Towards VLM-based Hybrid Explainable Prompt Enhancement for Zero-Shot Industrial Anomaly Detection

Weichao Cai[1], **Weiliang Huang**[2], **Yunkang Cao**[3], **Chao Huang**[4*], **Fei Yuan**[1*], **Bob Zhang**[2], **Jie Wen**[5]

[1]School of Information, Xiamen University
[2]Department of Computer and Information Science, University of Macau
[3]School of Robotics, Hunan University
[4]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[5]School of Computer Science & Technology, Harbin Institute of Technology, Shenzhen
caiweichao0914@stu.xmu.edu.cn, yc47492@um.edu.mo, caoyunkang@hnu.edu.cn
huangch253@mail.sysu.edu.cn, yuanfei@xmu.edu.cn, bobzhang@um.edu.mo, wenjie@hit.edu.cn

Abstract

Zero-Shot Industrial Anomaly Detection (ZSIAD) aims to identify and localize anomalies in industrial images from unseen categories. Owing to the powerful generalization capabilities, Vision-Language Models (VLMs) have achieved growing interest in ZSIAD. To guide the model toward understanding and localizing the semantically complex industrial anomalies, existing VLM-based methods have attempted to provide additional prompts to the model through learnable text prompt templates. However, these zero-shot methods lack detailed descriptions of specific anomalies, making it difficult to classify and segment the diverse range of industrial anomalies accurately. To address the aforementioned issue, we firstly propose the multi-stage prompt generation agent for ZSIAD. Specifically, we leverage the Multi-modal Language Large Model (MLLM) to articulate the detailed differential information between normal and test samples, which can provide detailed text prompts to the model through further refinement and anti-false alarm constraint. Moreover, we introduce the Visual Fundamental Model (VFM) to generate anomaly-related attention prompts for more accurate localization of anomalies with varying sizes and shapes. Extensive experiments on seven real-world industrial anomaly detection datasets have shown that the proposed method not only outperforms recent SOTA methods, but also its explainable prompts provide the model with a more intuitive basis for anomaly identification.

1 Introduction

Industrial Anomaly Detection (IAD) aims to accurately identify and locate industrial product defects by advanced computer vision models and algorithms, which improves production efficiency and guarantees product quality. This is significant in promoting industrial automation to a higher level of development.

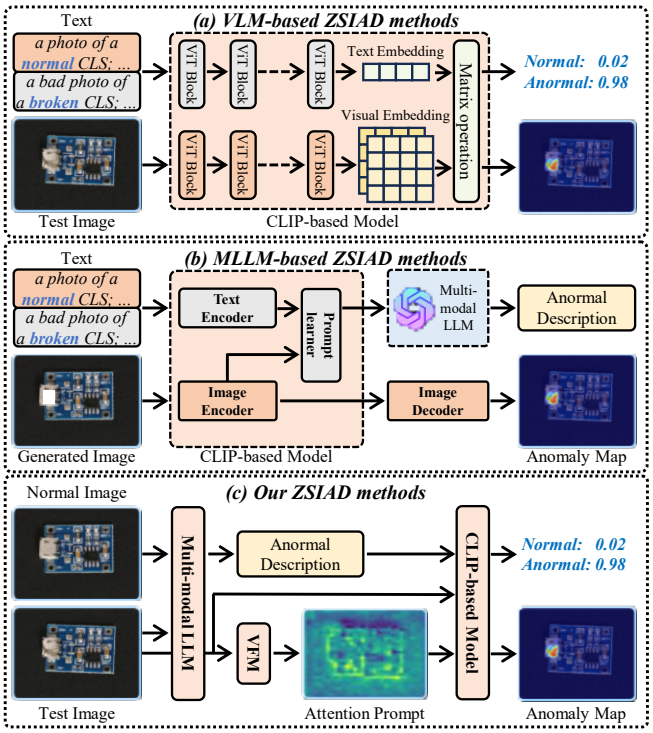

Figure 1: Comparisons between traditional VLM-Based method, existing Prompt-Enhancing methods and the proposed method in ZSIAD.

Given the scarcity of industrial anomaly samples in the real world, recent works have focused on how to achieve industrial anomaly detection by unsupervised and zero-shot methods. Unsupervised Industrial Anomaly Detection (UIAD) methods utilize normal samples to learn the distributional features of normal industrial products, which in turn identifies the differences between the input image and the normal

*Corresponding authors.

image. In this way, the detection and localization of industrial anomalies are mainly achieved through reconstruction [Huang *et al.*, 2021b; Huang *et al.*, 2022a] and feature embedding [Huang *et al.*, 2021a; Huang *et al.*, 2022b]. However, these methods usually require numerous normal samples for training to adequately learn the distributional features of normal data, which may be difficult to satisfy in real industrial scenarios. Moreover, the UIAD methods are limited in their effectiveness in detecting unseen anomalies, making it difficult to accurately identify novel anomalies with small differences from seen normal samples [Wang *et al.*, 2025b; Wang *et al.*, 2025a]. The zero-shot industrial anomaly detection methods are dedicated to achieve anomaly detection without training data on objects of the target category.

Inspired by the strong generalization ability of VLM on various downstream tasks [Sun *et al.*, 2024], Vision-Language Models (VLMs) have achieved growing interest in ZSIAD. WinCLIP [Jeong *et al.*, 2023] proposed the first VLM-based method for ZSIAD, combining state words and text prompt templates to achieve zero-shot anomaly classification and segmentation. It extracted and aggregated multi-scale spatial features aligned with language for efficient feature alignment [Huang *et al.*, 2024; Huang *et al.*, 2025].

Although these VLM-based ZSIAD methods show great promise, it is difficult to further improve the performance due to the lack of target-specific prompts for unseen anomalies. Specifically, we consider the following two perspectives as critical limitations to the accurate classification and localization of unseen anomalies by recent VLM-based ZSIAD methods: ***1)*** **Unpredictability and Complexity of Unseen Anomalies.** Industrial anomalies in the real world tend to be highly uncertain and diverse, which makes the model rely on additional prompts for the accurate identification and localization of unseen anomalies, as shown in Figure 1 (a). AdaCLIP [Cao *et al.*, 2025] introduced learnable hybrid prompts that allow the model to better recognize unseen anomalies, which consist of text prompt templates and learnable dynamic prompts. This dynamic prompt is usually generated in real time by the image encoder and linear projection layer for each test image to enhance the model's adaptability to different kinds of anomalies. However, the learnable prompts are difficult to capture the details of unseen anomalies, leading to models that may not be able to accurately identify and localize unseen anomalies. It naturally brings the second issue: ***2)*** **High annotation costs for detailed anomaly prompts.** Whether it is during training or inference, integrating detailed prompts can effectively improve the model's ability to identify and localize unseen anomalies. In practice, obtaining high-quality anomaly descriptions often faces numerous challenges, such as the scarcity of anomalous samples, the imbalance of categories, and the complexity of the annotation process. These factors lead to a significant increase in the cost of labeling anomaly data. AnomalyGPT [Gu *et al.*, 2024b] provides the model with more adaptability and accuracy in industrial anomaly detection tasks by generating the simulated anomalous regions and anomaly descriptions, as shown in Figure 1 (b). Specifically, it generates simulated anomalous regions on normal samples by image editing and generates detailed text descriptions for them as prompts, which alleviates the problem of high annotation cost to some extent. However, these methods rely on simulated anomalies, which may lead to poor generalization ability to industrial anomalies in real applications. Therefore, how to obtain reliable and detailed prompts in real-world datasets at low cost has become an urgent problem. In addition, how to fully utilize these prompts for industrial anomaly detection is also a worthwhile and challenging problem.

Facing the above issues, we attempt to address them from the following two perspectives, as shown in Figure 1 (c): ***(1)*** proposing a multi-stage prompt generation agent based on MLLM; ***(2)*** proposing an explainable prompt enhancement framework that can fully utilize the generated prompts for facilitating the understanding and detection of unseen industrial anomalies in VLM-based methods. Specifically, we first propose a multi-stage prompt generation agent for ZSIAD which can obtain semantic-rich detailed text prompts and attention prompts from unseen industrial samples. On the one hand, we compare the test samples with normal samples to obtain preliminary detailed anomaly information. To address the intrinsic illusion problem of MLLM, we obtain accurate and semantic-rich anomaly descriptions through further refinement and anti-false alarm constraints. On the other hand, we achieve anomaly-related attentional embedding through the Vision Foundation Model (VFM)-based Anomaly Feature Selection (AFS) module, which can facilitate the model to focus on potential anomalous regions in the foreground. Building on this foundation, we propose the Hybrid Explainable Prompts Enhancement (HEPE) framework, which jointly utilizes generated detailed text and attention prompts for further enhancing the performance of VLM-based methods in ZSIAD. The synergistic effect of hybrid explainable prompts allows maximizing the utilization of these prompts.

Therefore, our contribution can be summarized as follows.

1) We propose a multi-stage prompt generation agent. It can leverage MLLM and VFM to generate semantic-rich text prompts and attention prompts for unseen anomalies, thus providing specific anomaly prompts for VLMs at low cost.

2) We propose a hybrid explainable prompt enhancement framework to further improve the performance by jointly utilizing explainable generated text prompts and attention prompts in ZSIAD, which can enhance the model's understanding and detection of unseen anomalies.

3) We have conducted extensive experiments on seven industrial anomaly detection datasets. The experimental results show that our method has significant advantages over the recent state-of-the-art ZSIAD method.

2 Relative Work

2.1 Industrial Anomaly Detection Methods

The goal of industrial anomaly detection is to accurately identify and localize abnormal patterns or defects that deviate from normal operating conditions during the production process, thereby improving production efficiency, reducing losses, and ensuring system stability. According to the type of supervision, it can be divided into three categories: semi-supervised methods, unsupervised methods and zero-sample

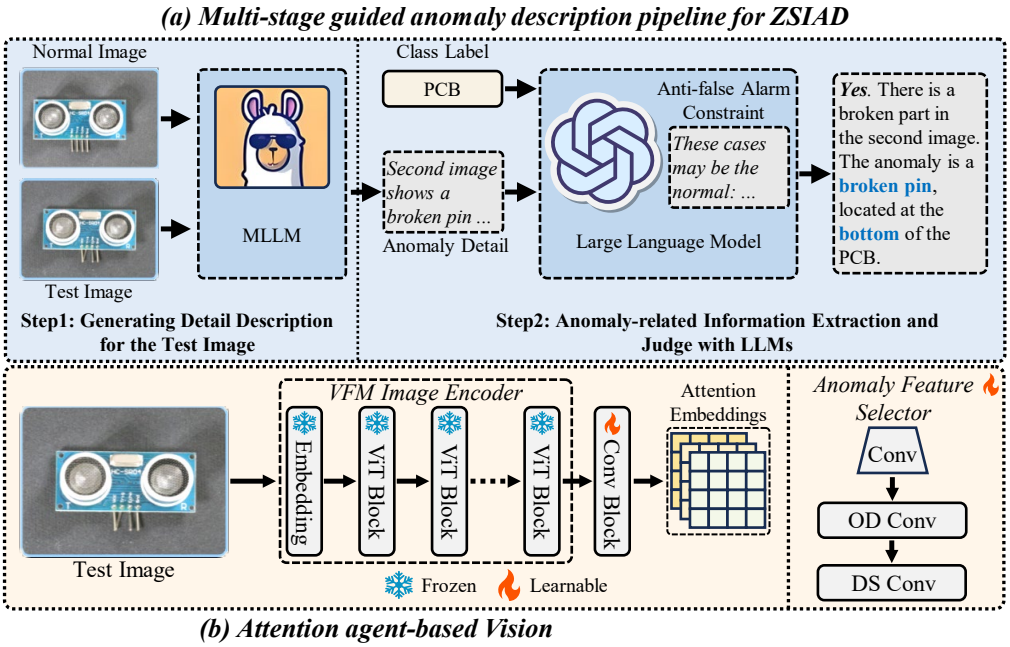

Figure 2: The proposed Multi-stage guided anomaly description pipeline and Attention agent-based VFM.

methods. Unsupervised IAD methods [Carratù *et al.*, 2023; Wang *et al.*, 2024a] detect industrial anomalies that deviate from the normal pattern mainly by learning the feature distribution of normal samples. The advantage of unsupervised methods is that they do not require anomalous samples for training and are suitable for industrial scenarios with scarce anomalous data. Semi-supervised Industrial Anomaly Detection (IAD) methods [Dong *et al.*, 2024] leverage a small number of labeled abnormal samples and a large number of unlabeled normal samples for training, thereby enhancing the model's ability to identify anomalies while reducing the dependence on large-scale labeled data.

Although these methods have achieved a promising performance, they often lack generalization ability when detecting unseen industrial anomalies. Therefore, an increasing number of studies [Zhu *et al.*, 2024; Gu *et al.*, 2024a] are focusing on zero-shot IAD methods. Since the pre-trained vision-language models (VLMs) naturally have strong generalization ability to various types of targets in the real world, they show significant advantages in ZSIAD. By leveraging vision-language alignment and transferring knowledge from pre-trained models, VLM-based ZSIAD methods [Chen *et al.*, 2024] can better understand and identify unseen industrial anomalies in real-world scenarios. Existing VLM-based methods typically guide anomaly detection through learnable prompts or manually designed text prompts. For instance, WinCLIP [Jeong *et al.*, 2023] guides anomaly detection by constructing text descriptions specifically tailored for the "normal" and "abnormal" categories. However, learnable prompts have limitations in capturing the fine-grained details of unseen anomalies. These prompts usually rely on limited labeled data and predefined text templates, making it difficult to accommodate the diversity and complexity of the real-world anomalies. Therefore, how to further improve the adaptability and detection accuracy of VLM-based models for unseen anomalies remains a challenge to be solved.

In this paper, we provide an innovative solution for ZSIAD through the proposed multi-stage prompt generation agent. The agent combines the powerful generalization capabilities of the pre-trained MLLM with the multi-stage prompt strategy to generate semantic-rich text prompts and attention prompts from unseen industrial anomalies.

2.2 Prompt Learning

Prompt learning is a technique for augmenting model inputs with task-specific prompts, thereby adapting pre-trained large models to downstream tasks. Specifically, model inputs can be augmented by well-designed prompts, which mainly include soft and hard prompts. Soft prompts [Jia *et al.*, 2022; Ren *et al.*, 2023] are learnable continuous vectors that are usually optimized for a specific dataset by gradient optimization methods. However, soft prompts have the disadvantage of lacking explainability and tend to perform worse than high-quality hard prompts [Wang *et al.*, 2024c]. Hard prompts are natural language instructions [Dong *et al.*, 2022], which have the advantage of being flexible and explainable but designing high-quality hard prompts requires a lot of manual effort.

Recently, several studies [Deng *et al.*, 2024; Qu *et al.*, 2024] have explored the application of prompts to VLM-based IAD methods. WinCLIP [Jeong *et al.*, 2023] have achieved promising performance using learnable soft prompts in IAD. In addition, some studies [Cao *et al.*, 2025; Zhou *et al.*, 2024] have jointly used both soft and hard prompts, improving the performance of ZSIAD by combining the advantages of both. However, further performance improvement is limited by the lack of high-quality hard prompts. In this

paper, we generate high-quality text prompts using the multi-stage prompt generation agent. This agent allows the model to better capture anomalous features while maintaining the model's ability to generalize to unseen anomalies.

3 Method

In this section, we will introduce the proposed hybrid explainable prompt enhancement method and its important components in detail, which mainly include (1) Overview; (2) Multi-stage Prompt Generation Agent; (3) Hybrid Explainable Prompt Enhancement Framework.

3.1 Overview

As shown in Figure 2, we first use the multi-stage prompt generation agent to generate semantic-rich text and attention prompts for unseen anomalies, thereby providing specific anomaly prompts for the model at low cost. For the text prompts, we input normal and test samples into the elaborated MLLM in the first step so that it describes in detail the differences between the test samples and the normal samples. To further obtain anomaly-related information as well as to prevent potential false alarm possibility, we refine the generated detailed anomaly descriptions under the anti-false alarm constraint in the second step to obtain accurate and anomaly-related text descriptions. In addition, to make the model pay more attention to the anomaly information in the foreground, we obtain anomaly-related attention prompts through the VFM-based anomaly feature selection module.

As shown in Figure 3, the proposed hybrid explainable prompts enhancement framework mainly consists of the visual branch, textual branch, and attention prompts branch. The proposed CLIP-based framework primarily explores the utilization of text and attention prompts for enhancing the model's ability to detect unseen anomalies. In the visual branch, we input the test industrial images into the pre-trained CLIP [Radford *et al.*, 2021] image encoder to extract patch-level visual embeddings. In the textual branch, we input the generated text prompts representing normal and abnormal respectively into the pre-trained CLIP text encoder to obtain text embeddings. Combining the dynamic and static prompt generators from previous work [Cao *et al.*, 2025], we obtain anomaly mapping by calculating the similarity between visual embeddings and textual embeddings. In addition, we combine the hybrid semantic fusion module (HSF) and generated attentional embeddings for facilitating the model to focus on anomalous regions in the foreground in order to enhance the accuracy and generalization capability of image-level anomaly detection.

3.2 Multi-Stage Prompt Generation Agent

For VLM-based ZSIAD methods, high-quality text descriptions of anomalies are essential to prompt the model for understanding and detection of unseen anomalies. However, existing methods either rely on predefined text templates or are annotation-costly. In addition, attention prompts can enhance the model's ability to focus on anomalous regions. Therefore, we propose the multi-stage prompt generation agent for generating semantic-rich text and attention prompts.

Text Prompt. To fully leverage the potential of MLLM in describing anomalies, we first input normal and test sample to MLLM in the first step, and the differences between the two samples are described in detail through MLLM in order to mine potential anomaly details. Subsequently, we bootstrap the process of refining anomaly details by introducing category labels in the second step, which aims at generating anomaly descriptions that are as relevant as possible to specific industrial products. It is noteworthy that LLMs are prone to misidentifying normal differences as abnormal regions due to the potential problems of illusion and misinterpretation. For instance, LLMs may misclassify nuts that are oriented differently from the normal samples as anomalies. Therefore, we further set the anti-false alarm constraint, which is mainly used to prevent LLMs from describing normal differences as abnormal.

Attention Prompt. To enhance the model's ability to focus on anomalous regions, we utilize the VFM and anomalous feature selection module to generate attentional prompts. Specifically, we first input the test industrial images into the VFM image encoder for attentional feature extraction. Then, the anomaly feature selection module can select the most useful regions for anomaly detection as attention prompts from the attentional features. which mainly contains Omni-dimensional Dynamic Convolution (OD Conv) [Li *et al.*, 2022] and Depthwise Separable Convolutions (DS Conv) [Chollet, 2017]. OD Conv significantly improves the dynamics of convolutional operations and feature extraction through the multidimensional attention mechanism that can dynamically enhance attention to the anomalous regions. DS Conv can achieve higher performance with a lower parameters through the property of deeply separable convolution.

3.3 Hybrid Explainable Prompt Enhancement

Due to its outstanding zero-shot generalization capability, the vision-language classic framework CLIP has achieved tremendous success across various domains of computer vision. Building upon prior CLIP-based ZSIAD methods [Cao *et al.*, 2025], we propose a Hybrid Explainable Prompts Enhancement that combines generated text descriptions and attention prompts for industrial anomaly detection. The following are each component of the proposed framework.

Static and Dynamic Prompts. To improve the ZSIAD performance effectively using the proposed hybrid explainable prompts, we introduce static and dynamic prompts. The static prompt $\mathbf{P}_S$ serves as a foundational learning cues shared across all images. Furthermore, we introduce dynamic prompts $\mathbf{P}_D$, allowing the semantic information contained in the hybrid prompts to be effectively integrated into the CLIP semantic space. $\mathbf{P}_D$ differs from $\mathbf{P}_S$ in that they are generated by the prompt generator $\mathcal{G}$ for each test image individually. In the proposed method, the composition of $\mathcal{G}$ includes class tokens from the CLIP image encoder and a learnable projection layer. This projection layer is used to map class tokens into dynamic prompts $\mathbf{P}_D$.

Hybrid Semantic Fusion Module (HSF). Previous anomaly detection methods [Chen *et al.*, 2023] select the maximum value of anomalies as the anomaly score, which

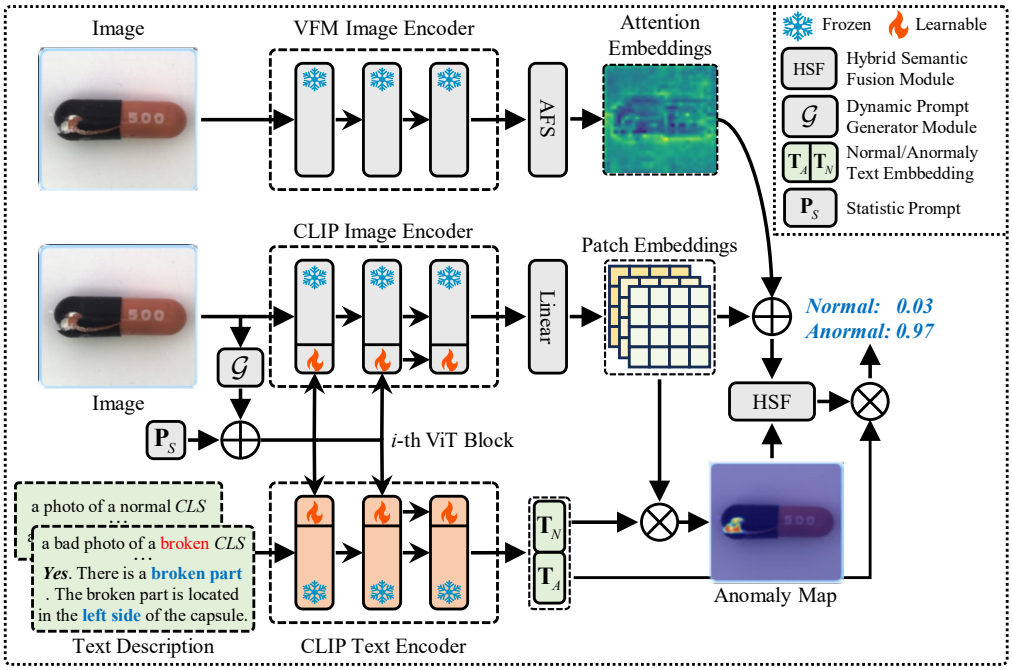

Figure 3: The proposed explainable prompt enhancement framework for ZSIAD, which can fully utilize the generated prompts for facilitating the understanding and detection of unseen industrial anomalies.

is sensitive to noise predictions. To enhance the robustness of the model and the reliability of the detection results, we introduce the HSF module [Cao *et al.*, 2025]. It is adopted to aggregate patch embeddings that are more likely to represent anomalies, which effectively consolidate region-level information for robust image-level anomaly detection. The HSF output integrates the semantic information most relevant to anomalies. Combined with the proposed hybrid explainable prompts, it can incorporate even more anomaly-related semantic information. Compared to the previous max-value-based anomaly detection methods, the resulting semantically enriched vision patch embeddings effectively improve image-level anomaly detection performance.

4 Loss Function

The pixel-level anomaly map in the proposed method is derived by calculating the cosine similarity between patch embeddings $\mathbf{Q}_E$ with integrated attention prompts and text embeddings $\mathbf{T}_A$ and $\mathbf{T}_N$ with integrated text prompts. The anomaly map is defined as follows:

$$\mathbf{M}_i = \psi\left(\frac{e^{cos(\mathbf{Q}_E^i, \mathbf{T}_A)}}{e^{cos(\mathbf{Q}_E^i, \mathbf{T}_A)} + e^{cos(\mathbf{Q}_E^i, \mathbf{T}_N)}}\right), \tag{1}$$

where ψ is an interpolate function. The pixel-level anomaly map $\mathbf{M}_i$ in i-th block is resized to match the dimensions of the test image. Anomaly maps are extracted from multiple blocks [24] and aggregated to produce the final anomaly map $\mathbf{M}$. In the training process, dice loss $\mathcal{L}_D$ and focal loss $\mathcal{L}_D$ is jointly applied as the object function, which is defined as follows:

$$\mathcal{L}_D = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i}, \tag{2}$$

$$\mathcal{L}_F = -\alpha(1-p_i)^{\gamma} log(p_i), \tag{3}$$

where α is the balancing factor, γ is the hyper-parameter that adjusts the weight of easily classified samples, p_i is the anomaly scores predicted by the model, and g_i is the ground truth label. The final object function $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_F. \tag{4}$$

5 Experiments

5.1 Experiment Setup

Datasets. We conduct experiments using the seven industrial anomaly detection datasets for all experiments: MVTec AD [Bergmann *et al.*, 2021], VisA [Zou *et al.*, 2022], MPDD [Jezek *et al.*, 2021], BTAD [Mishra *et al.*, 2021], KSDD [Tabernik *et al.*, 2020], DAGM [Wieler and Hahn, 2007], and DTD-Synthetic [Aota *et al.*, 2023].

Evaluation Metrics. Consistent with prior ZSIAD methods [Cao *et al.*, 2025], we adopted Area Under the Receiver Operating Characteristic Curve (AUROC) and maximum F1 score (max-F1) to evaluate the anomaly detection performance of both image-level and pixel-level. Beyond the dataset-specific results, we further reported the mean performance across all datasets. This result is quantified using AUROC and max-F1.

Implementation Details. We adopted QWen2-VL-72B [Wang *et al.*, 2024b] to generate detailed descriptions of the anomalies. Furthermore, QWen2.5-7B [Yang *et al.*, 2024] is

Metric	Dataset	w/o supervised training		w/ supervised training				
		SAA	WinCLIP	DINOv2	SAM	APRIL-GAN	AdaCLIP	Our
Image-level (AUROC, max-F1)	MVTec AD	(63.5, 87.4)	(91.8, **92.9**)†	(74.4, 87.4)	(70.8, 86.0)	(82.3, 88.9)	(89.2, 90.6)	(**91.9**, 92.4)
	VisA	(67.1, 75.9)	(78.1, 80.7)†	(75.2, 78.5)	(61.9, 73.9)	(81.7, 80.7)	(85.8, 83.1)	(**86.3**, **84.1**)
	MPDD	(42.7, 73.9)	(61.4, 77.5)	(62.4, 74.9)	(63.0, 77.0)	(66.0, 76.0)	(76.0, **82.5**)	(**83.4**, 82.4)
	BTAD	(59.0, 89.7)	(68.2, 67.6)	(79.3, 69.3)	(89.4, 85.7)	(85.2, 82.0)	(88.6, 88.2)	(**95.2**, **90.9**)
	KSDD	(68.6, 37.6)	(93.3, 79.0)	(94.9, 77.5)	(65.8, 37.9)	(95.7, 85.2)	(97.1, 90.7)	(**98.9**, **91.4**)
	DAGM	(87.1, 88.8)	(91.7, 87.6)	(90.7, 89.2)	(82.7, 83.6)	(93.5, 91.8)	(**99.1**, **97.5**)	(98.4, 96.5)
	DTD-Synthetic	(94.4, 93.5)	(95.1, 94.1)	(85.8, 93.5)	(81.9, 91.1)	(98.1, 96.8)	(95.5, 94.7)	(**96.5**, **95.9**)
	Average	(68.9, 78.1)	(82.8, 82.8)	(80.4, 81.5)	(73.6, 76.4)	(86.1, 85.9)	(90.2, 89.6)	(**92.9**, **90.5**)
	Rank	(6.3, 5.3)	(4.4, 4.0)	(5.0, 5.1)	(5.7, 6.0)	(3.0, 3.4)	(2.3, 2.1)	(**1.3**, **1.6**)
Pixel-level (AUROC, max-F1)	MVTec AD	(75.5, 38.1)	(85.1, 31.6)†	(85.9, 39.6)	(85.4, 29.4)	(83.7, 39.8)	(88.7, 43.4)	(**90.9**, **45.9**)
	VisA	(76.5, 31.6)	(79.6, 14.8)†	(95.0, 30.3)	(92.6, 18.2)	(95.2, 32.3)	(**95.5**, **37.7**)	(94.9, 35.2)
	MPDD	(81.7, 18.9)	(71.2, 15.4)	(95.6, 31.1)	(94.8, 22.1)	(95.1, 30.6)	(**96.1**, 34.9)	(96.0, **35.5**)
	BTAD	(65.8, 14.8)	(72.6, 18.5)	(91.9, 43.4)	(**93.8**, 46.9)	(89.5, 38.4)	(92.1, 51.7)	(93.1, **52.1**)
	KSDD	(78.8, 6.6)	(95.8, 21.3)	(**99.3**, 50.6)	(91.2, 18.4)	(98.2, **56.2**)	(97.7, 54.5)	(98.2, 54.2)
	DAGM	(62.7, 32.6)	(81.3, 13.9)	(90.9, 52.0)	(88.6, 40.7)	(90.3, 57.9)	(91.5, 57.5)	(**92.4**, **63.3**)
	DTD-Synthetic	(76.7, 60.6)	(79.5, 16.1)	(97.0, 63.4)	(95.0, 56.7)	(97.8, 72.7)	(97.9, 71.6)	(**98.6**, **72.8**)
	Average	(73.9, 29.0)	(80.7, 18.8)	(93.7, 44.3)	(91.7, 33.2)	(92.8, 46.9)	(94.2, 50.2)	(**95.1**, **51.0**)
	Rank	(6.9, 5.7)	(5.9, 6.4)	(3.0, 4.0)	(4.4, 5.4)	(3.6, 2.9)	(2.3, 2.1)	(**1.9**, **1.4**)

Table 1: Comparisons of ZSIAD methods in the 7 industrial anomaly datasets. The best performance is in bold and the second best is underlined. † denotes to results taken from original papers. Rank denotes the average performance rankings on all datasets.

Text Prompt	Attention Prompt		Industrial Dataset			
	VFM	AFS	Image-level		Pixel-level	
×	×	×	91.4	89.3	93.7	49.8
✓	×	×	91.7	90.3	94.5	50.5
✓	✓	×	92.3	89.4	94.8	50.8
✓	✓	✓	**92.9**	**90.5**	**95.1**	**51.0**

Table 2: Ablation study of each proposed strategy.

Anti-false Alarm Constraint	Normal Image	Industrial Dataset			
		Image-level		Pixel-level	
×	×	91.7	89.6	93.9	50.0
✓	×	92.1	89.9	94.2	50.2
×	✓	92.6	90.3	94.7	50.8
✓	✓	**92.9**	**90.5**	**95.1**	**51.0**

Table 3: Ablation study of different text prompt generation strategy.

utilized to extract anomalous information and judge the presence of anomalies. The pre-trained CLIP (ViT-L/14@336px) [Radford *et al.*, 2021] is employed as the backbone for subsequent ZSIAD models, extracting patch embeddings from the 6th, 12th, 18th, and 24th ViT blocks. DINOv2 (ViT-S) [Oquab *et al.*, 2024] is adopted as the VFM. We trained the proposed method for 5 epochs with a learning rate of 0.01. All experiments were performed with a single NVIDIA A100 GPU (80GB).

Comparison Methods. We compare the proposed method with two baselines: with and without training. Specifically, SAA [Cao *et al.*, 2023] and WinCLIP [Jeong *et al.*, 2023] are adopted as our non-trained baselines. For trained baselines, we selected APRIL-GAN [Chen *et al.*, 2023] and AdaCLIP [Cao *et al.*, 2025]. Following the previous work, we trained DINOv2 [Oquab *et al.*, 2024] and SAM [Kirillov *et al.*, 2023], which appends additional linear layers to multiple Transformer layers as a task head for anomaly detection.

5.2 Comparison with State-of-the-Art Methods

Table 1 presents the results of our comparative experiments, demonstrating the superior performance of the trained methods over the untrained ZSIAD baselines. WinCLIP and SAA with the manual text prompts exhibit inferior performance. It is noteworthy that fine-tuning the visual foundation models (DINOv2 and SAM) can achieve satisfactory pixel-level ZSIAD. The impressive performance of VFM-based ZSIAD methods fine-tuned on domain-specific data underscores the fact that pre-trained VFMs have essential knowledge for the anomaly. With minor adaptations, this foundational knowledge can be effectively leveraged for ZSIAD tasks.

Compared to other ZSIAD methods, the proposed method exhibits notable performance enhancements. For example, compared to the sub-optimal method AdaCLIP, the proposed method achieves 2.7% improvement in image-level AUROC and 0.9% improvement in max-F1. Moreover, the pixel-level AUROC reached 95.1% and the max-F1 score attained 51.0%. These results underscore the superiority of the proposed method and validate the efficacy of incorporating text-enhanced and attention-based prompts. Furthermore, we visualize the anomaly detection results across different datasets in Figure 4. The proposed method shows notably more precise pixel-level anomaly detection for unseen classes than other methods. This best pixel-level performance can be attributed to the detailed anomaly information provided by text-enhanced prompts and the more discriminative visual features captured by attention prompts.

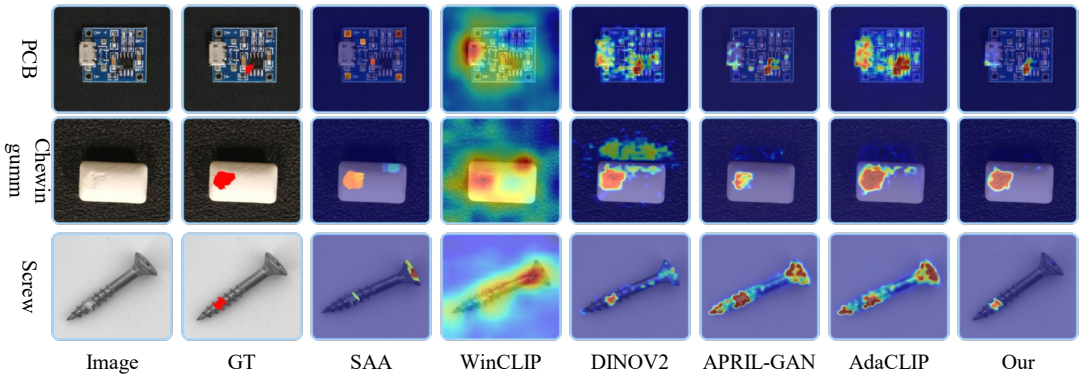

Figure 4: Visualization of anomaly maps of different ZSIAD methods. The proposed method can get the most precise segmentation results for novel categories.

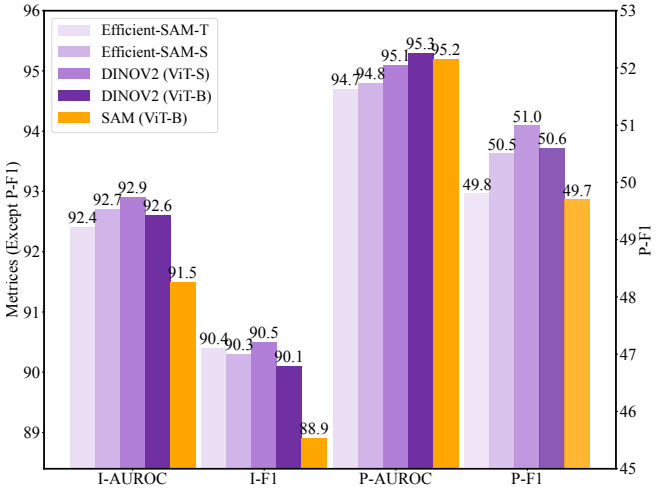

Figure 5: Ablation study of different VFM for attention prompts generation. 'I' and 'P' means the Image-level and Pixel-level, respectively.

5.3 Ablation Study

In this subsection, we conduct ablation experiments on the all datasets to validate the effectiveness of different strategies. We quantitatively evaluated the effectiveness of the proposed text and attention prompts by different combinations. The results of the ablation study are shown in Table 2. Compared to the baseline, each combination achieves a better performance. Therefore, we can conclude that both strategies contribute to the ZSIAD performance improvement. In addition, the final ablation study results demonstrate that the joint application of the proposed strategies can further result in greater performance gain, proving the effectiveness and rationality of the proposed strategies.

5.4 Qualitative Analysis

Text Prompt Generation Strategy. Table 3 presents the results of ablation experiments on various text prompt generation strategies. The experimental results clearly demonstrate that the model achieves optimal performance by contrasting anti-false alarm with normal samples in generating text prompts. In other words, the joint application of the proposed strategies mitigates the noise in generated text prompts. On the one hand, LLM can correct descriptions that mistakenly judge differences in location as errors by applying the anti-false positive strategy. It allows subsequent ZSIAD models to obtain accurate and semantically rich descriptions of anomalies. On the other hand, MLLM has a clear reference to exclude misjudgments caused by the specific attributes of the objects themselves (such as describing sesame-like spots on sugar as anomalies), thereby outputting accurate anomaly information. Consequently, the combined application of these strategies minimized noise in text prompt generation, resulting in optimal model performance.

Different VFM for Attention Prompt. We explored the impact of different VFM-generated attention cues on the proposed ZSIAD method, with results reported in Figure 5. The experimental results clearly shows that using DINOv2 (ViT-S) to generate attention prompts achieves the optimal ZSIAD performance for the proposed method. The pixel-level performance achieved with SAM and its variants EfficientSAM [Xiong *et al.*, 2024] remains suboptimal. We believe this is due to DINOv2 being a self-supervised VFM, aimed at learning general-purpose visual features. In other words, DINOv2 extracts discriminative and semantic-rich features from extensive image datasets through a self-supervised training method. These feature representations are essential for ZSIAD tasks. SAM and its derivatives are primarily designed for interactive segmentation, demonstrating strong performance guided by user-provided prompts. SAM exhibits a weaker overall semantic understanding of complex scenes in scenarios without clear prompts.

6 Conclusion

In this paper, we present a novel multi-stage prompt generation agent for ZSIAD to generate semantic-rich text prompts and attention prompts for unseen anomalies. Furthermore, we propose the Hybrid Explainable Prompt Enhancement method, which maximizes the potential of these prompts to enhance the model's understanding and detection of unseen anomalies. Extensive experiments on the widely-utilized benchmarks have validated the effectiveness of our method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62371404, in part by the National Natural Science Foundation of China under Grant 62301621, in part by Shenzhen Science and Technology Program (No. 20231121172359002), in part by Shenzhen General Research Project (No. JCYJ20241202125904007), and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011398).

References

[Aota *et al.*, 2023] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *WACV*, pages 5564–5572, 2023.

[Bergmann *et al.*, 2021] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *IJCV*, 129(4):1038–1059, 2021.

[Cao *et al.*, 2023] 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*, 2023.

[Cao *et al.*, 2025] 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*, pages 55–72. Springer, 2025.

[Carratù *et al.*, 2023] Marco Carratù, Vincenzo Gallo, Salvatore Dello Iacono, Paolo Sommella, Alessandro Bartolini, Francesco Grasso, Lorenzo Ciani, and Gabriele Patrizi. A novel methodology for unsupervised anomaly detection in industrial electrical systems. *IEEE TIM*, 2023.

[Chen *et al.*, 2023] Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/fewshot 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*, 2(4), 2023.

[Chen *et al.*, 2024] Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. In *IJCAI*, pages 17–33. Springer, 2024.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017.

[Deng *et al.*, 2024] Chenghao Deng, Haote Xu, Xiaolu Chen, Haodi Xu, Xiaotong Tu, Xinghao Ding, and Yue Huang. Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In *ACM MM*, pages 1761–1770, 2024.

[Dong *et al.*, 2022] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[Dong *et al.*, 2024] Hao Dong, Gaëtan Frusque, Yue Zhao, Eleni Chatzi, and Olga Fink. Nng-mix: Improving semi-supervised anomaly detection with pseudo-anomaly generation. *IEEE TNNLS*, 2024.

[Gu *et al.*, 2024a] 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 *ACM MM*, pages 2041–2049, 2024.

[Gu *et al.*, 2024b] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*, volume 38, pages 1932–1940, 2024.

[Huang *et al.*, 2021a] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE TII*, 18(8):5171–5179, 2021.

[Huang *et al.*, 2021b] Chao Huang, Zehua Yang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, and Yaowei Wang. Self-supervision-augmented deep autoencoder for unsupervised visual anomaly detection. *IEEE TCYB*, 52(12):13834–13847, 2021.

[Huang *et al.*, 2022a] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE TCYB*, 54(5):3197–3210, 2022.

[Huang *et al.*, 2022b] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE TNNLS*, 34(11):9389–9403, 2022.

[Huang *et al.*, 2024] Chao Huang, Weichao Cai, Qiuping Jiang, and Zhihua Wang. Multimodal representation distribution learning for medical image segmentation. In *IJCAI*, pages 4156–4164, 2024.

[Huang *et al.*, 2025] Chao Huang, Weiliang Huang, Qiuping Jiang, Wei Wang, Jie Wen, and Bob Zhang. Multimodal evidential learning for open-world weakly-supervised video anomaly detection. *IEEE TMM*, 2025.

[Jeong *et al.*, 2023] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pages 19606–19616, 2023.

[Jezek *et al.*, 2021] 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*, pages 66–71. IEEE, 2021.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022.

[Kirillov *et al.*, 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*, pages 4015–4026, 2023.

[Li *et al.*, 2022] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. In *ICLR*, 2022.

[Mishra *et al.*, 2021] 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*, pages 01–06. IEEE, 2021.

[Oquab *et al.*, 2024] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[Qu *et al.*, 2024] 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*, pages 301–317, 2024.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

[Ren *et al.*, 2023] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alexander J Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. *NeurIPS*, 36:12569–12588, 2023.

[Sun *et al.*, 2024] Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.13836*, 2024.

[Tabernik *et al.*, 2020] 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.

[Wang *et al.*, 2024a] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, pages 22883–22892, 2024.

[Wang *et al.*, 2024b] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[Wang *et al.*, 2024c] Wei Wang, Hanyang Li, Ke Shi, Chao Huang, Yang Cao, Cong Wang, and Xiaochun Cao. Optimal graph learning and nuclear norm maximization for deep cross-domain robust label propagation. In *IJCAI*, pages 1407–1415, 2024.

[Wang *et al.*, 2025a] Wei Wang, Hanyang Li, Cong Wang, Chao Huang, Zhengming Ding, Feiping Nie, and Xiaochun Cao. Deep label propagation with nuclear norm maximization for visual domain adaptation. *IEEE TIP*, 2025.

[Wang *et al.*, 2025b] Wei Wang, Mengzhu Wang, Chao Huang, Cong Wang, Jie Mu, Feiping Nie, and Xiaochun Cao. Optimal graph learning based label propagation for cross-domain image classification. *IEEE TIP*, 2025.

[Wieler and Hahn, 2007] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, volume 6, page 11, 2007.

[Xiong *et al.*, 2024] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *CVPR*, pages 16111–16121, 2024.

[Yang *et al.*, 2024] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

[Zhu *et al.*, 2024] 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 *ACM MM*, pages 48–57, 2024.

[Zou *et al.*, 2022] 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*, pages 392–408. Springer, 2022.