

Defect-Aware Hybrid Prompt Optimization for Zero-Shot Multi-Type Anomaly Detection and Segmentation

Nadeem Nazer^{1,2*}, Hongkuan Zhou^{1,3*†}, Lavdim Halilaj¹, Ylli Sadikaj⁴, and Steffen Staab^{3,5}

¹ Corporate Research, Robert Bosch GmbH, Renningen, Germany

[†] Hongkuan.Zhou@de.bosch.com

² Otto-von-Guericke-University Magdeburg, Magdeburg, Germany

³ Institute for Artificial Intelligence, University of Stuttgart, Stuttgart, Germany

⁴ Faculty of Computer Science, UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria

⁵ University of Southampton, Southampton, UK

Abstract. Recent vision-language models (VLMs) like CLIP have shown impressive anomaly detection performance under significant distribution shift by utilizing high-level semantic information through text prompts. However, these models often overlook fine-grained defect cues, e.g., hole, cut, or scratch, that are essential for understanding the anomaly’s nature. Moreover, the modality gap between images and text can lead to subtle visual evidence being poorly captured in textual descriptions. To address the gap, we enhance the representation of “abnormal” with structured semantics, bridging coarse anomaly signals and fine-grained defect categories. We propose a hybrid prompting mechanism that combines human-readable descriptions of defect types with learnable token embeddings. Building on these ideas, we introduce DAPO, a **Defect-aware Prompt Optimization** framework for zero-shot multi-type and binary anomaly detection and segmentation under distribution shift. DAPO aligns anomaly-relevant visual features with their corresponding textual semantics by learning hybrid defect-aware prompts that combine fixed textual anchors with trainable token embeddings. We conducted experiments on public benchmarks (MPDD, VisA, MVTec-AD, MAD, and Real-IAD) and an internal dataset. The results suggest that compared to the baseline models, DAPO achieves a 3.6% average improvement in AUROC and average precision metrics at the image level under distribution shift, and a 5.2% average improvement in AUROC and F1 when localizing novel anomaly types under zero-shot settings.

1 Introduction

Visual inspection is a crucial task in manufacturing, ensuring that the products are free of defects before reaching customers. Manual efforts in such inspections

* These authors contributed equally.

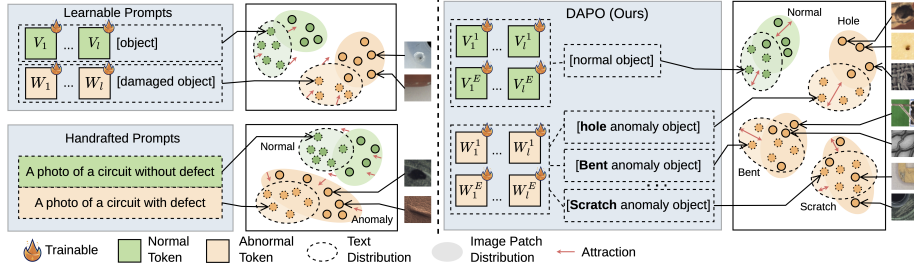


Fig. 1: Motivation. Existing CLIP-based prompting often yields coarse normal–abnormal separation, where different defect types collapse into overlapping anomaly representations. DAPO uses compositional defect-aware prompts to align visual cues with defect semantics, producing better-separated representations across defect types while avoiding manual prompt enumeration. In particular, the shared learnable context tokens capture a type-agnostic abnormal context that transfers across defect categories, while the defect-name anchors provide type-specific semantics.

incur high labor costs and attention. With advancements in deep learning and computer vision techniques, these complex processes are increasingly being automated, reducing manual involvement and enabling high-precision identification of defects. Therefore, visual inspection through anomaly detection (AD) and segmentation is essential in identifying and localizing defective regions.

Most conventional approaches adopted one-class or unsupervised anomaly detection [5, 7, 22, 28, 39], where they are trained solely on the volumes of normal (non-defective) images and classify any deviation from the learned distribution as an anomaly. Although such models perform well on their training distribution, the domain shift in industrial settings is large, so that these models often need to be trained separately on different domains and frequently retrained. Consequently, recent work increasingly leverages large-scale pre-trained VLMs (e.g., CLIP [21]), using their broad priors as an adaptable foundation that mitigates data scarcity and reduces extensive domain-specific annotation.

Despite their potential, existing approaches remain limited to coarse-grained anomaly detection, i.e., identifying whether an image is abnormal without differentiating defect types. This limitation matters in practice for two reasons. 1) type-level semantics provide an inductive bias that aligns local features with defect-specific prototypes, improving robustness under distribution shift [23]; 2) focusing on the nature of anomalies such as stains, holes, or missing components is essential for industrial operators to trace root causes and implement preventive measures. Prior work, such as MultiADS [23], attempts to incorporate defect-type knowledge via a curated knowledge base and hand-crafted prompts, but this process is time-consuming and difficult to scale as new defect categories emerge. Prompt learning offers an appealing alternative and has been successfully used in works like AnomalyCLIP [36]. However, relying solely on learnable (soft) prompt tokens can drift away from the natural-language manifold on which CLIP generalizes well thanks to its large-scale pretraining, thereby weakening zero-shot transfer ability when tested on the target domain.

To overcome these challenges, we propose DAPO, **Defect-aware Prompt Optimization**, where we learn hybrid prompts for each defect type, which consist of a set of shared learnable defect token embeddings together with the literal tokens of specific defect-type descriptions in textual format (e.g., $[W_1], \dots, [W_l]$ [“bent”] [“defect”]). Sharing learnable tokens across defect categories enables prompt construction for novel defect types by reusing the same learned abnormal tokens while simply swapping in the textual description of the new defect. The main contributions of this work are as follows:

- We propose DAPO, a defect-aware prompt optimization framework that composes hybrid prompts from fixed textual anchors and a small set of shared learnable defect tokens reused across defect types.
- We design a prompt composition and optimization scheme that explicitly bridges the coarse anomaly concept and fine-grained defect semantics, improving image-text alignment.
- We conduct comprehensive experiments on five public industrial benchmarks and an internal semiconductor dataset under significant distribution shifts, demonstrating consistent gains on both zero-shot multi-type anomaly segmentation and binary anomaly detection. Our code is publicly available.⁶

2 Related Works

2.1 Unsupervised Unimodal Approaches

Such approaches model normal distributions from training data, treating deviations as anomalies. Reconstruction-based methods [20, 24, 28], leverage architectures of GAN [8], Variational AEs [14], and AEs [16] to find anomalies. Other approaches, such as PatchCore [22], PaDIM [5], and SPADE [4], adopt representational or memory-bank-based strategies by storing normal image features and measuring distances between test samples and those reference features using different distance measurement techniques. STPM [39] and RD4AD [7] utilize a knowledge distillation-based student-teacher framework for AD, using feature matching performed utilizing a pre-trained classification model. Anomalies are detected by measuring feature discrepancies between the student and teacher networks. However, these methods face several limitations, including dependence on training data and a one-class classification focus. These challenges motivated the development of zero-shot approaches.

2.2 Zero-Shot Anomaly Detection

Zero-Shot Anomaly Detection is gaining attention due to advancements in large-scale vision-language models (VLMs), which have wide application in robotics [29, 30, 32] and autonomous driving [31, 37]. CLIP, in particular, is serving as a widely adopted backbone for this task because of its strong zero-shot visual recognition

⁶ <https://github.com/boschresearch/defect-aware-prompt-optimization>

capability. Several approaches emerged, adapting CLIP for AD, such as WinCLIP [11], which uses manually crafted normal and abnormal state prompts with a window-based segmentation approach. AprilGAN [3] introduces additional trainable linear layers to adapt CLIP for AD. Drawing inspiration from prompt learning success in NLP, methods such as CoOp [34] and CoCoOp [33] learn task-specific prompts. AnomalyCLIP [36], replaces manually defined prompts with learnable object-agnostic tokens to form a general text representation for normal and abnormal classes. Similarly, SimCLIP [6] uses implicit prompt tuning to refine text embeddings for better alignment with CLIP’s patch features, and AdaCLIP [2] combines static and dynamic prompts in a hybrid learning framework. More recently, CLIPSAM [17] introduced a novel multi-modal approach by integrating CLIP with the Segment Anything Model (SAM) [15] for enhanced anomaly segmentation and detection capabilities. However, all these methods focus on learning or crafting prompts that can capture only coarse-grained features, failing to identify anomalies with different specific types. FiLo [9] introduces fine-grained anomaly descriptors generated by LLMs and adaptively learned textual templates, replacing general “abnormal” semantics with category-specific descriptions, but it does not learn a shared learnable component reused across defect types. FAPrompt [38] learns compound prompts for fine-grained feature extraction, yet it has no explicit defect-name anchor and is formulated for binary AD rather than explicit defect-type segmentation. Recent work such as LTOAD [27] explores the use of multiple learnable concept-specific prompts, targeting On-line AD by updating prompts at test time to handle long-tailed distributions, whereas Anomaly-OV [26] is a multimodal large language model (MLLM)-based detection-and-reasoning framework. Both are complementary to, but not directly comparable with, our lightweight frozen-CLIP zero-shot setting. Furthermore, a major limitation across the aforementioned prompt-learning designs is their formulation for binary normal-vs-anomalous scoring. While MultiADS [23] represents a representative attempt at multi-defect type detection, it relies heavily on extensive hand-crafted prompts, which are difficult to scale.

Unlike AdaCLIP’s global-vs-dynamic (image-conditioned) prompts, DAPO composes a fixed defect-name anchor with a set of *shared* learnable tokens. Unlike FiLo and FAPrompt, it supports zero-shot instantiation of previously unseen defect types by simply swapping the textual anchor, without retraining or updating an external descriptor set. Our approach mitigates these limitations through hybrid defect-aware prompting, where learnable tokens are composed with fixed textual anchors, and through local alignment objectives that couple defect semantics to fine-grained visual evidence, reducing the discrepancy between visual and textual representations.

3 Preliminaries

We provide the formal definition of three anomaly detection tasks addressed in this work: (1) binary anomaly detection, (2) binary anomaly segmentation under distribution shift, and (3) zero-shot multi-type anomaly segmentation.

Note that the task of binary anomaly detection/segmentation under distribution shift is commonly referred to as *zero-shot* anomaly detection/segmentation in the literature. However, we argue that this term is *misleading*, as the objective is not to locate new types of anomalies, but to distinguish between normal and anomalous samples when applied to previously unseen products, which should be seen as a distribution shift.

Common setup. Let $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, \cdot)\}_{i=1}^{N_1}$ and $\mathcal{D}_{\text{target}} = \{(\mathbf{x}_i, \cdot)\}_{i=N_1+1}^{N_1+N_2}$ be source and target datasets whose images are drawn from different input distributions: $\mathbf{x}_i \sim \mathcal{P}(\mathcal{X})$ for $i \in \{1, \dots, N_1\}$ and $\mathbf{x}_i \sim \mathcal{P}(\mathcal{X}')$ for $i \in \{N_1 + 1, \dots, N_1 + N_2\}$, with $\mathcal{P}(\mathcal{X}) \approx \mathcal{P}(\mathcal{X}')$ and each image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$. A model $\hat{f}(\cdot)$ is trained solely on $\mathcal{D}_{\text{train}}$ and, at test time, adapted to $\hat{f}'(\cdot)$ using auxiliary target-domain metadata, such as the set of relevant defect types. The three tasks differ in their label space and prediction target.

Binary anomaly detection under distribution shift. Each image is labeled with $y \in \{0, 1\}$ indicating anomaly presence. For each $\mathbf{x} \in \mathcal{D}_{\text{target}}$, the detector $\hat{f}'(\cdot)$ classifies it as normal or anomalous.

Binary anomaly segmentation under distribution shift. Each image is paired with a binary mask $\mathbf{y} \in \{0, 1\}^{H \times W}$ indicating the presence of anomalies at each pixel. For each pixel of $\mathbf{x} \in \mathcal{D}_{\text{target}}$, the model $\hat{f}'(\cdot)$ classifies it as normal or anomalous.

Zero-shot multi-type anomaly segmentation. In $\mathcal{D}_{\text{train}}$, each image is paired with a mask $\mathbf{Y} \in \{0, 1\}^{H \times W \times (K_1+1)}$ encoding one normal class and K_1 defect types $\{d_1, \dots, d_{K_1}\}$; in $\mathcal{D}_{\text{target}}$, each image is labeled with a mask over one normal class and a possibly different set of K_2 defect types. Given the target defect types as external knowledge, $\hat{f}'(\cdot)$ predicts, for every $\mathbf{x} \in \mathcal{D}_{\text{target}}$, a mask $\hat{\mathbf{Y}} \in \{0, 1\}^{H \times W \times (K_2+1)}$ that locates anomalies of each defect type.

4 Methodology

We propose DAPO, a CLIP-based framework for anomaly detection and defect-type localization. DAPO learns compositional defect-aware prompts and aligns them with both global image features and patch-level features through global and local objectives (Fig. 2). At test time, prompts for unseen defect descriptors are instantiated by replacing the defect-description anchor $\langle D \rangle$.

4.1 Defect-Aware Prompt Design

We propose a defect-aware learnable prompts, which contain both human-readable words and learnable embeddings that can effectively capture abnormal and

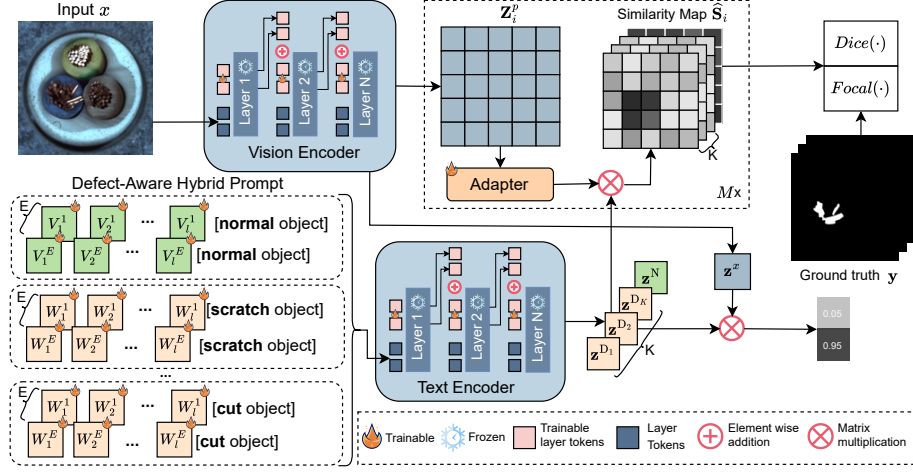


Fig. 2: The training pipeline of DAPO. We learn compositional defect-aware prompts by combining shared learnable context with defect-name anchors, and align them with global and patch-level visual features via global and local objectives. Only prompt tokens and lightweight adapters are updated; CLIP encoders remain frozen.

normal states (cf. Figure 2). Specifically, a prompt for the abnormal state is composed as:

$$\mathbf{P}^D = [W_1][W_2] \dots [W_l] \langle D \rangle [\text{anomaly}][\text{object}], \quad (1)$$

where $[W_i], i \in \{1, 2, \dots, l\}$ are learnable context tokens that model a shared, type-agnostic context for abnormality. $\langle D \rangle$ is a placeholder for the defect descriptor, such as {“bent”, “contamination”, ...}. The tokens [anomaly] and [object] are fixed literals corresponding to the words “anomaly” and “object”.

A key design choice is to share the learnable tokens $[W_i]$ across all defect types, yielding a single set of parameters that learns a defect-aware context composable with different descriptors: the explicit defect word $\langle D \rangle$ injects type-specific semantics (e.g., “scratch”, “crack”), while the shared tokens provide transferable anomaly knowledge that improves generalization across defect categories. The fixed anchor [anomaly] grounds the prompt in CLIP’s pre-trained semantic space, and [object] preserves object-level context while allowing the learnable tokens to focus on defect cues that generalize across object categories [36]. Similarly, the normal prompt can be represented as:

$$\mathbf{P}^N = [V_1][V_2] \dots [V_l][\text{normal}][\text{object}]. \quad (2)$$

where $[V_i], i \in \{1, 2, \dots, l\}$ are learnable context tokens for normal state. Building on this foundation, to further enhance the representation capability, for one normal and K abnormal states of different defect types, we generate E prompts, each capturing different aspects of the anomaly space. For any given defect type,

all E prompts are utilized, meaning that a single defect type (e.g., “scratch”) is described through E different learnable contextual embeddings. The prompt set can be written as:

$$\begin{aligned}\mathcal{B} &= \{\mathbf{P}_1^N, \mathbf{P}_2^N, \dots, \mathbf{P}_E^N\}, \\ \mathcal{A}_k &= \{\mathbf{P}_1^{D_k}, \mathbf{P}_2^{D_k}, \dots, \mathbf{P}_E^{D_k}\},\end{aligned}\quad (3)$$

where $\mathcal{B}$ is the prompt set for the normal state, while $\mathcal{A}_k$ is the prompt set for the abnormal state of the k -th defect.

4.2 Prompt Optimization

Instead of fine-tuning the parameters of the network, we optimize the prompt set we defined in Equation (3). In the following section, we discuss the encoding process of images and texts and the loss function design for local and global optimization.

Image Encoding Given an input image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, the image encoder generates intermediate image patch embeddings at M encoding stages or layers and a single global image embedding $\mathbf{z}^x$ in the shared vision-text semantic space $\mathcal{Z}$. For each stage $i \in \{1, 2, \dots, M\}$, a linear adapter transforms the corresponding patch embeddings into the vision-text semantic space $\mathcal{Z}$. We denote these adapted embeddings as $\mathbf{Z}_i^p \in \mathbb{R}^{C \times H' \times W'}$. $H' \times W'$ are the resolution of image patch embeddings. Formally, the output of the image encoder is defined by:

$$(\mathbf{z}^x, \mathbf{Z}_1^p, \dots, \mathbf{Z}_M^p) = f_\theta(\mathbf{x}), \quad (4)$$

where $f_\theta(\cdot)$ is the parameterized image encoder function.

Text Encoding For each prompt $\mathbf{P} \in \mathcal{A}_k$, the text encoder $f_\lambda(\cdot)$ encodes the prompt to a text embedding. The embedding representing the k -th defect state is

$$\mathbf{z}^{D_k} = \frac{1}{E} \sum_{\mathbf{P} \in \mathcal{A}_k} f_\lambda(\mathbf{P}), \quad k \in \{1, \dots, K\} \quad (5)$$

Similarly, the embedding for the normal state is

$$\mathbf{z}^N = \frac{1}{E} \sum_{\mathbf{P} \in \mathcal{B}} f_\lambda(\mathbf{P}). \quad (6)$$

In total, we have $K+1$ embeddings $\{\mathbf{z}^N, \mathbf{z}^{D_1}, \dots, \mathbf{z}^{D_K}\}$ represent one normal state and K abnormal state.

Following prior work on progressive prompt tuning, we adopt a progressive prompt injection strategy to stabilize optimization when inserting learnable tokens into both visual and textual encoders.

4.3 Loss Function Design

Our loss function consists of two components: The global loss uses cosine similarity to align the global image embedding with its corresponding normal or abnormal state, capturing high-level semantic differences between samples. This is achieved by calculating the cosine similarity between the global image embedding $\mathbf{z}^x$ with $\mathbf{z}^N$ and the averaged defect-aware prompts, $\frac{1}{K} \sum_{k=1}^K \mathbf{z}^{D_k}$ to get the similarity score $\mathbf{s} \in \mathbb{R}^2$. The global loss is defined by

$$\mathcal{L}_{\text{global}} = CE(\text{softmax}(\mathbf{s}), y), \quad (7)$$

where $CE(\cdot)$ represents the cross-entropy calculation.

The local loss aligns image patch embeddings with one of the $K + 1$ states at a fine-grained spatial resolution. Each adapted image patch embedding $\mathbf{Z}_i^p$ is compared with $K + 1$ text embeddings to get the similarity maps. Since we choose image patch embeddings at M different layers, we get M different similarity maps $\hat{\mathbf{S}}_i \in \mathbb{R}^{(K+1) \times H' \times W'}$. This allows the model to localize and differentiate between defect types and normal regions at a finer spatial resolution. The local loss is:

$$\begin{aligned} \mathcal{L}_{\text{local}} = \frac{1}{M} \sum_{i=1}^M & Focal(UP(\hat{\mathbf{S}}_i), \mathbf{Y}) + Dice(UP(\hat{\mathbf{S}}_i[0]), \mathbf{Y}[0]) \\ & + Dice(\mathbf{1} - UP(\hat{\mathbf{S}}_i[0]), \mathbf{1} - \mathbf{Y}[0]), \end{aligned} \quad (8)$$

where $Focal(\cdot)$, and $Dice(\cdot)$ are focal [19] and dice [18] losses, respectively. Focal loss is designed to address the class imbalance problem, which is especially relevant in AD, where anomalous samples are significantly fewer than normal ones. Dice loss encourages the model to learn accurate decision boundaries by measuring the overlap between the predicted and ground truth segmentation mask. $UP(\cdot)$ is the upsampling function to enlarge the resolution to the original image size $H \times W$. $\hat{\mathbf{S}}_i[0]$ represents the predicted probability map for the normal state at stage i . $\mathbf{Y} \in \mathbb{R}^{(K+1) \times H \times W}$ represents the ground truth of the multi-defect type segmentation and $\mathbf{Y}[0]$ is the layer for the normal part. The final loss is a weighted combination of the global and local losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \lambda \mathcal{L}_{\text{local}} \quad (9)$$

where λ is a hyperparameter used for balancing. It is important to note that the underlying feature spaces of both text and vision modalities are pre-trained to capture image-text semantics, rather than anomaly semantics. This motivates refining the embedding space to better align the modalities within the anomaly detection domain, as described in the following section.

4.4 Training and Inference

At inference, we instantiate $(K' + 1)$ prompts (normal plus the target defect descriptors), aggregate the similarity maps across the M stages, and obtain (i) a binary anomaly map as $UP(1 - \hat{\mathbf{S}}[0])$ and (ii) a multi-type mask by pixel-wise argmax over the $(K' + 1)$ channels.

5 Experiments

In this section, we describe the experimental settings and benchmarks, define the evaluation metrics, and analyze the performance of DAPO through comparisons, ablations, and qualitative examples.

5.1 Datasets

Five public industrial datasets are used: MVTec-AD [1], VisA [40], MPDD [12], MAD [35], and Real-IAD [25], as well as an internal semiconductor (ASIC) inspection dataset. Our internal dataset consists of Application-Specific Integrated Circuit (ASIC) inspection images collected from a real manufacturing line. Unlike public benchmarks that span multiple object categories and defect taxonomies, this dataset contains one product type and a single defect type (contamination), with pixel-level annotations of contaminated regions. Samples of this dataset are exemplified in Figure 3b. We evaluate both image-level anomaly detection and pixel-level segmentation and use this dataset as a focused case study to assess performance under realistic acquisition conditions. For datasets with defect-type annotations, we report multi-type anomaly segmentation results using the dataset-provided defect labels.

5.2 Implementation Details

We use a frozen OpenCLIP [10] ViT-L/14@336 backbone with OpenAI pre-trained weights, and optimize only the learnable prompts and M linear adapters while keeping both the vision and text encoders frozen throughout training. These trainable components constitute only $\sim 0.15\%$ of the model, allowing training in ~ 25 min on MVTec-AD and inference in ~ 55 min on the large-scale Real-IAD dataset on a single NVIDIA H200 GPU. Following a source $\rightarrow$ target transfer protocol, we train on MVTec-AD and evaluate on the remaining benchmarks, and additionally train on VisA to evaluate on MVTec-AD. Input images are resized to 518×518 and normalized with the standard CLIP preprocessing pipeline. We optimize with Adam [13] (learning rate 10^{-3} , batch size 8) for 5 epochs, implemented in PyTorch 2.7.0. We extract patch features from layers $\{6, 12, 18, 24\}$ ($M=4$) and set the prompt length $l=5$, the number of prompt instances $E=10$, and the local-loss weight $\lambda=4.0$. We initialize the learnable context tokens within CLIP’s embedding space, matching its token-embedding mean μ and standard deviation σ , which converges faster and yields better results than random initialization; empirically, performance peaks within the first one or two epochs, with later epochs tending to overfit the source distribution.

5.3 Evaluation Metrics

To evaluate the performance, we use the area under the receiver operating characteristic curve (AUROC), average precision (AP) for AD. For segmentation,

we use pixel-level AUROC and the area under per-region overlap (AUPRO). For multi-type segmentation, we use AUROC, AP, and F1 by macro-averaging. In addition, for deployment-oriented analysis on the internal dataset, we report operating-point metrics derived from the confusion matrix, including true positive rate (TPR) and false positive rate (FPR), where $\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ and $\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$. On the public benchmarks, all baseline numbers are *reported* from their original papers or, where unavailable under our protocol, from MultiADS [23], and DAPO is evaluated under the same source→target protocol; on the internal ASIC dataset, AnomalyCLIP and MultiADS are *reproduced* by us using the authors’ released code.

6 Results

We evaluate our approach on two primary tasks: (1) Binary Anomaly Detection and Segmentation under distribution shift. (2) Multi-Type Anomaly Detection, a task to evaluate fine-grained localization proposed in MultiADS [23].

6.1 Multi-type Anomaly Segmentation

We first evaluate our primary contribution: using prompt-learning for zero-shot multi-type anomaly segmentation. We compare our method with the handcrafted prompt baseline, MultiADS [23]. The performance comparison is detailed in Table 1. To the best of our knowledge, we are among the first to apply a prompt-learning approach to the zero-shot multi-type anomaly segmentation task. Our approach consistently outperforms MultiADS, measured by the AP metric, and achieves comparable or superior pixel-level F1 scores. The performance margin is the largest in MPDD with a +7.2% on F1-score and a +6.0% on MAD-Real datasets, showing the strength of our prompt-learning approach for localizing visually present anomalies such as “scratch” or “crack”. However, our method shows a slight decrease in the VisA and MAD-Sim dataset with respect to the AUROC metric, explained by dataset composition: the lower AUROC on VisA and MAD-Real stems from the high frequency of “missing” component anomalies, which require object-level semantic priors to identify. Our method is not guided by such class semantics, while MultiADS uses class names to gain an advantage on these specific “absence” defects. In contrast, the superior AP and F1 scores demonstrate that for visually present anomalies, our defect-aware prompts improve the alignment between the modalities, leading to more accurate localization. More generally, the pattern of a high pixel-level AUROC together with lower AP/F1 reflects the strong class imbalance of MTAS: AUROC rewards correctly ranking the abundant normal background pixels, whereas AP and F1 directly penalize false positives within the predicted defect regions, which favors DAPO’s conservative, high-precision predictions that concentrate on the most discriminative abnormal evidence.

To better understand the model’s generalization, we analyze individual defect types (Table 2). DAPO generalizes well to unseen(*) defects with a strong,

Table 1: Zero-Shot Multi-type Anomaly Segmentation performance. Best results are highlighted in bold. MultiADS results are directly reported from the original paper.

Dataset		Method	Pixel-Level		
Train	Test		AUROC	AP	F1-score
MVTec-AD	VisA	MultiADS	93.6	24.8	22.1
		DAPO (ours)	88.9	25.0	22.2
	MPDD	MultiADS	95.2	53	42.8
		DAPO (ours)	95.5	53.6	50.0
	MAD-Sim	MultiADS	92.1	31.5	27.9
		DAPO (ours)	91.9	33.3	25.0
	MAD-Real	MultiADS	89.2	52.3	52.5
		DAPO (ours)	86.1	58.5	58.5

visually-present signal, such as textural or geometric deviations: it reaches 66.6% F1 on the unseen “flattening” defect in MPDD, and on MAD-Real and VisA it identifies “stains” (40.4% F1 vs. 6.1%) and “breakage” (20.8% vs. 7.9%). Averaged over MultiADS across the representative unseen defect types in Table 2, this amounts to a 5.2% mean improvement in AUROC and F1, underscoring its ability to localize novel defects in the zero-shot setting. The strength fades for anomalies lacking a strong local signal. For fine-grained defects such as “particle” on VisA, detection is high (99% AUROC) but segmentation precision is low: the model recognizes that something is wrong yet cannot localize it precisely. It also struggles with “missing” components: the 1.8% F1 on MAD-Real directly explains the overall AUROC gap in Table 1, as a missing component leaves a visually normal patch (e.g., the background) that provides no signal for the patch-based ViT encoder to align with a text prompt.

6.2 Binary Detection and Segmentation under Distribution Shift

Table 3 reports binary AD performance under distribution shift at both image and pixel levels. Following prior protocols [23, 36], we train on MVTec-AD and evaluate on VisA, MPDD, and Real-IAD as target domains. DAPO achieves the strongest **image-level** performance across all targets, obtaining the highest AUROC and AP, which indicates robust transfer under substantial domain shift. Averaged over the two strongest defect-aware baselines (AnomalyCLIP and MultiADS) across VisA, MPDD, and Real-IAD, this amounts to a 3.6% mean improvement in image-level AUROC and AP. For reference, on VisA DAPO attains an image-level AUROC of 84.9, ahead of the recent fine-grained-prompt methods FiLo (83.9) and FAPrompt (84.6) and well above AdaCLIP (75.4) under the same source→target protocol (FiLo/FAPrompt as reported in their original papers). At the pixel level, our results are competitive, while AUPRO remains below some baselines. Since AUPRO rewards predictions that expand to cover the entire annotated mask, DAPO’s conservative heatmaps, which highlight the most discriminative abnormal evidence rather than the full region, are penalized, especially for large or ambiguous defects with broad ground-truth masks

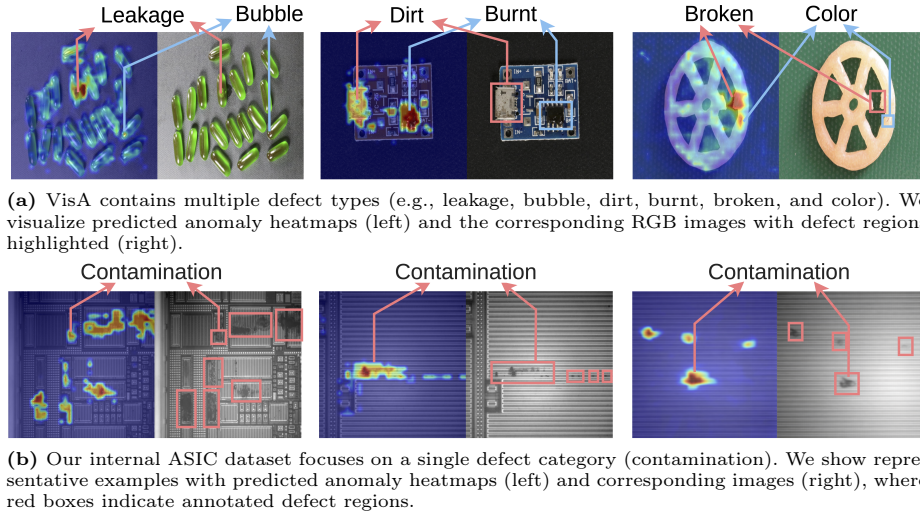


Fig. 3: Qualitative examples of DAPO from VisA and our internal ASIC dataset.

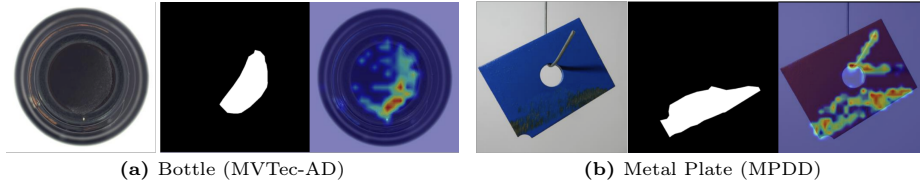


Fig. 4: Visualization of large-area defect predictions. DAPO highlights the most discriminative abnormal evidence rather than expanding to the full annotated region, which lowers AUPRO on broad ground-truth masks while suppressing false positives on normal regions.

(Fig. 4). This same conservative behavior is beneficial in deployment: on the internal ASIC dataset, it avoids excessive activation on normal regions and helps reduce false alarms.

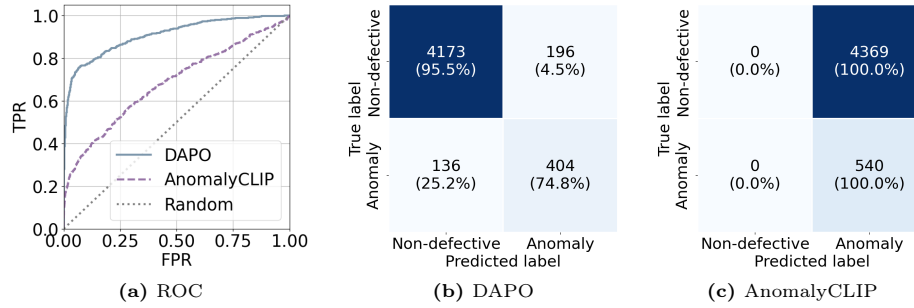
Importantly, we additionally evaluate on an internal ASIC dataset (Table 4), which reflects a more deployment-realistic setting than public benchmarks. On this dataset, AnomalyCLIP exhibits a marked performance degradation (68.2 AUROC / 35.3 AP), whereas defect-aware approaches (MultiADS and DAPO) remain substantially more stable. This gap suggests that explicitly incorporating defect semantics can improve robustness when moving from curated benchmarks to real production data, consistent with the qualitative examples in Fig. 3.

Fig. 5 further analyzes behavior on the internal dataset. The ROC curves (Fig. 5a) show that DAPO consistently dominates AnomalyCLIP across operating points, indicating better separability between normal and contaminated samples under production conditions. At a fixed threshold, the confusion matrices (Fig. 5b, 5c) reveal a balanced trade-off for DAPO (TN=4173, TP=404),

Table 2: Pixel-level segmentation for representative seen and unseen(*) defects across datasets, reported per defect type.

Dataset	Defect Type	MultiADS [23]		DAPO (ours)	
		AUROC	F1	AUROC	F1
VisA	Extra*	94.1	2.1	95.4	3.5
	Missing*	88.5	5.3	87.6	13.3
	Particle*	97.2	0.2	99.0	3.1
	Breakage*	98.5	7.9	97.1	20.8
MPDD	Flattening*	96.7	36.1	98.8	66.6
	Mismatch*	88.4	2.6	93.7	17.9
	Rust*	88.4	26.1	90.1	28.4
	Scratch	96.7	27.0	97.4	26.8
MAD-Real	Stains	97.0	6.1	98.7	40.4
	Missing	84.1	3.7	77.7	1.8
MAD-Sim	Burrs*	95.6	1.2	95.4	1.4
	Stains	98.2	15.0	99.2	30.0

operating at FPR=4.5% and TPR=74.8%, with low false alarms and strong defect recall, an important property for deployment. In contrast, AnomalyCLIP assigns the anomaly label to nearly all samples, yielding an impractically high false-alarm rate.

**Fig. 5:** Binary anomaly detection performance comparison. (a) ROC curves for DAPO, AnomalyCLIP, and Random Classifier. (b) and (c) are confusion matrices, demonstrating classification accuracy, with our approach DAPO.

6.3 Ablation Study

We further conduct a prompt decomposition study to isolate the contribution of each component in DAPO: (a) fixed defect-name anchors (*text-only*), (b) shared

Table 3: Binary anomaly detection and anomaly segmentation performance under distribution shift. Best results are in **bold**. The second-best results are underlined.

Binary AD and AS		Image-Level		Pixel-Level	
Dataset	Method	AUROC	AP	AUROC	AUPRO
VisA	CLIP-AC	65.0	70.1	24.8	17.3
	CoCoOp	78.1	-	93.6	-
	WinCLIP	78.1	81.2	79.6	56.8
	AprilGAN	78.0	81.4	94.2	86.8
	AnomalyCLIP	82.1	85.4	95.5	<u>87.0</u>
	AdaCLIP	75.4	79.3	<u>95.0</u>	-
	MultiADS	<u>83.6</u>	<u>86.9</u>	<u>95.0</u>	89.7
	DAPO (ours)	84.9	87.1	94.3	84
MPDD	CLIP-AC	56.2	66.0	58.7	29.1
	CoCoOp	61	-	95.1	-
	WinCLIP	63.6	69.9	76.4	48.9
	AprilGAN	73.0	80.2	94.1	83.2
	AnomalyCLIP	77.0	82.0	96.5	88.7
	AdaCLIP	66.3	75	<u>96.3</u>	-
	MultiADS	78.3	78.4	95.8	89.7
	DAPO (ours)	81.2	83.6	95.1	84.4
Real-IAD	WinCLIP	75	72.3	87.1	59.9
	AprilGAN	75.7	73.5	96	86.8
	AnomalyCLIP	78.4	76.7	96.2	85.7
	AdaCLIP	70.1	68.5	95.3	-
	MultiADS	78.7	79.1	96.6	87.1
	DAPO (ours)	84.3	84	<u>96.4</u>	80.3

learnable context tokens without defect anchors (*soft-only*), and (c) their hybrid composition (DAPO). As shown in Table 5, text-only prompts preserve semantic specificity but remain limited by the image-text modality gap, leading to weak transfer in practice. In contrast, soft-only prompts substantially improve binary anomaly detection by learning task-adaptive context, yet they lack explicit type cues and therefore fail to produce well-separated decision regions required for $(K+1)$ -channel multi-type anomaly segmentation (MTAS). Combining shared context tokens with defect-name anchors yields the best of both worlds: the anchors provide explicit defect semantics, while the learnable

Table 4: Binary anomaly detection performance under distribution shift on our internal semiconductor dataset.

Binary Anomaly Detection		Image-Level	
Dataset	Method	AUROC	AP
Semi-conductor (Internal Dataset)	AnomalyCLIP	68.2	35.3
	MultiADS	91.1	61.7
	DAPO (ours)	91.8	79.1

tokens capture transferable anomaly context and improve cross-domain alignment, resulting in the strongest AUROC/AP and enabling MTAS. The lower block of Table 5 isolates three further design choices. Removing token sharing and learning a separate prompt per defect type (*per-defect tokens*) drops AUROC to 70.9, confirming that a shared, type-agnostic context, rather than memorizing individual defect prompts, is what enables transfer; this directly validates our central design choice. Disabling the progressive injection of learnable tokens (*w/o progressive fine-tuning*) reduces AUROC to 79.7, and replacing the lightweight linear adapters with deeper 2-layer adapters brings no gain (80.5). Together these confirm that DAPO’s improvements stem from its compositional, shared-token design rather than from additional trainable capacity.

Table 5: Component ablations on MPDD (image-level AUROC/AP). The upper block isolates the prompt composition (defect-name anchor vs. shared learnable tokens); the lower block ablates token sharing, progressive fine-tuning, and adapter depth. ✓/✗ indicate whether the variant supports the $(K+1)$ -channel MTAS task.

Variant	AUROC	AP	MTAS
Defect-anchor only (text, non-learnable)	53.8	65.0	✓
Soft-only (no defect anchor)	73.7	78.1	✗
Per-defect tokens (no sharing)	70.9	75.6	✓
w/o progressive fine-tuning	79.7	81.5	✓
2-layer adapters	80.5	82.3	✓
DAPO (shared tokens + anchor)	81.2	83.6	✓

7 Conclusion

We presented DAPO, a defect-aware prompt optimization framework for binary anomaly detection/segmentation and zero-shot multi-type anomaly segmentation. DAPO learns compositional hybrid prompts that combine a small set of shared learnable context tokens with explicit defect-descriptor anchors, enabling prompt instantiation for new defect descriptors at inference. By aligning global and patch-level visual representations with defect-aware text semantics, it improves fine-grained localization and robustness under distribution shift, achieving consistent gains in image-level detection and competitive segmentation across five public benchmarks and an internal dataset. Future work includes extending DAPO to other vision-language backbones and improving localization of absence-based structures via object-level priors and few-shot adaptation.

Acknowledgements

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hongkuan Zhou.

References

1. Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9592–9600 (2019)
2. Cao, Y., Zhang, J., Frittoli, L., Cheng, Y., Shen, W., Boracchi, G.: Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In: European Conference on Computer Vision. pp. 55–72. Springer (2024)
3. Chen, X., Han, Y., Zhang, J.: April-gan: 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 (2023)
4. Cohen, N., Hoshen, Y.: Sub-image anomaly detection with deep pyramid correspondences. arxiv 2020. arXiv preprint arXiv:2005.02357 (2005)
5. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: International conference on pattern recognition. pp. 475–489. Springer (2021)
6. Deng, C., Xu, H., Chen, X., Xu, H., Tu, X., Ding, X., Huang, Y.: Simclip: Refining image-text alignment with simple prompts for zero-/few-shot anomaly detection. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1761–1770 (2024)
7. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9737–9746 (2022)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
9. Gu, Z., Zhu, B., Zhu, G., Chen, Y., Li, H., Tang, M., Wang, J.: Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2041–2049 (2024)
10. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., et al.: Openclip. If you use this software, please cite it as below **7** (2021)
11. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19606–19616 (2023)
12. Jezek, S., Jonak, M., Burget, R., Dvorak, P., Skotak, M.: Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: 2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT). pp. 66–71. IEEE (2021)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

16. LeCun, Y., et al.: Generalization and network design strategies. *Connectionism in perspective* **19**(143-155), 18 (1989)
17. Li, S., Cao, J., Ye, P., Ding, Y., Tu, C., Chen, T.: Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *Neurocomputing* **618**, 129122 (2025)
18. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855* (2019)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2980–2988 (2017)
20. Liu, Y., Zhuang, C., Lu, F.: Unsupervised two-stage anomaly detection. *arXiv preprint arXiv:2103.11671* (2021)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
22. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14318–14328 (2022)
23. Sadikaj, Y., Zhou, H., Halilaj, L., Schmid, S., Staab, S., Plant, C.: Multiads: Defect-aware supervision for multi-type anomaly detection and segmentation in zero-shot learning. *arXiv preprint arXiv:2504.06740* (2025)
24. Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly localization in images. In: *European Conference on Computer Vision*. pp. 485–503. Springer (2020)
25. Wang, C., Zhu, W., Gao, B.B., Gan, Z., Zhang, J., Gu, Z., Qian, S., Chen, M., Ma, L.: Real-riad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22883–22892 (2024)
26. Xu, J., Lo, S.Y., Safaei, B., Patel, V.M., Dwivedi, I.: Towards zero-shot anomaly detection and reasoning with multimodal large language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025)
27. Yang, C.A., Peng, K.C., Yeh, R.A.: Toward long-tailed online anomaly detection through class-agnostic concepts. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 23419–23430 (2025)
28. Yang, J., Shi, Y., Qi, Z.: Dfr: Deep feature reconstruction for unsupervised anomaly segmentation. *arXiv preprint arXiv:2012.07122* (2020)
29. Yao, X., Blei, T., Meng, Y., Zhang, Y., Zhou, H., Bing, Z., Huang, K., Sun, F., Knoll, A.: Long-horizon language-conditioned imitation learning for robotic manipulation. *IEEE/ASME Transactions on Mechatronics* pp. 1–12 (2025). <https://doi.org/10.1109/TMECH.2025.3547047>
30. Zhou, H., Bing, Z., Yao, X., Su, X., Yang, C., Huang, K., Knoll, A.: Language-conditioned imitation learning with base skill priors under unstructured data. *IEEE Robotics and Automation Letters* **9**(11), 9805–9812 (2024). <https://doi.org/10.1109/LRA.2024.3466076>
31. Zhou, H., Schmid, S., Li, Y., Halilaj, L., Yao, X., Cao, W.: Predicting the road ahead: A knowledge graph based foundation model for scene understanding in autonomous driving. In: Curry, E., Acosta, M., Poveda-Villalón, M., van Erp, M., Ojo, A., Hose, K., Shimizu, C., Lisena, P. (eds.) *The Semantic Web*. pp. 116–132. Springer Nature Switzerland, Cham (2025)

32. Zhou, H., Yao, X., Mees, O., Meng, Y., Xiao, T., Bisk, Y., Oh, J., Johns, E., Shridhar, M., Shah, D., Thomason, J., Huang, K., Chai, J., Bing, Z., Knoll, A.: Bridging language and action: A survey of language-conditioned robot manipulation. *CoRR abs/2312.10807* (2023), <https://doi.org/10.48550/arXiv.2312.10807>
33. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 16816–16825 (2022)
34. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *International Journal of Computer Vision* **130**(9), 2337–2348 (2022)
35. Zhou, Q., Li, W., Jiang, L., Wang, G., Zhou, G., Zhang, S., Zhao, H.: Pad: A dataset and benchmark for pose-agnostic anomaly detection. *Advances in Neural Information Processing Systems* **36**, 44558–44571 (2023)
36. Zhou, Q., Pang, G., Tian, Y., He, S., Chen, J.: Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In: *International Conference on Learning Representations* (2024)
37. Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, H., Knoll, A.C.: Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles* pp. 1–20 (2024). <https://doi.org/10.1109/TIV.2024.3402136>
38. Zhu, J., Ong, Y.S., Shen, C., Pang, G.: Fine-grained abnormality prompt learning for zero-shot anomaly detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2025)
39. Zolfaghari, M., Sajedi, H.: Unsupervised anomaly detection with an enhanced teacher for student-teacher feature pyramid matching. In: *2022 27th International Computer Conference, Computer Society of Iran (CSICC)*. pp. 1–4. IEEE (2022)
40. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: *European Conference on Computer Vision*. pp. 392–408. Springer (2022)