VMAD: Visual-Enhanced Multimodal Large Language Model for Zero-Shot Anomaly Detection

Huilin Deng, Hongchen Luo, Wei Zhai, Yanming Guo, Yang Cao, *Member, IEEE*, and Yu Kang, *Senior Member, IEEE*

***Abstract*—Zero-shot anomaly detection (ZSAD) enables the inspection of unseen objects by bridging textual prompts and visual features, showing great potential in flexible manufacturing. While existing ZSAD methods rely on predefined prompts and struggle with unseen defects, Multimodal Large Language Models (MLLMs) offer promising solutions through their generative and interpretative capabilities. However, adapting MLLMs to Industrial Anomaly Detection (IAD) remains challenging due to fine-grained anomaly patterns and subtle visual distinctions. We propose VMAD (Visual-enhanced MLLM Anomaly Detection), a framework that enriches MLLM with visual IAD knowledge through two key components: a Defect-Sensitive Structure Learning scheme that transfers patch-similarities for improved discrimination, and a Locality-enhanced Token Compression that leverages multi-level local features for fine-grained detection. We also introduce RIAD, a comprehensive IAD dataset with detailed anomaly annotations. Extensive experiments on MVTec-AD, Visa, WFDD, and RIAD demonstrate VMAD's superior performance. The dataset and code will be publicly available at https://github.com/denghuilin-cyber/VMAD**

***Note to Practitioners*—This research addresses zero-shot Industrial Anomaly Detection (IAD) challenges in dynamic production. Our VMAD framework, leveraging Multimodal Large Language Models (MLLMs), offers: 1. Flexibility: Adapts to new products and unknown defects without extensive labeled data. 2. Interpretability: Provides defect localization and explanations for analysis. 3. Interactivity: Supports follow-up queries for deeper insights. VMAD enhances MLLMs through novel cross-modal learning for improved defect recognition and a visual embedding compression method for fine-grained perception. However, industrial applications may require further optimization. Future work will focus on real-time performance and domain-specific integration. Furthermore, our RIAD dataset provides a valuable resource for developing and evaluating IAD systems, helping to bridge the gap between research and practical applications.**

***Index Terms*—Anomaly detection, zero-shot learning, multimodal large language models.**

Received 3 October 2024; revised 22 January 2025 and 15 June 2025; accepted 11 July 2025. Date of publication 22 July 2025; date of current version 11 February 2026. This article was recommended for publication by Editor H. Liu upon evaluation of the reviewers' comments. This work was supported by the National Natural Science Foundation of China under Grant 62033012 and Grant 62306295. *(Corresponding authors: Hongchen Luo; Yang Cao.)*

Huilin Deng and Wei Zhai are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230052, China (e-mail: huilin_deng@mail.ustc.edu.cn; wzhai056@ustc.edu.cn).

Hongchen Luo is with the College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110819, China (e-mail: luohongchen@ise.neu.edu.cn).

Yanming Guo is with the College of Systems Engineering, National University of Defense Technology, Changsha, Hunan 410073, China (e-mail: guoyanming@nudt.edu.cn).

Yang Cao and Yu Kang are with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230052, China, and also with the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, Anhui 230088, China (e-mail: forrest@ustc.edu.cn; kangduyu@ustc.edu.cn).

Digital Object Identifier 10.1109/TASE.2025.3591656

I. Introduction

INDUSTRIAL Anomaly Detection (IAD) aims to classify and localize defects in industrial manufacturing. By identifying abnormal patterns in industrial processes, IAD techniques enable timely intervention and optimization, thereby enhancing overall productivity. Collecting anomaly data is challenging due to their rarity and unpredictability [1]. Therefore, conventional IAD works [2], [3] mainly explore unsupervised techniques, measuring the deviation of test sample features from the learned normal distribution. However, these methods necessitate abundant training samples. Consequently, they exhibit limited generalization to novel classes and fail to adapt to dynamic production environments.

Recently, Zero-Shot Anomaly Detection (ZSAD) offers flexible inspection by leveraging text prompts as additional information, enabling anomaly detection of unseen objects. Mainstream Contrast-based methods, based on pre-trained CLIP [4], *compare image features to textual descriptors representing 'normal' and 'abnormal' respectively*, as illustrated in Fig. 1 (a). WinCLIP [5] initially adapts CLIP using manually crafted prompts. AnomalyCLIP [6] further substitutes manual templates with object-agnostic text vectors for generic representation. ClipSAM [7] leverages SAM to CLIP's anomaly localization. Despite the promise, predefined prompts subject them to the following limitations: 1) Limited Flexibility: generic descriptors (**e.g.**, 'damaged' in Fig.1 (a)) are insufficient to capture diverse anomalies in industrial manufacturing. 2) Bias and Incompleteness: fixed prompt templates (**e.g.**, "A xxx photo of xxx") are primarily designed for foreground object classification. The focus on overall object semantics can lead to biases and incomplete detection of normal and abnormal parts within objects. Therefore, these methods still perform ZSAD in a ***closed-world setting, executing binary classification in constrained semantic space with predefined prompts, thus struggling with unseen defects.***

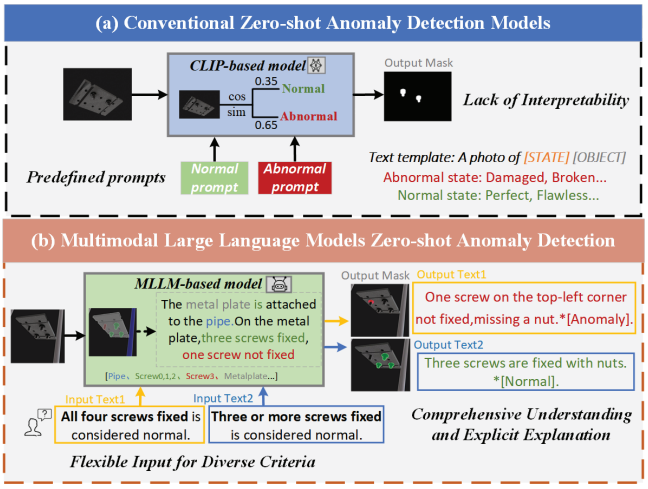


Fig. 1. **Comparison between previous ZSAD methods and MLLMs-based ZSAD methods.** (a) Previous ZSAD methods rely on predefined prompts, confining them to closed-world anomaly detection, which lacks the flexibility to adapt to new categories or capture specific types of anomalies. (b) MLLM-based methods leverage open-ended text interpretation and generation to offer a more open, adaptable, and comprehensive anomaly detection.

Most recently, Multimodal Large Language Models (MLLMs) have emerged as a promising solution to closed-world limitations. Unlike contrast-based methods with predefined prompts, MLLMs enable open-ended visual interpretation through generative text analysis. Their flexibility in processing arbitrary textual prompts with visual inputs facilitates dynamic anomaly analysis under varying criteria (as shown in Fig. 1 (b), the same sample yields different inspection results based on varying criteria). However, ***anomalies are visually confusing, characterized by minimal discriminative feature variances between normal and anomalous samples, mostly manifesting in object-localized regions.*** While MLLMs' general visual interpretation, they often struggle with abnormality discriminability and fine-grained perception.

This weak discriminability often stems from minor manufacturing flaws or imperfections that complicate the detection process. Our investigation uncovers a promising approach to address this issue by focusing on the pronounced disparities in patch-similarity distributions between normal and anomalous samples. Specifically, despite subtle global visual variances, we reveal that while normal patches frequently exhibit visual parallels across multiple unmarked samples, anomalous areas rarely find such correspondences, an observation similar to MuSc [8]. Leveraging this patch-similarity disparity as valuable IAD-specific knowledge, we propose a **D**efect-**S**ensitive **S**tructure **L**earning (**DSSL**) scheme to amplify the abnormality distinction in LLM representation. DSSL transfers patch-similarity from visual to LLM space, maintaining text-visual semantic consistency. Specifically, DSSL calculates and aligns two similarity distributions: Visual Patch Similarity (between local input image patches and normal global features) and Text-Visual Similarity (between visual patches and semantic tokens of normal samples in LLM space). These distributions are aligned through contrastive learning [9]. Notably, semantic tokens combine both linguistic and visual cues [10] from normal descriptions with projected visual features.

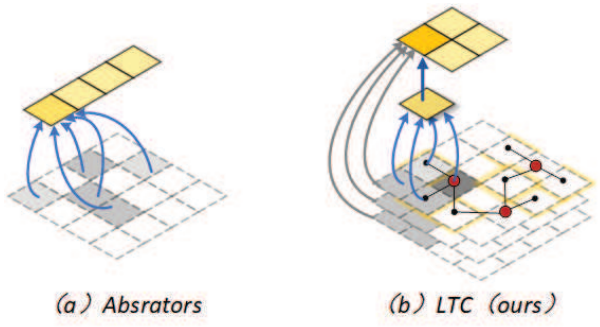


Fig. 2. **Various visual projectors.** Abstractors compress limited information, while LTC mines multi-level local cues.

The second challenge concerns the fine-grained semantics of anomalies. Our analysis reveals that visual projectors, which transform visual signals into LLM-compatible tokens, fundamentally influence MLLMs' perception of subtle anomalies. Existing MLLMs employ abstractors for computational efficiency [11], [12], [13], but they compromise the integrity of visual features by aggregating information from constrained regions (Fig. 2 (a)). To address this limitation, we propose a **L**ocality-enhanced **T**oken **C**ompression (**LTC**) method which preserves rich semantics while reducing tokens. As outlined in Fig. 2 (b), LTC mines multi-level features in local contexts using coarse-to-fine injection with multi-level features' integration.

To this end, We propose VMAD (**V**isual-enhanced **M**LLM **A**nomaly **D**etection), a novel framework offering simultaneous anomaly localization and explainable text, with interactive engagement for follow-up inquiries. As illustrated in Fig. 3 (a), VMAD comprises an MLLM and a visual branch. The MLLM processes image-text inputs, generating '[seg]' tokens for prompt segmentation.

Moreover, we present **R**eal **I**ndustrial **A**nomaly **D**etection (**RIAD**), an extensive IAD dataset, encompassing 28,040 images across 24 object categories and 15 defect types. As shown in Fig. 5, RIAD features paired image-text data with masks, providing detailed anomaly descriptions, impact analyses, and recommendations, offering crucial IAD knowledge. Extensive experiments on zero-shot benchmarks, including MVTec-AD, VisA, WFDD, and our RIAD, demonstrate VMAD's superior performance over state-of-the-art methods. Plus, VMAD exhibits an exceptional ability to provide assessments and elaborative insights for industrial defects.

Our contributions are summarized as follows: **1)** We propose VMAD, a novel framework for IAD that simultaneously localizes anomalies and generates explanatory text. **2)** We design a cross-modal learning scheme,**DSSL**, which integrates the MLLM with visual-similarity cues for enhanced anomaly discrimination. **3)** We introduce a novel visual projector, the LTC mechanism, **LTC**, which produces high-quality tokens for fine-grained defect perception. **4)** We collect a dataset named RIAD containing plenty of paired image-text data with anomaly masks, providing a comprehensive resource for MLLM-based IAD development.

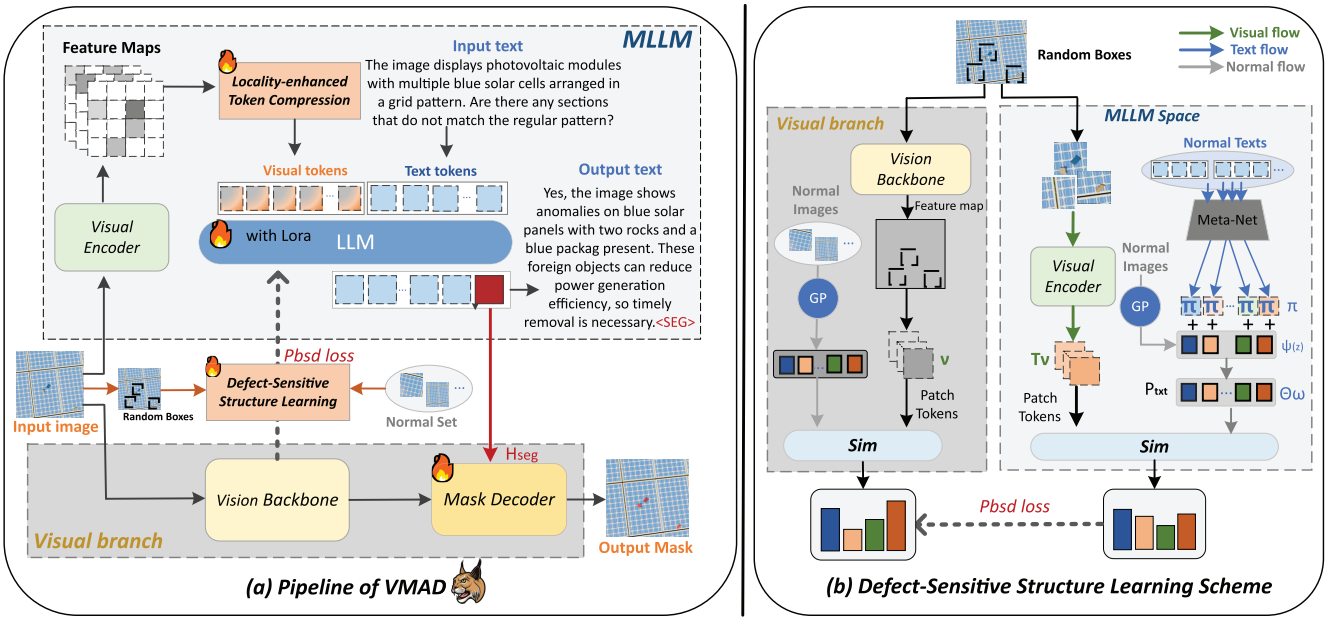


Fig. 3. **(Left) Overview of VMAD.** VMAD incorporates a visual branch for anomaly localization (Sec. III-B), with Locality-enhanced Token Compression serving as a visual projector (Sec. III-D). **(Right) Defect-Sensitive Structure Learning.** It aligns visual and text-visual patch similarity distributions using PBSD loss, enhancing MLLM's sensitivity to anomalous structures (Sec. III-C). **GP:** Global Pooling, **Sim:** Similarity computation.

II. Related Work

A. Industrial Anomaly Detection

Given the scarcity of anomalies, conventional Industrial Anomaly Detection (IAD) research mainly explores unsupervised and self-supervised techniques. Unsupervised approaches fall into two main streams. Reconstruction-based methods ([14], [15]) use encoder-decoder architectures to minimize input-reconstruction discrepancies. The Embedding-based branch ([16], [17]) identifies anomalies through feature disparities. It encompasses several sub-categories: a) One-class techniques ([2], [3]). b) Memory-augmented models ([18], [19]). c) Knowledge distillation frameworks ([20], [21]). On the other hand, Self-supervised methods leverage proxy tasks to learn discriminative representations between nominal and anomalous patterns [22], [23]. These conventional methods follow the 'one-class-one-model' paradigm and struggle with novel object classes. In contrast, our VMAD enables in-context learning for multiple novel object categories.

B. Zero-Shot Anomaly Detection

Zero-Shot Anomaly Detection (ZSAD) [5], [24], utilizing text prompts as auxiliary information to detect anomalies across unseen object categories, has emerged as a promising solution for flexible industrial inspection. WinCLIP [5] pioneers language-driven zero-shot IAD using CLIP [4]. Building on this, AnomalyCLIP [6] further substitutes manual templates with object-agnostic text vectors for generic representation. KAnoCLIP [25] introduces a knowledge-driven (KD) loss to create learnable anomaly prompts, removing the reliance on fixed prompts. Meanwhile, FILo [26] enhances anomaly prompts using learnable prompts tailored to specific anomaly types. As for another line, MuSc [8] uses a mutual scoring mechanism to compare patches in unlabeled images, assigning anomaly scores without additional labeled data or predefined prompts. Despite their promise, predefined prompts impose limited flexibility, with broad descriptors failing to represent the specific types of defects across object categories adequately. Fixed prompt templates can introduce biases and result in incomplete detection of abnormal parts. Moreover, these methods lack explainable analysis. In contrast, our proposed VMAD offers flexible detection and explainable analysis with a novel MLLM-based framework.

C. Multimodal Large Language Models (MLLMs)

Large Language Models (LLMs) have shown remarkable capabilities in language tasks. Building on this, Multimodal Large Language Models (MLLMs) extend these capabilities to the visual domain. Early efforts like Flamingo [27] and BLIP-2 [28] pioneered the use of Q-Former and Resampler to bridge vision and language. Subsequent models such as LLaVA [29] and MiniGPT4 [30] enhanced instruction following through visual instruction tuning. Recent advancements, including Ferret [31], LISA [32] and GLaMM [33] have pushed MLLMs towards fine-grained visual understanding and grounded conversation generation.

In Industrial Anomaly Detection, AnomalyGPT [34] and Myriad [35] pioneer MLLM applications. AnomalyGPT converts anomaly maps to learnable embeddings while Myriad encodes anomaly maps into LLM-compatible tokens via Q-Former to enable LLM's visual comprehension. ALFA [36] introduces a run-time prompt adaptation strategy for informative anomaly prompts to leverage the visual understanding capabilities of LLMs. However, both of them rely on pre-trained visual models for anomaly localization, confining LLMs to text responses and thus restricting overall generalization. Our VMAD extends LLMs to both anomaly localization

and analysis, introducing a novel patch-based scheme for cross-modal generalization.

Visual projectors are crucial for real-time IAD tasks, efficiently transforming visual features for LLM compatibility. Existing methods like QPN [37] and DeepStack [11] focus on query enhancement and token restructuring, while others use pixel shuffle [12] or nearby concatenation [38] to reduce visual tokens. However, these approaches compromise visual feature integrity in pursuit of efficiency [13]. Our LTC mechanism incorporates multi-level visual cues through a coarse-to-fine scheme, balancing efficiency with structural integrity.

III. Method

This section outlines the problem setup (Sec. III-A) and provides an overview of VMAD with its loss function (Sec. III-B). It then introduces the Defect-Sensitive Structure Learning scheme for MLLM and visual branch (Sec. III-C). Finally, we describe the Locality-enhanced Token Compression (LTC) mechanism as MLLM's visual projector (Sec. III-D).

A. Problem Setup

We extend traditional mask-only zero-shot anomaly detection by incorporating text responses. This task aims to detect anomalies in novel object categories without support images while providing textual explanations. Training (D_{train}) and testing (D_{test}) sets contain disjoint object classes. Query images and text instructions are fed into the network in pairs. The text instruction T_q includes *a description of normal scenes* and *questions about anomalies*. ***e.g.,'The gray metal plates in the picture should be secured with four pairs of screws, which must not be loose or missing. Is there any anomaly in this picture?'*** Given multi-modal inputs (I_{img}, T_q) from D_{train}, the model learns to associate semantic information to provide text responses and segmentation masks. During testing, the model is evaluated without further optimization.

B. Overall Architecture

Our network consists of a standard MLLM and a visual branch. The Defect-Sensitive Structure Learning (DSSL) scheme spans both branches, using visual patch similarity to enhance MLLM's anomaly discrimination. The Locality-enhanced Token Compression (LTC) mechanism serves as the MLLM's visual projector, generating compact, detail-rich visual tokens. Subsequent sections elaborate on each module.

1) MLLM Framework: The proposed MLLM framework comprises three pivotal components: **(1) Visual Encoder F_I:** the input image $I_{\text{img}} \in \mathbb{R}^{H\times W\times 3}$ is transformed into a set of visual embeddings $I_v \in \mathbb{R}^{N\times C}$ through the widely-utilized CLIP-ViT-L/14 [39] vision encoder. **(2) Visual Projector $\Psi_{I\to T}$:** it is responsible for projecting visual embeddings I_v into visual token T_v in the textual embedding space T, ensuring they have the appropriate dimension for a subsequent language model. The visual projector takes the N visual embeddings I_v and converts them to M visual tokens T_v, where $M < N$. **(3) LLM denoted as $\Theta(T_v, T_t)$:** it processes the visual token T_v and the textual token T_t, producing the coherent textual response auto-regressively. Formulated as:

$$I_v = F_I(I_{\text{img}}), \tag{1}$$

$$T_v = \Psi_{I\to T}(I_v), \tag{2}$$

$$\hat{y}_{txt} = \Theta(T_v, T_t). \tag{3}$$

The embedding-as-mask scheme [32] is leveraged to equip our MLLM with segmentation ability for anomaly localization. In particular, we augment the LLM's vocabulary with a specialized < seg > token, which serves as the bridge between the MLLM and downstream mask decoder. Moreover, the text output y_{txt} is ensured to end with the < seg > token.

2) Visual Branch: The < seg > token's last-layer embedding $\overline{H}_{seg}$ in the LLM is extracted and passed through an MLP projection layer $\Gamma(\cdot)$ to obtain H_{seg}. H_{seg} is then used to inform the mask decoder about the semantic knowledge necessary for anomaly discrimination. H_{seg} and f are fed into the mask decoder to obtain the final segmentation mask, where the image I_{img} are processed by the visual backbone (*e.g.*, SAM [40]) to extract visual feature maps f. The process can be expressed as:

$$H_{seg} = \Gamma(\overline{H}_{seg}), f = V_{backbone}(I_{\text{img}}), \tag{4}$$

$$\overline{M} = M_{decoder}(H_{seg}, f). \tag{5}$$

3) Training Objectives: The model is optimized end-to-end through the text generation loss $\mathcal{L}_{txt}$, segmentation mask loss $\mathcal{L}_{seg}$ and the specially designed defect-structure loss $\mathcal{L}_{struc}$. The overall objective $\mathcal{L}$ is the weighted sum of these losses, formally defined as:

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{pbsd}}\mathcal{L}_{\text{pbsd}}. \tag{6}$$

To encourage high-quality segmentation results, the segmentation loss L_{seg} is computed as a combination of per-pixel binary cross-entropy (BCE) loss and DICE loss with corresponding loss weights λ_{bce} and λ_{dice}. Concretely, given the predicted map $\hat{M}$ and query mask M_q, segmentation loss $\mathcal{L}_{seg}$ is formally defined as:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}} BCE\left(\hat{M}, M_q\right) + \lambda_{\text{dice}} DICE\left(\hat{M}, M_q\right). \tag{7}$$

For text generation, $\mathcal{L}_{\text{txt}}$ is the auto-regressive cross-entropy loss between predicted text $\hat{y}_{txt}$ and the ground-truth text y_{txt}. The patch distribution similarity loss $\mathcal{L}_{pbsd}$ (Sec. III-C Eq. 17) aims to enhance anomalous structure discriminability by transferring patch-level similarity knowledge from visual domain to multimodal space, enabling consistent normal-abnormal patch distinction across modalities.

C. Defect-Sensitive Structure Learning

While LLMs excel at vision-language understanding, their limited capacity in local visual pattern comprehension leads to misalignment between local semantic and visual distributions. DSSL transfers patch-level similarity from visual to LLM space, maintaining text-visual semantic consistency. Specifically, DSSL calculates and aligns two similarity distributions: Visual Patch Similarity (between local patches and global features) and Text-Visual Patch Similarity (between visual

patches and semantic tokens in LLM space). Concretely, we randomly generate a set of patch boxes $B_i[j]_{j=1}^{N_b}$ from the input image *Iimg*, where N_b is the number of patch boxes per image.

1) Visual Patch Similarity Distribution: To obtain feature embeddings of patch boxes $v[j]_{j=1}^{N_b}$, we apply ROI pooling with the patch coordinates, and then project the pooled features through a normalization head:

$$f = V_{backbone}(I_{img}), v^j = g_\alpha(\text{ROI}(f; B[j])), \tag{8}$$

where $V_{backbone}$ denotes the visual backbone, I_{img} is the input image, $B[j] \in \mathbb{R}^4$ refers to the spatial coordinates of j-th patch box, and g_α is the projection head for local embeddings.

To establish reliable global representations z, we apply global average pooling followed by a projection head:

$$z = g_\beta(GAP(h)). \tag{9}$$

Based on the global normalized features, we establish a comprehensive memory queue M_{img} and a normal-only reference set P_{img}, formulated as:

$$\text{M}_{\text{img}} = \{z_m : y_m = y^+ \text{ or } y_m = y^-\}, \text{P}_{\text{img}} = \{z_t : y_t = y^+\}. \tag{10}$$

where y^+ and y^- denote the normal and abnormal class labels, respectively. The M_{img} contains normal and abnormal samples from the same class as I_{img}, while positive set P_{img} *contains only global features of normal images from the same class.*

Finally, the similarity relationship between local patch and normal global representation is calculated as a conditional probability [41], using a pre-defined temperature parameter τ:

$$p(\boldsymbol{z_t}|v^j) = \frac{\exp(\boldsymbol{z_t} \cdot \boldsymbol{v^j}/\tau)}{\sum_{\boldsymbol{z_m} \in \text{M}_{\text{img}}} \exp(\boldsymbol{z_m} \cdot v^j/\tau)}. \tag{11}$$

Eq.11 encodes similarity between local patch (v^j) of input images and randomly sampled global representations (z_t) of positive set P_{img}. This local-global distinction is then used to guide normal-abnormal differentiation in LLM space.

2) Text-Visual Patch Similarity Distribution: It calculates the similarity between visual-patch and normal semantic tokens in LLM space. Notably, semantic tokens combine both linguistic and visual cues for a comprehensive representation of normality. To obtain patch tokens $\{T_v^j\}_{j=1}^{N_b}$ in LLM space, we crop image patches $\{I_{box}^j\}_{j=1}^{N_b}$ from the original image and project them into the LLM space:

$$I_{box}^j = \text{ROI}(I_{img}; B[j]), T_v^j = \boldsymbol{\Psi_{I \to T}}(F_I(I_{box}^j)). \tag{12}$$

To obtain semantic tokens for image x, we propose a three-stage process. First, we generate text tokens π by encoding normal descriptions through Meta-Net [10]. Next, we extract global visual features z using MLLM's visual encoder and project them into the LLM space. Finally, the text and visual tokens are fused through addition to obtain multimodal semantic tokens:

$$\boldsymbol{\theta_\omega(x)} = \boldsymbol{\Psi_{I \to T}}(z) + \pi, \quad \pi = \boldsymbol{Meta}(t_x), \tag{13}$$

where $\boldsymbol{\theta_\omega(x)}$ represents semantic tokens for the image x. The global normalized feature z is obtained consistently with our visual branch (Eq.9), while the feature is extracted through MLLM's visual encoder. In this work, the Meta-Net is built with a two-layer bottleneck (Linear-ReLU-Linear). Based on Eq. 13, we then define the positive set P_{txt} and memory queue M_{txt} as follows:

$$\text{P}_{\text{txt}} = \{\boldsymbol{\theta_\omega}(x_t) : y_t = y^+\}, \tag{14}$$

$$\text{M}_{\text{txt}} = \{\boldsymbol{\theta_\omega}(x_m) : y_m = y^+ \text{ or } y_m = y^-\}. \tag{15}$$

Similar to Eq. 11, the similarity distribution between patch tokens T_v^j and normal semantic tokens $\boldsymbol{\theta_\omega(z)}$ is computed as:

$$p(\boldsymbol{\theta_\omega}^t|T_v^j) = \frac{\exp(\boldsymbol{\theta_\omega}^t \cdot \boldsymbol{T_v}^j/\tau)}{\sum_{\boldsymbol{\theta_\omega^m} \in \text{M}_{\text{txt}}} \exp(\boldsymbol{\theta_\omega}^m \cdot v^j/\tau)}. \tag{16}$$

To align local-global visual distributions and multimodal semantic distributions obtained by Eq. 11 and Eq. 16, respectively, we introduce the Patch-Based Similarity Distribution (PBSD) loss. Mathematically, PBSD loss can be expressed as:

$$\mathcal{L}_{pbsd} = \frac{1}{N_b} \sum_{j=1}^{N_b} \sum_{\boldsymbol{x_t} \in \text{P}} -p(\boldsymbol{z_t}|v^j) \log p(\boldsymbol{\theta_\omega}^t|T_v^j), \tag{17}$$

where $p(\boldsymbol{z_t}|v^j)$ is detached from the computation graph to prevent gradient flow. The PBSD loss effectively transfers the knowledge of patch-level similarities from the visual domain to multimodal space, ensuring the model learns to distinguish normal and abnormal patches consistently across modalities.

D. Locality-Enhanced Token Compression

In real-time industrial inspection, abstractors sacrifice spatial information for efficiency (Fig. 2 (a)). Our Locality-enhanced Compression balances spatial details and efficiency by mining multi-level features in local contexts (Fig. 2 (b)).

1) Locality-Enhanced Downsampling: We downsample MLLM's last-layer visual features $I_v \in \mathbb{R}^{N \times C}$ to $I_v^\downarrow \in \mathbb{R}^{M \times C}$ with our locality-aware resampler, where ρ is downsampling rate. It combines convolution and deformable attention to handle diverse anomaly shapes and sizes. As shown in Fig. 4 (right) (a), it comprises a *Local-Context Learner* to enhance flexibility towards geometric variations, and a *Spatial-Preserving Downsampler* to retain local context while abstracting features. Each learnable query gathers visual cues through a 2D coordinate-based sampling using reference points and sampling offsets:

$$p' = p + \Delta p, \tag{18}$$

$$O(i, j) = \sum_{k=0}^{K-1} W_k * X(i + s * p', j), \tag{19}$$

where s denotes stride and Δp is the learned offset. As illustrated in Fig. 4 (right) (b), Local-Context Learner stacks $\mathcal{L}_c$ blocks, each containing self-attention and deformable-attention layers. The downsampler comprises $\mathcal{L}_c$ ResNet blocks, followed by adaptive average pooling, and another $\mathcal{L}_c$ blocks.

2) Coarse-to-Fine Refinement: The refinement process begins with coarse-to-fine mappings, integrating detailed cues from high-resolution features into the downsampled representation. As the coarse representation $I_v^\downarrow$, ***each pixel in*** $\boldsymbol{I_v^\downarrow} \in \mathbb{R}^{\boldsymbol{1 \times M \times C}}$ ***corresponds to*** $\boldsymbol{\rho \times \rho}$ ***sub-region in*** $\boldsymbol{I_v} \in \mathbb{R}^{\boldsymbol{\rho^2 \times M \times C}}$.

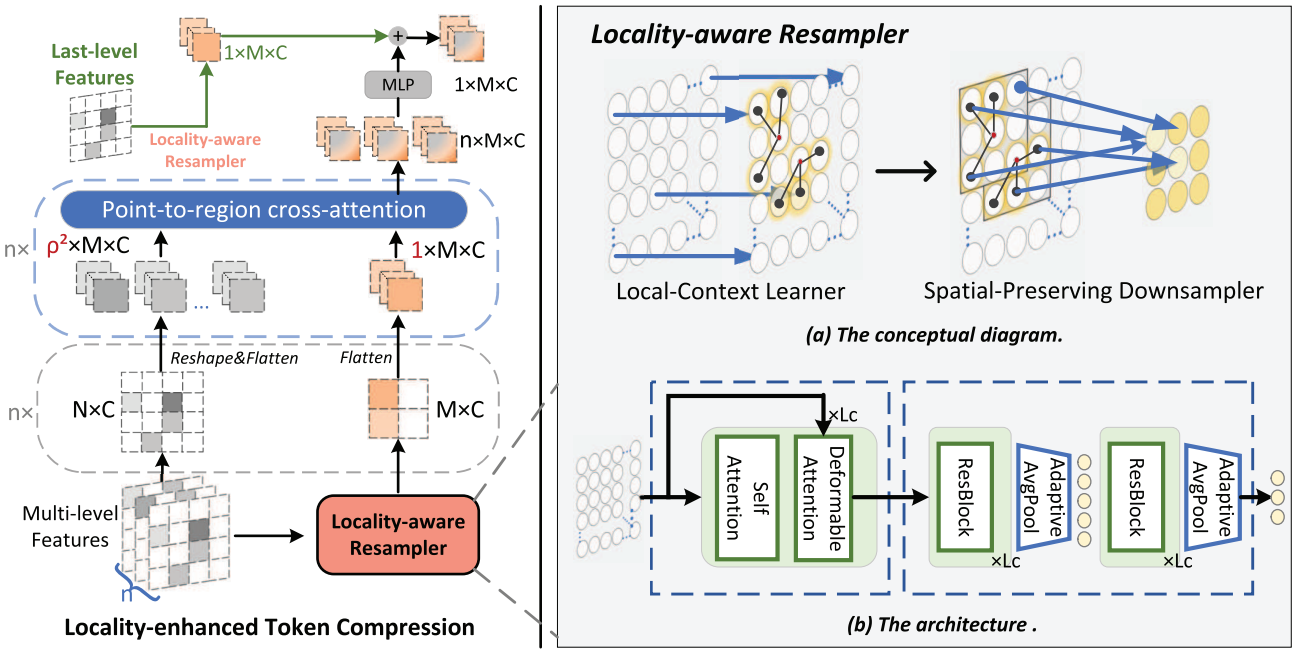


Fig. 4. **Overview of LTC mechanism.** LTC incorporates multi-level visual cues through a coarse-to-fine scheme, preserving comprehensive image information to the LLM.

This mechanism enables low-resolution queries to assimilate fine-grained keys and values. Specifically, we consider the low-resolution $I_v^{\downarrow} \in \mathbb{R}^{1\times M\times C}$ as point-based queries while I_v as region-based keys and values. Formulated as:

$$V_{att,l} = softmax\left(\frac{Q_l^{\downarrow} K_l^T}{\sqrt{d}}\right) V_l, \tag{20}$$

where $Q_l^{\downarrow}$ represents the queries derived from coarse features, K_l, V_l are keys and values from the original ones.

3) Multi-Level Feature Integration: To achieve hierarchical representations, multi-level features are integrated as enriched reference keys and values. This leverages inherent biases of different CLIP encoder layers: shallow layers capture low-level details, while deeper layers encode semantic understanding. Concretely, we incorporate the past n levels:

$$V_{att,all} = Linear(cat([V_{att,L-n-1}, \dots V_{att,L-2}])), \tag{21}$$

where L is the total levels in CLIP encoder. Finally, we enhance last-layer coarse representation with aggregated values:

$$I_v^* = I_v^{\downarrow} + V_{att,all}. \tag{22}$$

LTC maintains rich semantic cues while reducing visual tokens, simultaneously performing coarse-to-fine injection and multi-level feature integration.

IV. Dataset

A. Collection Details

We collect the Real Industrial Anomaly Detection Dataset (RIAD), extracted from various practical industrial datasets including MIAD [44], Realiad [45], Vision [46], MAD-sim [47], MVTec-3D [48], and PKU-market-iphone [49]. RIAD pairs each image with an anomaly mask and text data (Fig. 5 (a)). As illustrated in Fig. 5 (b) (top), RIAD encompasses a rich variety of industrial scenarios, including outdoor environments such as wind power equipment, indoor industrial components, and food products. This diverse range provides a more comprehensive representation of real-world industrial defects. The RIAD contains 11,627 normal and 16,413 defective images across 24 object classes and 15 defect types, providing extensive resources for IAD model development.

B. Annotation Details

We utilize Anylabeling to label defect types. Semantic masks are recorded in COCO format with defect-category IDs, encompassing 15 distinct anomaly categories. Notably, we've transformed original binary (normal/abnormal) masks into semantic masks that distinguish different defect types. Meanwhile, images are also annotated with bounding boxes of anomaly regions. To provide rich, contextual textual data for each anomaly instance, we feed the anomaly images and COCO-format masks to GPT-4 [12], generating three types of question-answer pairs: anomaly descriptions, impact analyses, and suggestions. This comprehensive textual information distinguishes RIAD from existing datasets, offering a holistic resource for IAD models leveraging both visual and textual inputs. Paired examples are shown in Fig. 5 (a).

C. Statistic Analysis

Fig. 5 (b) shows the count and distribution of anomalies in images. The top of Fig. 5 (b) displays image samples, showcasing diverse scenarios from outdoor industrial settings to indoor electronics, ceramics, building blocks, and food

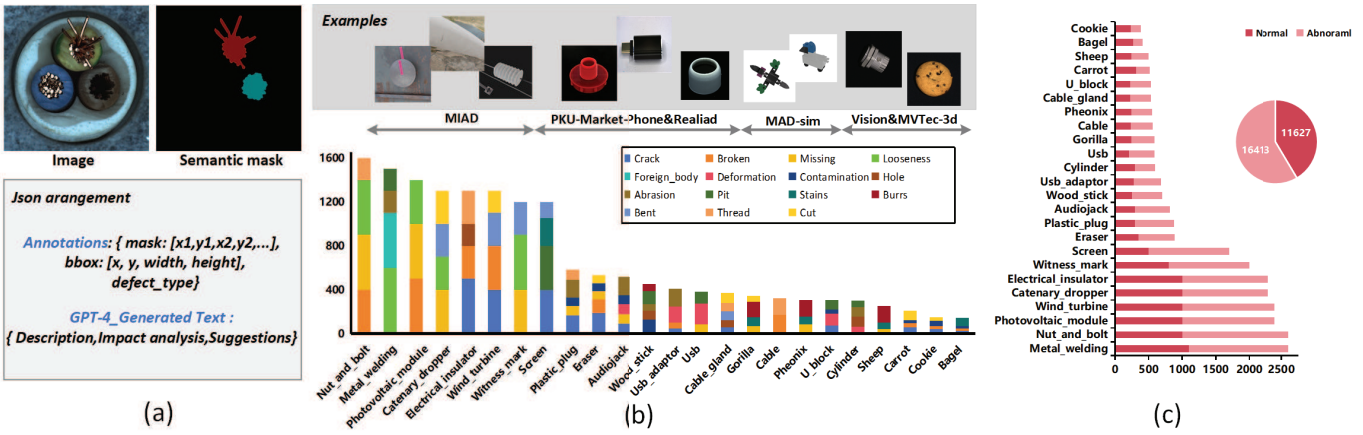


Fig. 5. **Properties of the RIAD dataset.** (a) RIAD data pairs: images, semantic segmentation masks encoding defect types, and GPT-generated text. Masks and text are stored in JSON format. (b) The horizontal axis represents the categories of objects, the vertical axis represents quantity, and different colors represent different defect types. The top part shows sample RIAD images with their source datasets indicated. (c) The ratio of normal images and abnormal images in each object class.

items. Fig. 5 (c) illustrates the ratio of normal images and abnormal images in each object category. In the cross-category setting, we train on 16 classes (15,274 images) and test on 8 unseen classes (8,342 images) during training.

V. Experiments

This section elaborates on the experiments' details. Section V-A.1 presents the experimental protocol. Section V-B analyzes the main results. Section VI demonstrates the ablation study.

A. Experimental Protocol

1) Evaluation Protocols: We evaluate models' adaptability to unseen scenarios using two complementary settings:

- *Cross-dataset Evaluation:* Following [34], [35], we adopt cross-dataset zero-shot anomaly detection, training on RIAD, and evaluating three widely-used public datasets without fine-tuning: **MVTec-AD** [50]: 5,354 high-resolution images across 15 classes. **Visa** [51]: The largest industrial anomaly detection dataset, with 10,821 images across 12 categories of colored industrial parts. **WFDD** [17]: 4,101 images of woven fabrics with different textures and patterns across 4 categories.
- *Cross-category Evaluation:* We randomly split RIAD into non-overlapping subsets: 16 classes (15,274 images) for training and 8 classes (8,342 images) for testing.

2) Training Data Formulation: Our multi-modal training approach incorporates three distinct task types:

- *Anomaly Segmentation Task.* The model outputs only segmentation masks. We employ a question-answer template like "USER: < IMAGE > Are there any abnormalities present in the OBJECT? If so, please output the defect segmentation result. ASSISTANT: It is < SEG >."
- *Anomaly Segmentation and Answering Task.* The model produces both anomaly masks and corresponding textual answers. We employ a question-answer template: "USER: < IMAGE > Are there any abnormalities present in the OBJECT? If so, please output the defect segmentation result and provide the QUERY TYPE. ASSISTANT: It is < SEG >. ANSWERING" Queries fall into three types: description, impact analysis, and suggestion analysis, aiming to provide comprehensive insights for anomalies. The example responses are illustrated in Fig.7
- *Visual Question Answering Task.* To preserve the original Visual Question-answering ability of LLM, we include the VQA task during training. We generate question-answer pairs based on defect descriptions using GPT-4.

3) Baselines: For a comprehensive evaluation, we compare VMAD against 8 typical baseline methods: two MLLM-based anomaly detection models (**Myriad** [35] and **AnomalyGPT** [34]), four zero-shot anomaly detection models (**WinClip** [5], **Musc** [8],**VAND** [43] and **AnomalyClip** [6]), and one supervised anomaly detection model (**BGAD** [42]) and one multi-modal pixel-grounding model (**LISA** [32]). The anomaly detection performance is evaluated using the **A**rea **U**nder the **R**eceiver **O**perating **C**haracteristic Curve (**AUROC**). Additionally, **A**verage **P**recision (**AP**) for **Im**age-level detection (**Img**) and AUPRO for pixel-level segmentation are also included to provide more in-depth analysis.

4) Implementations: We employ LLaVA-7B-v1-1 [29] as our multi-modal LLM and utilize the ViT-H SAM for the visual backbone. We keep the vision encoders and visual backbone frozen. Concurrently, the LLM is fine-tuned using the LoRA [52] technique. The memory bank (Sec. III-C) contains 1/20 of both normal and anomalous samples from each class. We adjust the down-sampling ratio $\rho = 2$ in Sec. III-D, and set $n = 4$. Models are trained for 10 epochs using an AdamW optimizer with a cosine learning rate schedule on an 8×4090 GPU cluster, processing two input pairs per iteration. The batch size is set to 2, and the learning rate is 1e-6. The LoRA rank is configured to 8, and the ROI-patch size is set to 32×32.

B. Results and Analysis

This section compares VMAD with previous IAD works through numerical analysis and visual examples.

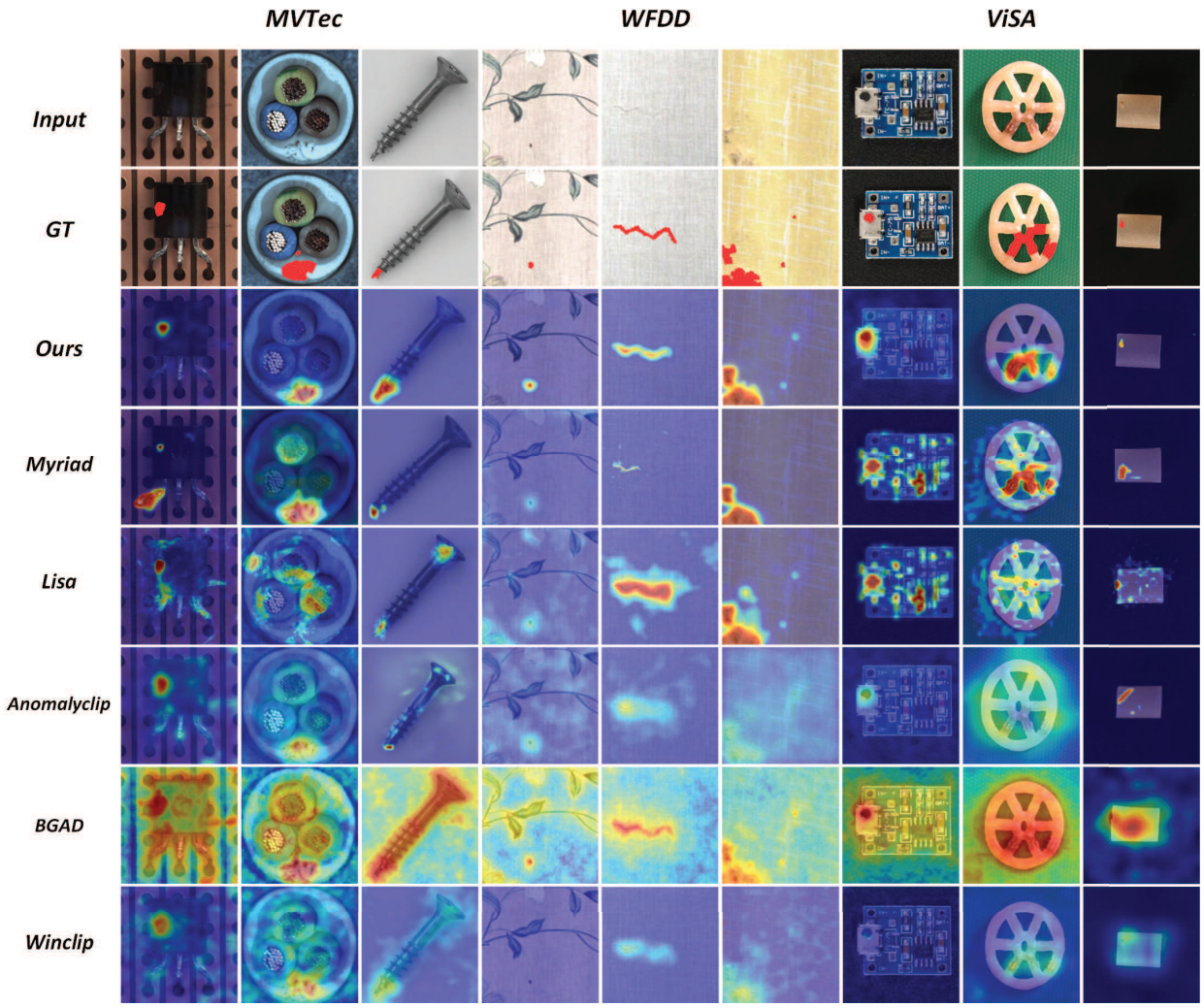


Fig. 6. Zero-shot anomaly segmentation on MVTec-AD, WFDD, and ViSA datasets.

TABLE I

ZERO-SHOT PERFORMANCE ON CROSS-DATASET EVALUATION SETTING ON MVTEC-AD, VISA, AND WFDD DATASETS. THE BEST RESULTS ARE IN WHILE THE SECOND BEST ARE UNDERLINED. 'IMG' AND 'PIXEL' REPRESENT THE MEAN IMAGE-LEVEL AND PIXEL-LEVEL AUROC. ^{O} DENOTES USING THE LAST-LAYER FEATURE INJECTION IN THE LTC MODULE

Methods	MVTec-AD				ViSA				WFDD				Inference Time(s)
	Img	Pixel	PRO	AP	Img	Pixel	PRO	AP	Img	Pixel	PRO	AP	
BGAD [42]	90.1	91.3	87.3	85.6	87.4	88.6	82.0	80.6	95.0	96.2	86.8	88.1	0.57
AnomalyClip [6]	91.5	94.7	86.2	83.4	82.3	87.1	84.2	80.3	94.1	95.3	84.9	83.8	0.37
VAND [43]	91.8	92.5	88.2	86.8	83.5	90.7	81.9	82.8	94.6	96.1	84.3	88.5	0.34
WinClip [5]	92.5	93.4	89.8	87.5	84.8	91.3	80.4	82.4	93.2	96.2	86.1	83.5	0.52
Musc [8]	90.5	92.1	89.2	84.2	85.2	90.4	80.8	81.0	91.7	92.2	87.8	84.1	0.68
LISA [32]	89.1	91.6	82.2	81.2	81.3	85.2	77.8	79.3	92.1	93.5	86.0	87.2	0.97
AnomalyGPT [34]	92.1	93.9	86.2	85.1	86.6	92.4	81.7	75.6	93.7	**96.9**	85.7	84.2	0.89
Myriad [35]	93.2	92.3	87.9	**88.2**	85.9	90.5	81.3	83.5	91.8	93.9	87.3	82.7	0.87
VMAD^{o}	93.9	94.0	89.7	87.2	87.1	92.1	83.2	82.1	94.2	95.4	87.1	89.2	0.75
VMAD	**95.8**	**96.1**	**91.2**	87.6	**89.7**	**93.8**	**84.6**	**85.2**	**95.3**	96.7	**89.2**	**90.1**	0.84

1) Cross-Dataset Evaluation: Table I summarizes the zero-shot results on MVTec-AD, Visa, and WFDD datasets. Tab. I reveals that VMAD outperforms competing methods in the presence of large domain gaps, particularly excels in pixel-level and region-level analyses, with margins of **3.3%**p and **3.8%**p compared to Myriad [35]. Additionally, VMAD achieves the fastest inference time among MLLM-based methods, surpassing Myriad, AnomalyGPT, and LISA by 0.12-0.23s. Notably, leveraging final layer feature injection and optimizing with the LTC, we accelerate inference by 0.09s while maintaining stable performance. This demonstrates that LTC reduces the number of visual tokens while preserving

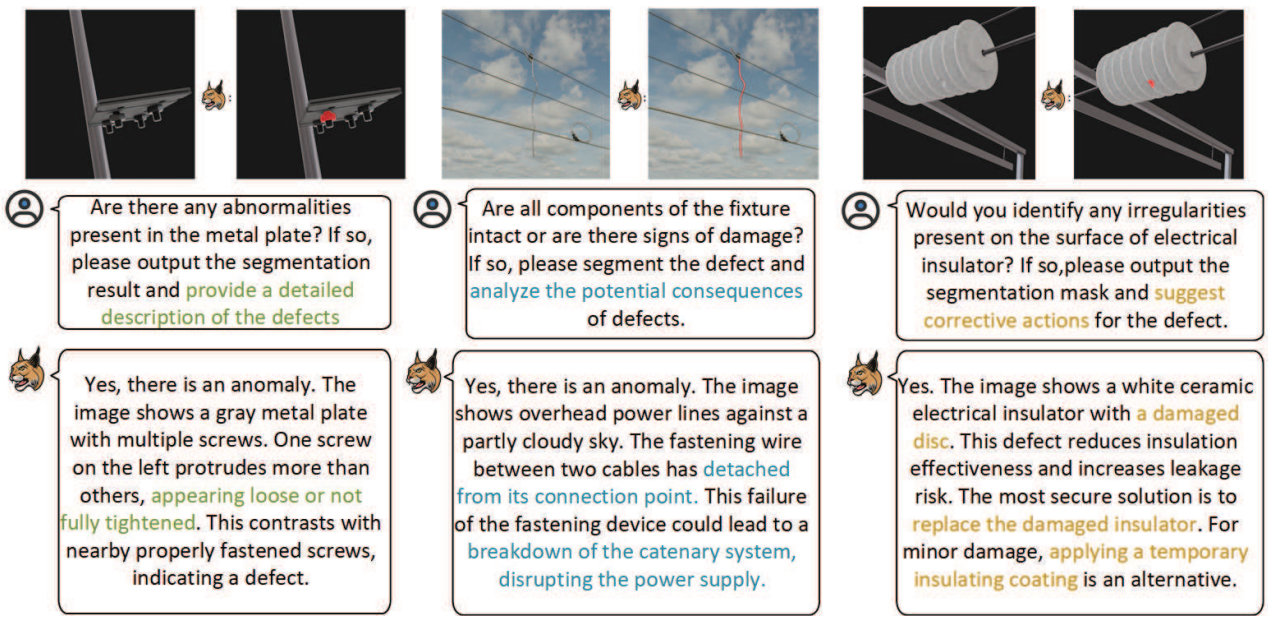


Fig. 7. **Qualitative examples of VMAD in cross-category setting.** For each set, we show the input image (left), VMAD's binary prediction mask (right, thresholded visual output), and VMAD's inspection analysis (bottom). The binary masks visualize VMAD's pixel-level anomaly localization results, while the text demonstrates its capability to provide comprehensive analysis, including image description, anomaly impact assessment, and improvement suggestions.

TABLE II

ZERO-SHOT ANOMALY DETECTION PERFORMANCE ON CROSS-CATEGORY EVALUATION ACROSS EACH CLASS IN RIAD. THE BEST RESULTS ARE IN BOLD WHILE THE SECOND BEST ARE UNDERLINED. 'IMG' AND 'PIXEL' REPRESENT THE MEAN IMAGE-LEVEL AND PIXEL-LEVEL AUROC

Methods	metal_welding			u_block			toy_brick			...	wind_turbine			Average		
	Img	Pixel	PRO	Img	Pixel	PRO	Img	Pixel	PRO		Img	Pixel	PRO	Img	Pixel	PRO
LISA [32]	84.1	90.5	81.0	89.1	91.2	86.6	80.4	82.3	76.1		81.8	89.5	82.7	79.2	82.7	78.9
BGAD [42]	91.6	94.7	87.2	79.1	82.4	77.2	86.2	87.3	85.2		91.8	93.5	87.1	89.2	91.4	87.1
AnomalyClip [6]	90.2	92.2	91.1	87.4	90.7	89.2	89.1	93.7	89.2	...	88.9	90.6	89.1	92.7	96.6	90.5
VAND [43]	96.2	95.4	94.2	91.2	96.3	89.7	91.7	92.3	86.9		78.8	82.6	88.1	92.1	94.7	89.4
WinClip [5]	94.8	96.1	90.8	90.8	95.2	87.6	90.4	91.5	85.3		89.2	94.6	87.9	94.2	93.1	86.1
Musc [8]	87.8	91.2	89.2	86.3	88.9	86.4	88.3	96.2	83.8		87.4	82.2	88.4	90.8	94.2	89.2
AnomalyGPT [34]	90.1	86.2	84.6	83.7	87.1	88.2	89.5	92.4	80.1		78.2	78.1	67.8	87.9	89.4	87.1
Myriad [35]	95.2	93.1	88.1	84.6	84.4	89.4	93.3	95.8	90.8		89.1	81.9	87.4	91.1	92.4	85.2
VMAD	**97.2**	**98.6**	**91.6**	**91.1**	**98.9**	**90.2**	**97.2**	**98.1**	**92.3**		**93.7**	**98.1**	**89.5**	**94.9**	**98.9**	**92.3**

accuracy. Fig. 6 visually displays the anomaly localization results, highlighting VMAD's improved segmentation precision.

2) Cross-Category Evaluation: Tab. II details the performance, demonstrating state-of-the-art results of VMAD. Notably, our model excels in pixel-level analyses, achieving a substantial **2.0%∼14.5%** AUC-PR improvement over other multimodal IAD methods. This underscores VMAD's exceptional capability in anomaly localization. Fig. 8 shows visualization results, while Fig. 7 displays VMAD's dialogue and segmentation on RIAD. Beyond precise localization, VMAD demonstrates remarkable zero-shot understanding and reasoning in anomaly detection, offering potential for production guidance through impact analysis and suggestions.

VI. ABLATION STUDY

In this section, we conduct a comprehensive ablation study to validate the contributions of major modules. All experiments are performed under two settings: cross-dataset setting on MVTec-AD or cross-category setting on RIAD.

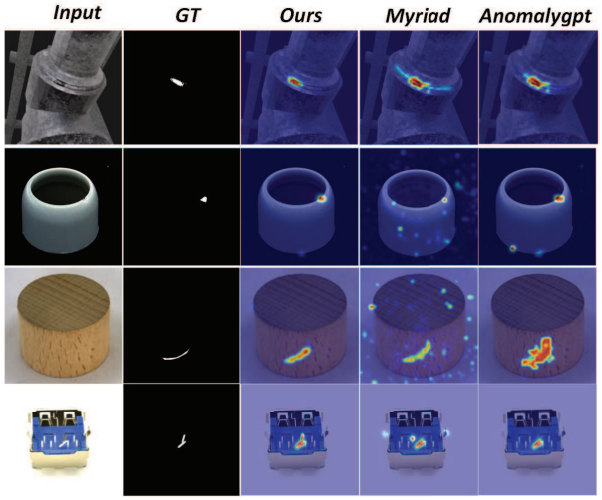


Fig. 8. Comparative visualization of zero-shot anomaly segmentation on RIAD dataset.

We report the experimental results of selectively activating modules in Tab. III. It reveals that activating DSSL alone

TABLE III
COMPONENT-WISE EXPERIMENTAL RESULTS

LTC	DSSL	MVTec			RIAD		
		Img	Pixel	PRO	Img	Pixel	PRO
✗	✗	90.5	91.3	87.3	91.3	93.4	89.5
✗	✓	93.6	94.3	89.1	92.1	94.5	90.8
✓	✗	92.9	93.7	90.5	93.2	96.1	91.5
✓	✓	**95.8**	**96.1**	**91.2**	**94.9**	**98.9**	**92.3**

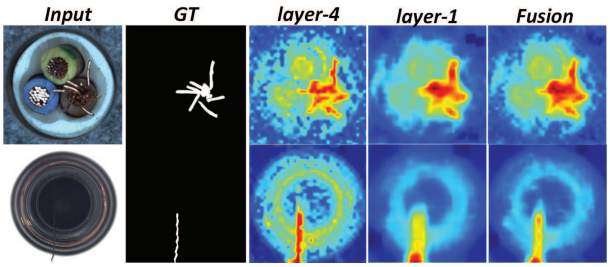


Fig. 9. LTC Multi-Level Feature Integration: attention maps from layer-1 (last) and layer-4 (fourth-to-last), and their fused result. See Sec. VI-A.2 for ablation study.

TABLE IV
ABLATION STUDY OF SAMPLED-PATCH NUMBER N_b ON MVTEC RIAD AND VISA

N_b	MVTec			RIAD			Visa		
	Img	Pixel	PRO	Img	Pixel	PRO	Img	Pixel	PRO
4	92.5	93.3	87.3	91.3	93.4	89.5	85.2	90.3	83.4
8	93.2	94.1	88.5	92.1	94.2	90.2	86.1	91.2	84.3
16	94.0	94.8	89.0	92.9	95.9	90.7	87.9	92.7	85.1
32	95.8	96.1	**91.2**	94.9	**98.9**	92.3	**89.7**	**93.8**	**84.6**
64	96.1	96.3	91.0	95.5	98.3	92.6	88.8	92.9	84.0
128	**96.4**	**96.7**	90.7	**95.7**	98.5	93.5	89.2	93.1	84.2
256	96.2	95.5	90.4	95.4	98.4	**93.8**	89.5	93.0	84.1

TABLE V
ABLATION STUDY ON DIFFERENT VISUAL PROJECTORS ON MVTEC DATASET. WE ADOPT TOKENS PER SECOND (TPS) TO EVALUATE THE THROUGHPUT OF LLM DURING INFERENCE

Method	#Tokens	#TPS	Img	Pixel	PRO
Baseline	144	16.9	92.9	94.3	89.1
+ *Injection*	144	–	93.5	95.4	90.9
+ *Multi-level Feature*	144	–	93.8	95.9	91.0
LTC(Ours)	114	26.8	95.8	96.1	91.2
c *MLP*	**576**	5.5	91.4 ↓1.5	92.6 ↓1.7	87.2 ↓1.9
c *Average-Pooling*	144	31.6	**94.9** ↑2.0	**95.6** ↑1.3	**91.9** ↑2.8
c *Resampler*	144	21.7	91.7 ↓1.2	92.4 ↓1.9	87.6 ↓1.5
c *LDP-v2*	144	22.8	91.5 ↓1.4	92.5 ↓1.8	87.8 ↓1.3

improves RIAD's AUROC by **3.1%**p (image-level) and **3.0%**p (pixel-level). LTC enhances MVTec's performance by **2.7%**p (pixel-level AUROC) and **2.0%**p (region-level AURPRO). MVTec's single-object, fine-grained defects benefit more from LTC. Conversely, RIAD's complex scenarios show greater improvements with DSSL's multi-modal semantic learning.

Tab. V presents the LTC's ablation results on MVTec. Adding (**darkgreen+**) the injection module obtains **+0.6**%, **+1.1**% and **+1.8**% gains over the baseline method, respectively. Fig. 9 demonstrates the effectiveness of multi-level feature fusion. Our LTC surpasses the previous best method LDP-v2 [53] by **+1.4%**, **1.8%** and **1.3%** LTC achieves a notable TPS improvement over the baseline (26.8 vs. 16.9), allowing the model to process more tokens within the same timeframe. While MLP supports more tokens, its TPS is limited to 5.5, indicating the inefficiency of visual token processing. In contrast, LTC effectively balances spatial structure with computational efficiency.

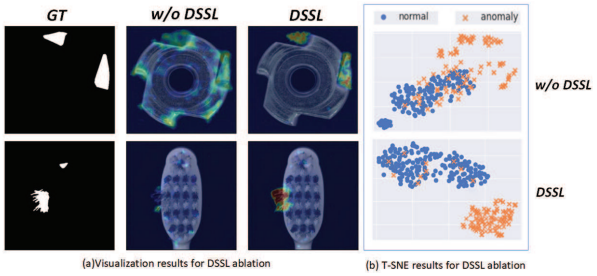


Fig. 10. **Ablation study on Defect-Sensitive Structure Learning.** (Left) Visualization with and without DSSL. (Right) t-SNE plot of patch tokens. 'w/o' denotes without.

Fig. 10 presents the ablation study of Defect-Sensitive Structure Learning module. Left: Visualization demonstrates DSSL's enhanced sensitivity to anomalous structures. Right: t-SNE plot of normalized features $\Psi_{I\to T}(z)$ in LLM space (Eq. 13) shows improved discrimination between normal and anomalous samples through DSSL's patch-based similarity learning. Tab. IV displays the ablation study of Sampled-patch Number N_b on MVTec RIAD and Visa. Increasing the number of sample patches N_b boosts model performance on MVTec, RIAD, and Visa datasets, enhancing detail capture and detection accuracy. However, beyond $N_b = 32$, gains are marginal and sometimes decline, possibly due to redundant information affecting generalization. At $N_b = 32$, the model hits peak performance efficiently, making it the optimal choice for balancing effectiveness and computational cost.

VII. LIMITATIONS

While VMAD shows promising results in structural defect detection, it faces significant limitations in logical anomaly detection, characterized by high false positive rates. This limitation primarily stems from insufficient visual reasoning capabilities, where the model struggles to effectively integrate image features during inference. Additionally, the SFT-guided text generation performs well in fixed scenarios but lacks generalizability in dynamic contexts, affecting the quality of reasoning chains. Future work should explore reinforcement learning post-training to boost visual reasoning for both structural and logical anomaly detection [54], [55].

VIII. REPRODUCIBILITY STATEMENT

To promote reproducibility and facilitate future research, we will publicly release the implementation code at https://github.com/denghuilin-cyber/VMAD, and the proposed RIAD dataset at https://drive.google.com/drive/u/1/folders/1dnPNsho7s6ZC3pJ3XlfGgEL3DRoqE15Y

IX. Conclusion

In this paper, we present VMAD, a visual-enhanced MLLM framework for zero-shot anomaly detection, enabling explainable analysis of unseen defects. Our contributions include: (1) a cross-modal learning scheme that transfers patch-based visual similarity to LLM space, and (2) a locality-enhanced token compression mechanism for fine-grained defect detection. We also introduce RIAD, a comprehensive industrial anomaly dataset supporting multiple tasks, including VQA, segmentation, and reasoning. Extensive experiments demonstrate VMAD's robust performance across public and real-world datasets. We believe our approach opens new avenues for interpretable and flexible industrial inspection systems. Future work includes exploring multimodal incremental learning for dynamic production environments.

References

[1] Q. Yu, K. Zhu, Y. Cao, F. Xia, and Y. Kang, "TF^2: Few-shot text-free training-free defect image generation for industrial anomaly inspection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 11825–11837, Nov. 2024.

[2] X. Deng and J. Li, "Deep one-class classification model assisted by radius constraint for anomaly detection of industrial control systems," *Eng. Appl. Artif. Intell.*, vol. 138, Dec. 2024, Art. no. 109357.

[3] M. Yang, J. Liu, Z. Yang, and Z. Wu, "SLSG: Industrial image anomaly detection with improved feature embeddings and one-class classification," *Pattern Recognit.*, vol. 156, Dec. 2024, Art. no. 110862.

[4] A. Radford et al., "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[5] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19606–19616.

[6] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," 2023, *arXiv:2310.18961*.

[7] S. Li, J. Cao, P. Ye, Y. Ding, C. Tu, and T. Chen, "ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation," 2024, *arXiv:2401.12665*.

[8] X. Li, Z. Huang, F. Xue, and Y. Zhou, "MuSc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images," 2024, *arXiv:2401.16753*.

[9] S. Xuan and S. Zhang, "Decoupled contrastive learning for long-tailed recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 6396–6403.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, May 2022, pp. 16816–16825.

[11] L. Meng et al., "DeepStack: Deeply stacking visual tokens is surprisingly simple and effective for LMMs," 2024, *arXiv:2406.04334*.

[12] Z. Chen et al., "How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites," 2024, *arXiv:2404.16821*.

[13] W. Li et al., "TokenPacker: Efficient visual projector for multimodal LLM," 2024, *arXiv:2407.02392*.

[14] J. Guo, S. Lu, L. Jia, W. Zhang, and H. Li, "ReContrast: Domain-specific anomaly detection via contrastive reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 10721–10740.

[15] J. Jiang et al., "Masked Swin transformer unet for industrial anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2200–2209, Feb. 2023.

[16] H. Deng, H. Luo, W. Zhai, Y. Guo, Y. Cao, and Y. Kang, "Prioritized local matching network for cross-category few-shot anomaly detection," *IEEE Trans. Artif. Intell.*, vol. 5, no. 9, pp. 4550–4561, Sep. 2024.

[17] Q. Chen, H. Luo, C. Lv, and Z. Zhang, "A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization," 2024, *arXiv:2407.09359*.

[18] Y. Huo, X. Cheng, S. Lin, M. Zhang, and H. Wang, "Memory-augmented autoencoder with adaptive reconstruction and sample attribution mining for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–18, 2024.

[19] W. Jiang, K. Yang, C. Qiu, and L. Xie, "Memory enhancement method based on skip-GANomaly for anomaly detection," *Multimedia Tools Appl.*, vol. 83, no. 7, pp. 19501–19516, Jul. 2023.

[20] Z. Yu, C. Zhang, and L. Sun, "KDSeg: Segmentation guided knowledge distillation for anomaly detection," in *Proc. 4th Int. Symp. Comput. Technol. Inf. Sci. (ISCTIS)*, Jul. 2024, pp. 815–820.

[21] Y. Zang, A. Lu, B. Li, and W. Hu, "Revisiting segmentation-guided denoising student–teacher in anomaly detection," *Vis. Comput.*, vol. 40, no. 6, pp. 4023–4038, 2024.

[22] H. Hojjati, T. K. K. Ho, and N. Armanfard, "Self-supervised anomaly detection in computer vision and beyond: A survey and outlook," *Neural Netw.*, vol. 172, Apr. 2024, Art. no. 106106.

[23] W. Zhai, Y. Cao, J. Zhang, H. Xie, D. Tao, and Z.-J. Zha, "On exploring multiplicity of primitives and attributes for texture recognition in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 403–420, Jan. 2024.

[24] H. Deng, Y. Guo, Z. Xu, and Y. Kang, "PTMNet: Pixel-text matching network for zero-shot anomaly detection," in *Proc. 9th Int. Conf. Big Data Inf. Analytics (BigDIA)*, Dec. 2023, pp. 781–787.

[25] C. Li, S. Zhou, J. Kong, L. Qi, and H. Xue, "KAnoCLIP: Zero-shot anomaly detection through knowledge-driven prompt learning and enhanced cross-modal integration," 2025, *arXiv:2501.03786*.

[26] Z. Gu et al., "FiLo: Zero-shot anomaly detection by fine-grained description and high-quality localization," in *Proc. 32nd ACM Int. Conf. Multimedia*, Feb. 2024, pp. 2041–2049.

[27] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Apr. 2022, pp. 23716–23736.

[28] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2023, pp. 19730–19742.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 34892–34916.

[30] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.

[31] K. You et al., "Ferret-UI: Grounded mobile UI understanding with multimodal LLMs," 2024, *arXiv:2404.05719*.

[32] X. Lai et al., "LISA: Reasoning segmentation via large language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 9579–9589.

[33] H. Rasheed et al., "GLaMM: Pixel grounding large multimodal model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13009–13018.

[34] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting industrial anomalies using large vision-language models," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 3, pp. 1932–1940.

[35] Y. Li et al., "Myriad: Large multimodal model by applying vision experts for industrial anomaly detection," 2023, *arXiv:2310.19070*.

[36] J. Zhu, S. Cai, F. Deng, B. C. Ooi, and J. Wu, "Do LLMs understand visual anomalies? Uncovering LLM's capabilities in zero-shot anomaly detection," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 48–57.

[37] Y.-Q. Yu, M. Liao, J. Wu, Y. Liao, X. Zheng, and W. Zeng, "TextHawk: Exploring efficient fine-grained perception of multimodal large language models," 2024, *arXiv:2404.09204*.

[38] X. Dong et al., "InternLM-XComposer2–4KHD: A pioneering large vision-language model handling resolutions from 336 pixels to 4K HD," 2024, *arXiv:2404.06512*.

[39] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, Apr. 2021, pp. 8748–8763.

[40] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[41] N. Pu, Z. Zhong, and N. Sebe, "Dynamic conceptional contrastive learning for generalized category discovery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7579–7588.

[42] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang, "Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection," 2022, *arXiv:2207.01463*.

[43] X. Chen, Y. Han, and J. Zhang, "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," 2023, *arXiv:2305.17382*.

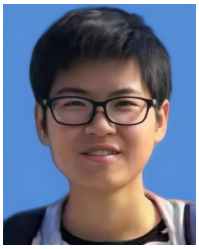
[44] T. Bao et al., "MIAD: A maintenance inspection dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Paris, France, Oct. 2023, pp. 993–1002.
[45] C. Wang et al., "Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 22883–22892.
[46] H. Bai et al., "VISION datasets: A benchmark for vision-based industrial inspection," 2023, *arXiv:2306.07890*.
[47] Q. Zhou et al., "PAD: A dataset and benchmark for pose-agnostic anomaly detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 44558–44571.
[48] P. Bergmann, X. Jin, D. Sattlegger, and C. Steger, "The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization," 2021, *arXiv:2112.09045*.
[49] J. Zhang, R. Ding, M. Ban, and T. Guo, "FDSNeT: An accurate real-time surface defect segmentation network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3803–3807.
[50] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9584–9592.
[51] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 392–408.
[52] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.
[53] X. Chu et al., "MobileVLM V2: Faster and stronger baseline for vision language model," 2024, *arXiv:2402.03766*.
[54] H. Deng, D. Zou, R. Ma, H. Luo, Y. Cao, and Y. Kang, "Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning," 2025, *arXiv:2503.07065*.
[55] X. Wang et al., "ViCrit: A verifiable reinforcement learning proxy task for visual perception in VLMs," 2025, *arXiv:2506.10128*.

Wei Zhai received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2022. He is a Post-Doctoral Fellow with the School of Information Science and Technology, USTC. His research interests mainly include computer vision and deep learning. He has published more than 20 articles in these areas with a series of publications in top journals and conferences, such as CVPR, ICCV, and NeurIPS.

Yanming Guo received the B.S. and M.S. degrees from the National University of Defense Technology in 2011 and 2013, respectively, and the Ph.D. degree from Leiden Institute of Advanced Computer Science (LIACS), Leiden University, in 2017. He is an Associate Professor with the College of Systems Engineering, National University of Defense Technology. His current interests include computer vision, natural language processing, and deep learning.

Huilin Deng was born in 2000. She received the B.E. degree in information engineering from Nanjing University of Science and Technology, Nanjing, China, in 2021. She is currently pursuing the Ph.D. degree in control science and engineering with the Department of Automation, University of Science and Technology of China, Hefei, China. Her current research interests include machine learning and computer vision.

Yang Cao (Member, IEEE) received the B.S. and Ph.D. degrees in information engineering from Northeastern University, Shenyang, China, in 1999 and 2004, respectively. Since 2004, he has been with the Department of Automation, University of Science and Technology of China, Hefei, China, where he is currently an Associate Professor. His current research interests include machine learning and computer vision.



Hongchen Luo received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2019, where she is currently pursuing the Ph.D. degree in control science and engineering with the Department of Automation. She is an Associate Professor with the College of Information Science and Engineering, Northeastern University. She has published several papers in conferences and journals, such as CVPR, ICCV, IJCAI, IJCV, and IEEE TRANSACTIONS ON ARTIFICIAL INTELLIGENCE. Her research interests mainly include computer vision and deep learning.

Yu Kang (Senior Member, IEEE) received the Dr.Eng. degree in control theory and control engineering from the University of Science and Technology of China, Hefei, China, in 2005. From 2005 to 2007, he was a Post-Doctoral Fellow with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China. He is currently a Professor with the Department of Automation, University of Science and Technology of China.