.

Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *ICML*, 2023.

Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. An effective baseline for robustness to distributional shift. *arXiv preprint arXiv:2105.07107*, 2021.

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE TPAMI*, 30(11):1958–1970, 2008.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). In *NeurIPS*, 2023.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *ICCV*, 2023.

Stack Exchange User. Simple outlier detection for time series. https://stats.stackexchange.com/questions/427327/simple-outlier-detection-for-time-series, 2021. Accessed: 2024-07-02.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022a.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022b.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *ICLR*, 2025.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022a.

Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? In *NeurIPS*, 2021.

Hongzhi Wang, Mohamed Jaward Bah, and Mohamed Hammad. Progress in outlier detection techniques: A survey. *IEEE Access*, 7:107964–108000, 2019a.

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, 2023a.

Lei Wang, Du Q Huynh, and Moussa Reda Mansour. Loss switching fusion with similarity search for video classification. In *ICIP*, 2019b.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023b.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, 2018.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022b.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Peng Wu, Chengyu Pan, Yuting Yan, Guansong Pang, Peng Wang, and Yanning Zhang. Deep learning for video anomaly detection: A review. *arXiv preprint arXiv:2409.05383*, 2024a.

Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *CVPR*, 2024b.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

Ruiyao Xu and Kaize Ding. Large language models for anomaly and out-of-distribution detection: A survey. In *NAACL Findings*, 2025.

Hui-Yue Yang, Hui Chen, Ao Wang, Kai Chen, Zijia Lin, Yongliang Tang, Pengcheng Gao, Yuming Quan, Jungong Han, and Guiguang Ding. Promptable anomaly segmentation with sam through self-perception tuning. In *AAAI*, 2025.

Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *IJCV*, 131(10): 2607–2622, 2023.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *IJCV*, pp. 1–28, 2024.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *CVPR*, 2025.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Jihun Yi and Sungroh Yoon. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *ACCV*, 2020.

Geng Yu, Jianing Zhu, Jiangchao Yao, and Bo Han. Self-calibrated tuning of vision-language models for out-of-distribution detection. In *NeurIPS*, 2024.

Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, 2020.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

EungGu Yun, Heonjin Ha, Yeongwoo Nam, and Bryan Dongik Lee. Language-assisted feature transformation for anomaly detection. In *ICLR*, 2025.

Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *CVIU*, 249:104163, 2024a.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *CVPR*, 2024b.

Giacomo Zara, Subhankar Roy, Paolo Rota, and Elisa Ricci. Autolabel: Clip-based framework for open-set video domain adaptation. In *CVPR*, 2023.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021.

Fanhu Zeng, Zhen Cheng, Fei Zhu, Hongxin Wei, and Xu-Yao Zhang. Local-prompt: Extensible local prompts for few-shot out-of-distribution detection. In *ICLR*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.

Boxuan Zhang, Jianing Zhu, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han. What if the input is expanded in ood detection? In *NeurIPS*, 2024a.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *CVPR*, 2025a.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *DMLR*, 2024b.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *CVPR*, 2025b.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *ICLR*, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *ECCV*, 2022.

Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *MICCAI*, 2024c.

Yabin Zhang and Lei Zhang. Adaneg: Adaptive negative proxy guided ood detection with vision-language models. In *NeurIPS*, 2024.

Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *ECCV*, 2024d.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *IEEE TPAMI*, pp. 1–14, 2024e.

Zihan Zhang, Zhuo Xu, and Xiang Xiang. Vision-language dual-pattern matching for out-of-distribution detection. In *ECCV*, 2024f.

Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*, 2021.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 40(6):1452–1464, 2017.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE TPAMI*, 45(4):4396–4415, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022c.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *ICLR*, 2024a.

Zhi Zhou, Ming Yang, Jiang-Xin Shi, Lan-Zhe Guo, and Yu-Feng Li. Decoop: Robust prompt tuning with out-of-distribution detection. In *ICML*, 2024b.

Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Junran Wu. Do llms understand visual anomalies? uncovering llm's capabilities in zero-shot anomaly detection. In *ACMMM*, 2024a.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, 2024.

Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. In *NeurIPS, Datasets and Benchmarks Track*, 2024b.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.

Zuo Zuo, Yao Wu, Baoqiang Li, Jiahao Dong, You Zhou, Lei Zhou, Yanyun Qu, and Zongze Wu. Clip-fsac: Boosting clip for few-shot anomaly classification with synthetic anomalies. In *IJCAI-24*, 2024.