SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2026, Vol. 69, Iss. 9, 192103:1–192103:14
https://doi.org/10.1007/s11432-025-4976-0

Myriad: a large multimodal model applying vision experts for industrial anomaly detection

Yuanze LI[1,2†], Haolin WANG[1,2†], Shihao YUAN[1,2], Ming LIU[1,2*], Debin ZHAO[1], Yiwen GUO[3], Chen XU[2], Guangming SHI[2] & Wangmeng ZUO[1,2]

[1]*Faculty of Computing, Harbin Institute of Technology, Harbin 15001, China*
[2]*Pazhou Lab Huangpu, Guangzhou 510555, China*
[3]*Independent researcher*

Received 14 January 2025/Revised 13 September 2025/Accepted 2 November 2025/Published online 13 August 2026

Abstract Industrial anomaly detection (IAD) has developed strong methods for one-class, zero-shot, and few-shot deployments, each effective in its own setting, yet no single approach adapts quickly across them. Large multimodal models (LMMs) offer a different route, since broad knowledge and strong visual-language understanding allow an LMM to interpret the outputs of diverse IAD methods and to calibrate their predictions using image evidence. To realize such an adaptive system, we present a novel large multimodal model applying vision experts for industrial anomaly detection (abbreviated as Myriad). Myriad treats conventional IAD models as VEs and converts their anomaly maps into lightweight prompts that steer a frozen Q-Former toward suspicious regions, while a compact low-rank adapter shapes features for IAD. The language pathway then fuses VE positional cues with visual evidence to produce calibrated, machine-usable decisions, effectively regularizing noisy or ambiguous anomaly maps. By simply switching the VE (e.g., one-class or zero-/few-shot experts) without modifying the architecture, Myriad adapts uniformly across deployment scenarios and remains robust to the choice of expert. Extensive experiments on MVTec-AD, VisA, and PCB Bank benchmarks demonstrate that our proposed method not only performs favorably against state-of-the-art methods under one-class and few-shot settings, but also inherits the flexibility and instruction-following ability of LMMs in the field of IAD. Source code and pre-trained models are publicly available at https://github.com/tzjtatata/Myriad.

Keywords anomaly detection, large multimodal model, vision expert, zero-shot, few-shot

Citation Li Y Z, Wang H L, Yuan S H, et al. Myriad: a large multimodal model applying vision experts for industrial anomaly detection. Sci China Inf Sci, 2026, 69(9): 192103, https://doi.org/10.1007/s11432-025-4976-0

1 Introduction

Industrial anomaly detection (IAD) focuses on identifying and localizing manufacturing defects, playing a critical role in ensuring reliable, uninterrupted industrial processes. By automatically detecting anomalies, IAD enables timely intervention, efficient maintenance, and process optimization, ultimately boosting productivity and minimizing downtime. In recent years, substantial research attention has been devoted to IAD, resulting in the release of numerous datasets [1–5] and methods [6–13].

Despite steady progress, research on IAD has splintered across deployment settings, such as one-class, zero-shot, and few-shot detection, each demonstrating advantages within its own scope. Nevertheless, an open challenge remains, i.e., how to incorporate these capabilities into a single model that can seamlessly adapt across scenarios. Large multimodal models (LMMs) offer a promising direction toward this goal. With their strong perception and reasoning abilities, LMMs can potentially combine generalizable knowledge with expert-derived visual cues, paving the way for a unified framework for IAD.

However, applying LMMs directly to anomaly detection is far from straightforward. On the one hand, existing LMMs are trained for general purposes, leaving their anomaly-related knowledge unlinked to the vision modality (see Subsection 3.1 for details). On the other hand, the scarcity of paired vision-language data in IAD makes full fine-tuning impractical. Moreover, anomaly maps produced by conventional IAD methods, while informative, are noisy and lack semantic clarity. For example, background scratches or subtle color shifts may be incorrectly highlighted as defects. Without additional regularization, these cues can lead to unstable predictions.

* Corresponding author (email: csmliu@outlook.com)
† These authors contributed equally to this work.

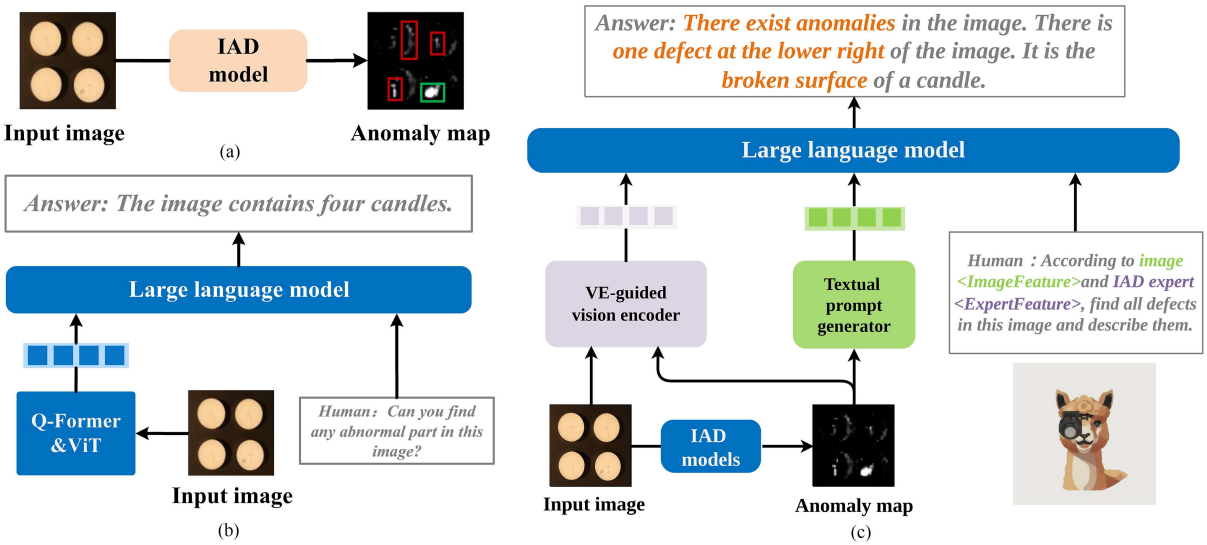

Figure 1 (Color online) Comparison between traditional IAD and LMM-based IAD. (a) Traditional methods output anomaly maps and global scores that are threshold-sensitive and can mis-highlight background, making it hard to decide whether and where defects exist. Green boxes mark ground-truth defects; red boxes mark overkill (false positives). (b) Off-the-shelf LMMs (e.g., MiniGPT-4) do not naturally generalize to IAD because their domain knowledge is weakly grounded in the visual modality. (c) Our Myriad couples pre-trained IAD models as vision experts to enhance visual modality and provide textual cues. The LMM then validates and refines these cues to produce standardized, explainable decisions and adapts across scenarios with plug-and-play experts.

To address these challenges, we propose Myriad, a large multimodal model applying vision experts for industrial anomaly detection. As shown in Figure 1, Myriad treats existing IAD methods as vision experts (VEs) and couples their anomaly cues with the broad knowledge embedded in LMMs, thereby bridging the domain gap and providing a unified solution across one-class, zero-shot, and few-shot scenarios. To begin with, following MiniGPT-4 [14], we stack the pre-trained ViT backbone from EVA-CLIP [15] and the Q-Former from BLIP-2 [16] to form a fundamental vision module. On this basis, we construct a VE-guided vision encoder that produces IAD-suitable visual features under the anomaly maps generated by vision experts. To adapt these features for IAD, we deploy a low-rank residual adapter (LoRRA) between the ViT backbone and the Q-Former. For leveraging the auxiliary visual information from the anomaly maps, we introduce a trainable visual prompt generator (VPG) that derives a set of IAD expert query tokens. These tokens extend the Q-Former's pre-trained query tokens, which remain frozen (along with the ViT and Q-Former parameters) to avoid overfitting the limited training data. By incorporating the anomaly maps, the Q-Former is guided to pay more attention to regions with higher anomaly scores. This strategy compensates for the lack of IAD-specific knowledge in pre-trained language-vision models (LVMs), ultimately enhancing anomaly detection performance.

Beyond the VE-guided visual pathway, we introduce lightweight auxiliary cues to the language pathway. In particular, a textual prompt generator (TPG) converts anomaly maps into coarse positional and extent descriptions, complementing visual features that may overlook such information. By jointly leveraging VE-guided features and these auxiliary cues, Myriad can consistently integrate outputs from different types of vision experts, enabling a single trained model to operate across one-class, zero-shot, and few-shot deployments without architectural modifications.

To evaluate Myriad, we conduct extensive evaluations on the MVTec-AD [1], VisA [2], and PCB Bank [17] datasets. The experimental results show that Myriad not only outperforms both domain-specific vision experts and state-of-the-art industrial anomaly detection methods, but also inherits the flexibility and instruction-following abilities of LMMs in the context of IAD. Qualitative results reveal that Myriad can follow human instructions to generate detailed descriptions of defects and discuss potential causes. Our contributions can be summarized as follows.

- A new framework, Myriad, is proposed for IAD. By introducing existing IAD models as vision experts into LMMs, Myriad effectively stimulates the IAD-relevant textual domain knowledge in LMMs and is adapted to the anomaly detection task. Besides, Myriad inherits the flexibility and instruction-following ability of LMMs and can produce comprehensive descriptions of anomalies following diverse human instructions.
- For extracting proper visual features and bridging domain gaps between general-purpose LMMs and anomaly detection tasks, we design a vision expert-guided vision encoder. In particular, a visual query generated from the

vision expert guidance is embedded into the Q-Former, and a LoRRA is also deployed to coordinate with the visual feature modulation process.

• Extensive experiments show that our proposed Myriad can appropriately receive prior knowledge from vision experts and outperforms state-of-the-art methods on MVTec-AD [1], VisA [2], and PCB Bank [17] datasets.

2 Related work

2.1 Industrial anomaly detection

Given industrial RGB or grayscale images, IAD methods aim to determine potential anomalies and their location. Feature embedding-based methods [6–8, 10, 18] have made great progress recently. Some methods [7, 8, 18] extract the features of normal samples with pre-trained models [19–21] and construct a memory bank, which is used for comparison to detect anomalies during the inference phase. Under the guidance of pre-trained models, some student networks [10, 22, 23] can also be learned to represent the normal sample space, and anomalies are then predicted by comparing characteristic representations between the student networks and the pre-trained networks. Though effective, pre-trained models may suffer from domain biases when being applied to industrial images. SimpleNet [6] addresses this issue by mapping the pre-trained feature space to a domain-specific feature space using a fully connected layer. Additionally, a binary discriminator is trained to detect outliers.

Apart from feature embedding-based methods, there have been other explorations toward more general and effective industrial anomaly detection in recent years. To eliminate the cost of training individual models for each anomaly type [6, 8, 11, 13], some methods [24–26] propose using a unified model to achieve multiple class anomaly detection. Reconstruction-based methods [11–13, 27] aim to reconstruct normal images from samples in the training phase. In the inference phase, the anomaly maps can be predicted by comparing the original and reconstructed images pixel by pixel. Language-guided methods [9, 28, 29] leverage the domain alignment abilities of pre-trained multimodal models, such as CLIP [30], RegionCLIP [31], and ImageBind [32]. For example, WinCLIP [28] and AprilGAN [9] construct two sets of linguistic prompts for abnormal and normal samples, respectively. Given an industrial image, they extract visual features of the image, which are compared against the textual features of the normal and abnormal prompts at each spatial position. Since the visual and textual domains are well aligned, such a comparison yields the pixel-wise anomaly scores. AnoVL [29] further designs test-time adaptation (TTA) to refine features and improve anomaly localization.

However, these IAD methods can only generate image- or pixel-level anomaly scores without deterministic detection, and they lack the flexibility to adapt to different practical scenarios (one-class and zero-/few-shot). To solve these issues, we propose incorporating large vision-language models for a unified industrial anomaly detection model.

2.2 Large multimodal models

Following the success of LLMs, researchers have explored transferring these capabilities to visual perception [14, 16, 33–38]. BLIP-2 [16] employs an image encoder to extract visual features and feed them into LLMs with text prompts. LLaVA [36] and Mini-GPT4 [14] further establish image-text alignment followed by instruction tuning. Shikra [34] and Kosmos-2 [33] enhance grounding via coordinate-based or token-based region referencing. More recently, multimodal studies such as the survey on multimodal referring segmentation [39], MOSEv2 [40, 41], and MeViS [42] highlight a shift toward fine-grained, semantics-aware cross-modal grounding, particularly in complex or dynamic scenes. These studies underscore the increasing importance of precise language-guided localization, which conceptually aligns with our goal of enabling LMMs to reason over localized visual evidence, though their tasks focus on supervised referring segmentation rather than anomaly-centric visual understanding. In the IAD domain, AnomalyGPT [43] introduces a language-guided image decoder to produce anomaly maps and a prompt learner to derive expert prompts for LLMs. However, it mainly relies on global visual features and does not effectively exploit fine-grained local representations. By contrast, our work leverages vision experts as priors to capture local anomaly-related cues, augments them with positional information from a textual prompt generator, and integrates these signals through the reasoning ability of LMMs, thereby enabling a unified model for one-class, zero-shot, and few-shot scenarios.

Table 1 Comparison between LMM-based and other existing methods of industrial anomaly detection across various functionalities.

Method	Inference efficiency	Flexibility	Instruction-following	Analysis and feedback
AOI	Fast	By manual efforts	✗	Predefined items
Learning-based	Medium	By model tuning/memory update	✗	✗
LMM-based	Slow	Ready	✓	✓

3 Method

In this section, we first discuss the paradigm of anomaly detection in Subsection 3.1, then introduce the architecture of Myriad in Subsection 3.2, which consists of the vision expert-guided vision encoder (Subsection 3.3) and the vision expert guidance adapter (Subsection 3.4). Finally, the learning objective and details about IAD instruction are presented in Subsections 3.5 and 3.6.

3.1 Discussion on anomaly detection paradigm

Automated optical inspection (AOI). Over the past decades, AOI has been the dominant technique for industrial anomaly detection. With tens of years of development, the AOI systems are highly optimized and efficient, which are suitable for mass production in modern industry. However, for a different product, even when only minor changes are applied to the product, the hyperparameters and configurations should be manually adjusted, making it extremely cumbersome and expensive, and they can only inspect anomaly types with predefined rules. As a result, the AOI systems are typically deployed in high-end manufacturing, and they are gradually struggling to satisfy the requirements of rapid product iteration and constantly changing demands.

Learning-based methods. Therefore, recent efforts have been directed toward learning-based methods to leverage the power of data-driven optimization schemes, and two main technical routes have been applied in the literature. The first category is to perform IAD directly, where the model predicts the results solely from the test sample [6, 9] or compares the features of certain normal samples against the test sample [2, 7–9]. The second category reconstructs a "normal" counterpart from the test sample via generative models for comparison [11–13, 24]. Nonetheless, as shown in Table 1, the abilities of these methods (e.g., flexibility and instruction-following capability) are still unsatisfactory.

Large multimodal model-based method. In real-world deployment scenarios, the demands constantly change due to product iteration, design updates, standard changes, etc. To satisfy these requirements, AOI and existing learning-based methods require extensive manual efforts and model tuning, as shown in Table 1. Furthermore, the requirements and queries may be open-ended, which may not be achieved by simply adjusting the discrimination principles or fine-tuning the models. To this end, we propose leveraging the eminent perception and comprehension abilities of LMMs.

As further shown in Figure 2, we find that even if not specifically designed for anomaly detection, the pre-trained language models utilized in the LMMs can still obtain sufficient professional knowledge about anomaly detection from the large-scale training corpora, including the causes, characteristics, and impacts of the industrial anomalies. However, such knowledge exists in the language modality only, and the general-purpose LMMs are unable to associate the knowledge with the vision modality due to the limited pairwise vision-language training data for IAD. In this paper, we identify the lack of specialized IAD-related knowledge as the main factor hindering the application of LMMs for IAD, and propose addressing this issue by introducing existing IAD models as vision experts and fusing the vision expert guidance into LMMs.

3.2 Model architecture

To begin with, we build the LMM framework following MiniGPT-4 [14], which comprises two modalities (i.e., vision and language). For the vision modality, a pre-trained ViT backbone is taken from EVA-CLIP [15], followed by the query transformer (Q-Former) from the pre-trained BLIP-2 [16]. For the language modality, we directly use the pre-trained Vicuna [44] model. As analyzed in Subsection 3.1, these models are trained from general-domain training data, and the LMM framework they combine lacks IAD-related knowledge, especially in the vision modality.

Considering that the language model holds a wealth of IAD-related knowledge, to enhance the IAD capability of the LMM framework, we focus on adjusting the vision modality and the vision-to-language knowledge transmission. For the former, we leverage existing IAD methods as vision experts, whose outputs (i.e., the anomaly map) are utilized to guide the feature extraction from the industrial image $\mathbf{I}$ in the vision modality. The proposed VE-guided vision encoder not only extracts IAD-suitable visual features (denoted by $\mathbf{t}_v$) under the guidance of vision experts,

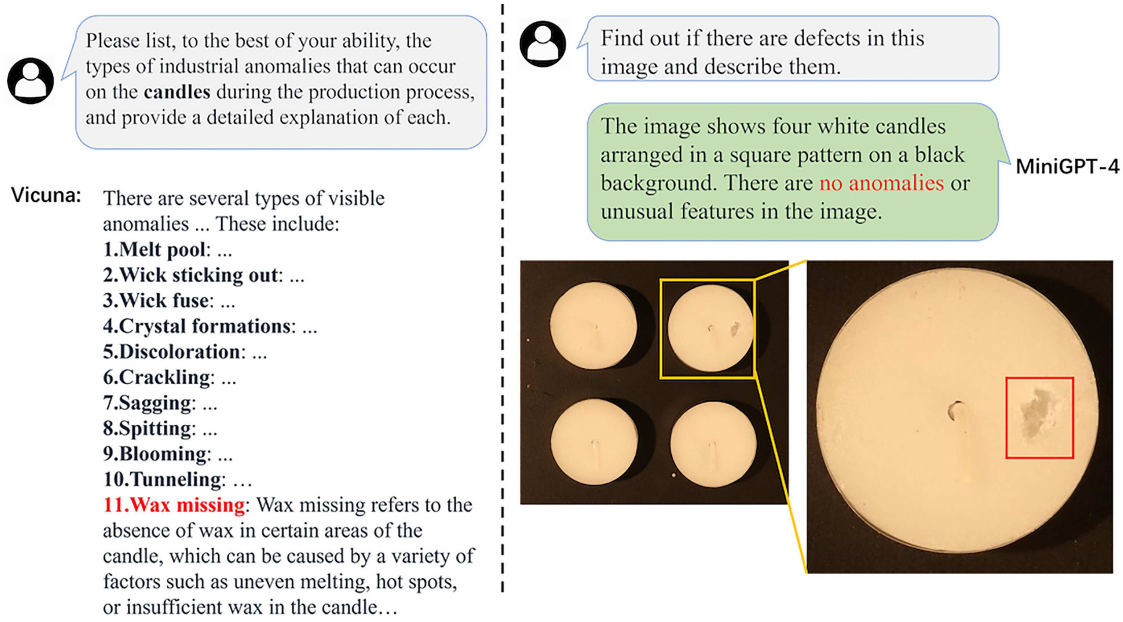

Figure 2 (Color online) MiniGPT-4 [14] fails to utilize the IAD knowledge in Vicuna [44] to recognize missing wax on the candles.

but can also learn from anomaly maps that contain the underlying cognition and comprehension of IAD models. For the latter, the anomaly map is fed into the language model as another input (denoted $\mathbf{t}_e$), hoping that it can deliver additional information to the language model (e.g., positional information), as the vision encoder usually extracts more visual appearance features.

In this way, given a text instruction $\mathbf{t}_i$, Myriad is capable of generating a text response $\mathbf{R}$ corresponding to the instruction, such as the anomaly detection results and detailed descriptions about the anomalies, i.e.,

$$R = \mathrm{LLM}(\mathbf{t}_i, \mathbf{t}_v, \mathbf{t}_e), \tag{1}$$

where both $\mathbf{t}_e$ and $\mathbf{t}_v$ are obtained from $\mathbf{I}$, and LLM denotes a large language model.

3.3 Vision expert-guided vision encoder

To obtain an IAD-suitable $\mathbf{t}_v$ from $\mathbf{I}$, we design the vision expert-guided vision encoder as follows. First, we let the vision encoder take anomaly maps as input as well, by introducing a visual prompt generator that produces expert query tokens $\mathbf{q}_e$ tailored to the original Q-Former query tokens, i.e.,

$$\mathbf{q}_e = G_Q(\mathrm{VE}(\mathbf{I}); \theta_{G_Q}), \tag{2}$$

where G_Q and θ_{G_Q} denote the visual prompt generator for Q-Former and its parameters, respectively. Second, we introduce a trainable low-rank residual adapter A (LoRRA) that is cascaded to the pre-trained ViT to generate visual representations (denoted by $\mathbf{v}_e$) that are modulated for the IAD task. To avoid overfitting to a certain dataset, the architecture takes inspiration from LoRA [45], which optimizes rank decomposition weight matrices instead of dense layers. The proposed LoRRA consists of two convolutional layers. The first layer compresses the visual features to four dimensions, and the second layer remaps them back to their original dimensionality as part of a residual connection

$$\mathbf{v}_e = \mathbf{v}_\mathbf{I} + A(\mathbf{v}_\mathbf{I}). \tag{3}$$

Note that $\mathbf{v}_\mathbf{I}$ is the visual feature extracted from the image $\mathbf{I}$ through pre-trained ViT. Together with the pre-trained query tokens $\mathbf{q}_0$ in Q-Former, the expert query tokens $\mathbf{q}_e$ then jointly interact with the adapted visual representations of industrial images (obtained from the ViT and the visual adapter) through cross-attention, to generate visual representations $\mathbf{f}_p$ following prior knowledge from the experts, i.e.,

$$\mathbf{f}_p = Q(\mathbf{q}_0, \mathbf{q}_e; \mathbf{v}_e). \tag{4}$$

With $\mathbf{f}_p$, finally, we use the pre-trained projection layer MLP in MiniGPT-4 to obtain the visual representations that are (1) suitable for the IAD task and (2) understandable by the LLM:

$$\mathbf{t}_v = \mathrm{MLP}(\mathbf{f}_p). \tag{5}$$

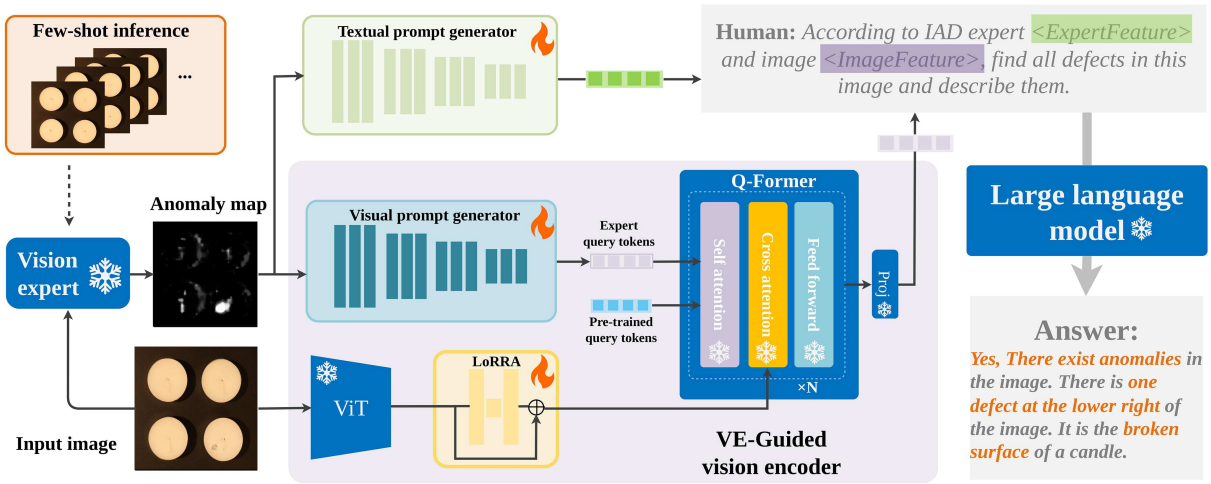

Figure 3 (Color online) The architecture of Myriad. Given an input industrial image, a pre-trained VE produces an anomaly map M as a prior. To adapt the visual pathway to IAD, a VE-guided vision encoder aligns generic features to industrial textures and concentrates attention on suspected regions using expert prompts from the VPG. In parallel, the TPG distills M into compact VE tokens that encode position, extent, and confidence, enabling the LMM backbone to exploit the prior. The LMM fuses VE-guided features and VE tokens to output a standardized, explainable decision with refined anomaly predictions. The framework is plug-and-play with different experts, so one model operates across one-class, zero-shot, and few-shot settings without per-product retraining.

With these designs, Myriad not only performs accurate IAD with high-quality anomaly maps from the vision experts but also achieves favorable performance when the vision experts fail. See evidence in Subsection 4.2.

3.4 Prompt generator

Traditional IAD models provide anomaly maps, which contain sufficient prior knowledge of the IAD task. To enable LLMs to receive extra auxiliary information, such as positional prior, we design the prompt generator, which aims to transform the anomaly maps into prompts $\mathbf{t}_e$ that compress priors from IAD vision experts.

$$\mathbf{p}_* = G_*(\mathbf{M}; \theta_{G_*}), \tag{6}$$

where G_* and θ_{G_*} are denoted as the prompt generator and its parameters, respectively, and $\mathbf{M} \in \mathbb{R}^{H \times W}$ represents the anomaly map of $\mathbf{I}$ produced by a certain vision expert. $\mathbf{p}_*$ is denoted as $\mathbf{p}_{\text{LLM}}$ or $\mathbf{p}_Q$. Note that in Myriad, we use two prompt generators (G_T and G_V) to prompt LLM and Q-Former, respectively. As shown in Figure 3, the prompt generator contains several blocks, each consisting of a convolution with 3×3 kernel, a ReLU as the activation function, and a max pooling, to map its input anomaly map $\mathbf{M}$ into expert prompts $\mathbf{p}_* \in \mathbb{R}^{D_{\text{VE}} \times D_{in,*}}$, where $D_{in,*}$ is the input dimension of the target module (LLM or Q-Former) and D_{VE} is the number of the expert prompt. In our experiments, D_{VE} is set to 9 for LLM and 49 for Q-Former.

Since the prompt generator only takes anomaly maps as inputs, it makes the vision expert tokens decoupled from the appearance and domain of colorful images to some extent, which can be more robust to domain shift in practice.

3.5 Learning objective

The training of our Myriad contains two stages. At the first stage, we pre-train the weights of the VE-guided vision encoder and textual prompt generator for LLM separately, such that $\mathbf{t}_v$ and $\mathbf{p}_{\text{LLM}}$ (also denoted as $\mathbf{t}_e$) contain reliable representations and can also be well received by LLM, i.e.,

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{n} y_k \log(\text{LLM}(\mathbf{t}_i, \mathbf{t}_*)), \tag{7}$$

where y_k is given to represent the target text sequence. $\mathbf{t}_*$ is denoted as $\mathbf{t}_e$ or $\mathbf{t}_v$. Cross-entropy loss $\mathcal{L}_{\text{CE}}$ is commonly employed for training language models. At the end of the stage, we obtain pre-trained weights of both the textual prompt generator and the VE-guided vision encoder.

At the second stage, the pre-trained weights of the textual prompt generator and VE-guided vision encoder are loaded and fine-tuned as shown in Figure 3. Myriad is activated to make use of both visual information guided by

experts and direct auxiliary hints from the vision expert outputs. Formally, we jointly train all parameters with combined losses:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{n} y_k \log(\text{LLM}(\mathbf{t}_i, \mathbf{t}_v, \mathbf{t}_e)). \tag{8}$$

3.6 IAD instruction data preparation

Due to the lack of detailed annotations and the small amount of IAD datasets, we train Myriad with only binary anomaly detection tasks. To construct the task, we design several instruction templates. The templates are given by paraphrasing, for example, "According to IAD expert ⟨Img⟩ ⟨ExpertFeature⟩ ⟨/Img⟩ and image ⟨Img⟩ ⟨ImageFeature⟩ ⟨/Img⟩, find all defects in this image.", where ⟨ImageFeature⟩ is a placeholder of the visual tokens produced from the VE-guided vision encoder and ⟨ExpertFeature⟩ is another placeholder for prompts extracted from the textual prompt generator. While pre-training the textual prompt generator and VE-guided vision encoder separately, the instruction templates are "According to IAD expert ⟨Img⟩ ⟨ExpertFeature⟩ ⟨/Img⟩, find all defects in this image." and "According to the image ⟨Img⟩ ⟨ImageFeature⟩ ⟨/Img⟩, find all defects in this image.", respectively. For the ground truth response, if anomalies are detected, the response is "Yes, there exist anomalies in this image." Otherwise, it is "No, there are no anomalies in this image."

To produce vision-language pairs in the one-class setting where only normal images are available, we follow prior studies [43, 46] to generate synthetic anomalies from normal images. These simulated anomalies act as proxies for abnormal patterns, allowing the model to learn the distinction between normal and abnormal features, which can then generalize to real anomalies. More specifically, the synthetic anomalies are generated using the cut-paste method, where several random regions in the source images are cut and pasted to a random position in each target image.

4 Experiments

Datasets. Our experiments are performed on the MVTec-AD [1] and VisA datasets [2]. The MVTec-AD dataset [1] consists of 3629 samples in the train set and 1725 samples in the test set, respectively. MVTec-AD contains 15 sub-datasets across different types of industrial products, including 5 texture sub-datasets and 10 object sub-datasets, making it the classical dataset in industrial anomaly detection. The VisA dataset [2] contains 9621 normal and 1200 anomalous samples, and covers 12 objects. VisA has multiple objects, a complex background, and fewer abnormal samples, making it a dataset closer to actual industrial applications than MVTec-AD. Therefore, VisA has become a popular and more challenging benchmark. For both MVTec-AD and VisA, we train Myriad following the standard one-class data split setting [2], where only normal samples are visible during training, and all abnormal samples are used for testing.

Evaluation settings. This work performs experiments on three settings: one-class setting, zero-shot setting, and few-shot setting. For the one-class setting, models are allowed to train with normal images of products. For the zero-shot setting, products are unseen for the IAD models. For the few-shot setting, "N-shot" means that N normal images of a product are visible during inference.

Evaluation metrics. By applying the vision experts, Myriad can also provide an anomaly map; thus, we can first follow previous studies to report its I-AUROC and P-AUROC performance with the experts, although (considering that our Myriad utilizes previous vision experts to obtain anomaly maps) it is easy to match its I-AUROC and P-AUROC performance to previous state-of-the-art methods. In addition, we report the mean accuracy across products based on the binary text responses of LMMs. Unlike raw VE anomaly scores that require thresholding, the LMM produces stable image-level binary judgments, making accuracy a practical and deployment-relevant metric that complements AUC. The threshold used to compute accuracy for conventional IAD methods is determined with the max scores of normal samples over k-fold evaluations. Note that the process is only performed on the train set. In practice, we set $k = 3$.

Implementation details. We utilize MiniGPT-4 as the base LMM and test in three different settings: the zero-shot setting, the few-shot setting, and the typical one-class setting. Note that MiniGPT-4 is pretrained on conceptual captions [47,48], SBU [49], and LAION [50]. In each setting, we use different vision experts. In the zero-shot setting, AprilGAN [9] is used as a baseline. We introduce MuSc [51] to evaluate the plug-and-play capability of the prompt generators and to serve as a more effective expert for improving Myriad's anomaly detection performance during inference. In the classical one-class settings, we use the image decoder and ImageBind backbone [32] in AnomalyGPT [43] as the vision expert. In the few-shot setting, the anomaly maps are obtained in the same way as

Table 2 One-class anomaly detection results on the MVTec-AD dataset. The best-performing method is in bold, and the second-best is underlined. Note that the overall accuracy is defined over all test samples, while the accuracy is defined as the mean accuracy over products.

Method	Image-AUC	Pixel-AUC	Accuracy (%)	Overall accuracy (%)
PaDiM	95.3	<u>97.4</u>	76.5	75.3
PatchCore	<u>99.0</u>	**98.1**	89.2	88.1
SimpleNet	**99.6**	**98.1**	93.0	92.2
UniAD	97.6	97.0	89.3	88.4
AnomalyGPT	97.4	93.1	93.3	92.1
Myriad (AnomalyGPT)	97.4	93.1	<u>94.2</u>	<u>93.2</u>
Myriad (SimpleNet)	**99.6**	**98.1**	**94.6**	**94.7**

Patchcore [8], by adapting ImageBind [32] as the backbone and estimating cosine similarity between the features of the test images and those of the few-shot normal samples.

Our training strategy consists of a pre-training and a fine-tuning stage. In the pre-training stage, we train the vision-expert guidance adapter and the vision-expert guided visual feature extraction module independently. Both modules would be combined and fine-tuned in the fine-tuning stage with a lower learning rate and real abnormal data. The pre-training process lasts for 32000 steps with a batch size of 16. In each training batch, half of the samples are real normal samples, and the other half contains simulated abnormal samples [46]. AdamW, with a weight decay of 0.05, is employed as the optimizer. No warmup step is used. The cosine learning rate decay strategy is applied, and the minimum learning rate is 0. In the fine-tuning phase, the training process lasts for 16000 steps with a batch size of 16. AdamW, with a learning rate of 3e−5, is used for fine-tuning. The cosine learning rate decay strategy is also used, and the final learning rate is set to 1e−5. All experiments are conducted on 2 NVIDIA A100 GPUs.

4.1 Quantitative results

In this subsection, we report quantitative comparison results between Myriad and previous IAD studies, including feature-embedding-based methods, reconstruction-based methods, and language-guided methods. Given different vision experts, Myriad can perform as zero-shot, few-shot, and one-class anomaly detectors.

One-class IAD. We report one-class IAD results on MVTec-AD in Table 2. Myriad achieves accuracy superior to that of the state-of-the-art methods on MVTec-AD. Our solution outperforms PatchCore by about 5.0% (94.2% vs. 89.2%) and outperforms SimpleNet by about 1.2% (94.2% vs. 93.0%). Moreover, with the same anomaly maps predicted by AnomalyGPT, Myriad outperforms AnomalyGPT by about 0.9% (94.2% vs. 93.3%). To ensure a fair comparison, we use the same training data and identical prompt-generator architecture for the LLM as in the prompt learner of AnomalyGPT. These results indicate that the VE-guided vision encoder extracts superior IAD visual features using the 0.4-billion-parameter EVA-CLIP model compared to the single-class image token produced by the 1.2-billion-parameter backbone employed in AnomalyGPT. Furthermore, more advanced vision-expert models (e.g., SimpleNet) can also be integrated into the Myriad framework to further enhance anomaly-detection accuracy.

Zero-shot IAD. With zero-shot vision experts, such as AprilGAN [9] and MuSc [51], Myriad can also perform as a zero-shot anomaly detector. As demonstrated in Table 3, Myriad with vision expert AprilGAN achieves 7.9% performance (72.2% vs. 64.3%) on MVTec-AD and 3.3% (63.2% vs. 59.9%) on VisA. Without fine-tuning, MuSc can be used as a plug-and-play vision expert. Myriad with MuSc achieves 79.3% on MVTec-AD and 78.7% on VisA. Compared with AprilGAN, Myriad gains 7.1% and 15.5% improvement on MVTec-AD and VisA, respectively. The result indicates that better vision experts are able to boost Myriad without any additional cost.

PCB Bank [17] is a novel benchmark for printed circuit board (PCB) anomaly detection. As in Table 4, we apply zero-shot evaluation to its test set and report results across several state-of-the-art general-domain LMMs and LMM-based methods. Myriad, based on a 7-billion-parameter LLM, achieves performance comparable to the 72-billion-parameter LLaVA-OneVision model, which is trained on 1.3 million multimodal data samples.

Compared to AnomalyGPT, Myriad improves accuracy by 11.7% (61.8% vs. 50.1%). For a fair comparison, we replace the input anomaly maps for the prompt learner in AnomalyGPT with binarized anomaly maps predicted by MuSc. This modification effectively boosts AnomalyGPT's accuracy from 50.1% to 52.3%. However, Myriad still outperforms AnomalyGPT by 9.5%. The performance gap between boosted AnomalyGPT and Myriad is attributed to the vision modality. We notice the detailed nature of PCB Bank, which includes numerous components and multiple labels. AnomalyGPT only provides a single class token with pre-trained ImageBind to the LLM, representing a global visual feature that loses many detailed aspects of the PCB. In contrast, Myriad supplies patch

Table 3 Zero/few-shot IAD results on MVTec-AD and VisA datasets. Results are listed as the average of 5 runs, and the best-performing method is in bold. The results for SPADE and WinCLIP are reported from [28]. The models are cross-validated across two datasets: MVTec-AD [1] and VisA [2].

Setup	Method	MVTec-AD						VisA						VisA D&R
		Image-AUC	Pixel-AUC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Image-AUC	Pixel-AUC	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	GPT-score
0-shot	MuSc	**97.8** (±0.1)	**97.1** (±0.1)	–	–	–	–	**92.6** (±0.2)	**98.7** (±0.2)	–	–	–	–	–
	Myriad (MuSc)	97.8 (±0.1)	97.1 (±0.1)	**79.3** (±0.0)	**95.0**	**75.0**	**80.6**	**92.6** (±0.2)	**98.7** (±0.2)	**78.7** (±0.2)	**95.7**	**62.2**	**72.2**	**4.21**
1-shot	AprilGAN	**86.2** (±0.5)	**87.6** (±0.1)	64.3 (±0.3)	**95.1**	52.3	62.5	**78.0** (±0.4)	**94.2** (±0.3)	59.9 (±0.2)	93.1	27.2	36.7	–
	Myriad (AprilGAN)	86.2 (±0.5)	87.6 (±0.1)	**72.2** (±0.2)	93.1	**64.7**	**72.4**	78.0 (±0.4)	94.2 (±0.3)	**63.2** (±0.2)	**96.1**	**34.7**	**47.8**	**3.03**
	SPADE	81.0 (±2.0)	91.2 (±0.4)	–	–	–	–	79.5 (±4.0)	95.6 (±0.4)	–	–	–	–	–
	PaDiM	75.5 (±1.0)	90.0 (±0.4)	57.4 (±1.6)	–	–	–	59.6 (±1.8)	91.3 (±0.3)	45.0 (±0.6)	–	–	–	–
	PatchCore	84.2 (±1.2)	92.4 (±0.5)	65.0 (±1.8)	–	–	–	76.8 (±1.6)	93.5 (±0.6)	60.0 (±1.0)	–	–	–	–
	WinCLIP	93.1 (±2.0)	95.2 (±0.5)	–	–	–	–	83.8 (±4.0)	**96.4** (±0.4)	–	–	–	–	–
	AprilGAN	92.0 (±0.3)	95.1 (±0.1)	67.5 (±0.2)	**97.6**	55.2	66.1	**91.2** (±0.8)	96.0 (±0.0)	75.6 (±0.1)	94.9	59.9	67.9	–
	AnomalyGPT	**94.1** (±1.1)	**95.3** (±0.1)	86.1 (±1.1)	89.2	92.1	90.1	87.4 (±0.8)	96.2 (±0.1)	77.4 (±1.0)	90.6	68.2	75.4	3.40
	Myriad (AprilGAN)	92.0 (±0.3)	95.1 (±0.1)	76.5 (±0.3)	93.9	71.0	78.2	91.2 (±0.8)	96.0 (±0.0)	72.7 (±0.1)	**99.7**	50.1	53.8	3.21
	Myriad (AnomalyGPT)	**94.1** (±1.1)	**95.3 (±0.1)**	**87.4** (±0.9)	90.7	**92.8**	**91.1**	87.4 (±0.8)	96.2 (±0.1)	**80.0** (±0.4)	80.2	**90.0**	**83.5**	**4.89**
2-shot	SPADE	82.9 (±2.6)	92.0 (±0.3)	–	–	–	–	80.7 (±5.0)	96.2 (±0.4)	–	–	–	–	–
	PaDiM	78.2 (±0.6)	92.1 (±0.4)	56.9 (±0.8)	–	–	–	65.5 (±1.5)	93.2 (±0.1)	46.4 (±0.7)	–	–	–	–
	PatchCore	87.1 (±0.8)	94.1 (±0.2)	68.4 (±2.3)	–	–	–	80.4 (±0.7)	95.0 (±0.2)	61.8 (±1.2)	–	–	–	–
	WinCLIP	94.4 (±1.3)	**96.0** (±0.3)	–	–	–	–	84.6 (±2.4)	**96.8** (±0.3)	–	–	–	–	–
	AprilGAN	92.4 (±0.3)	95.5 (±0.0)	67.6 (±0.1)	**98.0**	55.2	65.9	**92.2** (±0.3)	96.2 (±0.0)	74.2 (±0.2)	94.6	58.1	67.3	–
	AnomalyGPT	**95.5** (±0.8)	95.6 (±0.2)	84.8 (±0.8)	93.9	85.8	88.3	88.6 (±0.7)	96.4 (±0.1)	77.5 (±0.3)	93.5	62.5	71.5	3.46
	Myriad (AprilGAN)	92.4 (±0.3)	95.5 (±0.0)	77.1 (±0.2)	94.1	72.4	79.2	92.2 (±0.3)	96.2 (±0.0)	75.3 (±0.4)	**98.7**	55.0	69.0	3.27
	Myriad (AnomalyGPT)	**95.5** (±0.8)	95.6 (±0.2)	**86.2** (±0.7)	92.5	**89.0**	**89.8**	88.6 (±0.7)	96.4 (±0.1)	**82.3** (±0.5)	83.6	**87.6**	**84.5**	**5.02**
4-shot	SPADE	84.8 (±2.5)	92.7 (±0.3)	–	–	–	–	81.7 (±3.4)	96.6 (±0.3)	–	–	–	–	–
	PaDiM	80.9 (±0.9)	94.0 (±0.2)	57.9 (±1.2)	–	–	–	69.6 (±1.5)	94.4 (±0.1)	48.0 (±1.4)	–	–	–	–
	PatchCore	89.5 (±1.3)	94.9 (±0.2)	72.5 (±1.8)	–	–	–	82.2 (±0.8)	96.0 (±0.1)	63.1 (±0.4)	–	–	–	–
	WinCLIP	95.2 (±1.3)	96.2 (±0.3)	–	–	–	–	87.3 (±1.8)	**97.2** (±0.2)	–	–	–	–	–
	AprilGAN	92.8 (±0.2)	95.9 (±0.0)	67.2 (±0.1)	**98.6**	54.5	66.0	**92.6** (±0.4)	96.2 (±0.0)	75.7 (±0.1)	95.5	59.4	67.5	–
	AnomalyGPT	**96.3** (±0.3)	**96.2** (±0.1)	85.0 (±0.3)	96.7	80.6	85.6	90.6 (±0.7)	96.7 (±0.1)	77.7 (±0.4)	94.5	56.5	66.1	3.54
	Myriad (AprilGAN)	92.8 (±0.2)	95.9 (±0.0)	77.8 (±0.2)	94.5	73.1	80.0	92.6 (±0.4)	96.2 (±0.0)	76.5 (±0.2)	**98.7**	58.3	71.8	3.31
	Myriad (AnomalyGPT)	**96.3** (±0.3)	**96.2** (±0.1)	**86.0** (±0.3)	94.5	**85.5**	**88.6**	90.6 (±0.7)	96.7 (±0.1)	**83.5** (±0.3)	87.6	**84.5**	**85.3**	**5.54**

Table 4 Zero-shot IAD results on PCB Bank datasets. AnomalyGPT and Myriad are trained on MVTec-AD, because PCB Bank contains examples of VisA. The best-performing method is in bold.

Setup	Method	Params (B)	Image-AUC	Pixel-AUC	Accuracy (%)
0-shot	MuSc	–	80.4	97.6	–
	Qwen2-VL	2	–	–	49.9
	LLaVA-1.6	7	–	–	38.9
	LLaVA-OneVision-SI	8	–	–	60.8
	LLaVA-OneVision-OV	8	–	–	59.6
	Qwen2-VL	7	–	–	59.5
	Qwen2-VL	72	–	–	59.6
	LLaVA-OneVision-OV	72	–	–	60.9
	AnomalyGPT	7	53.2	82.5	50.1
	AnomalyGPT (MuSc)	7	80.4	97.6	52.3
	Myriad (MuSc)	7	80.4	97.6	**61.8**

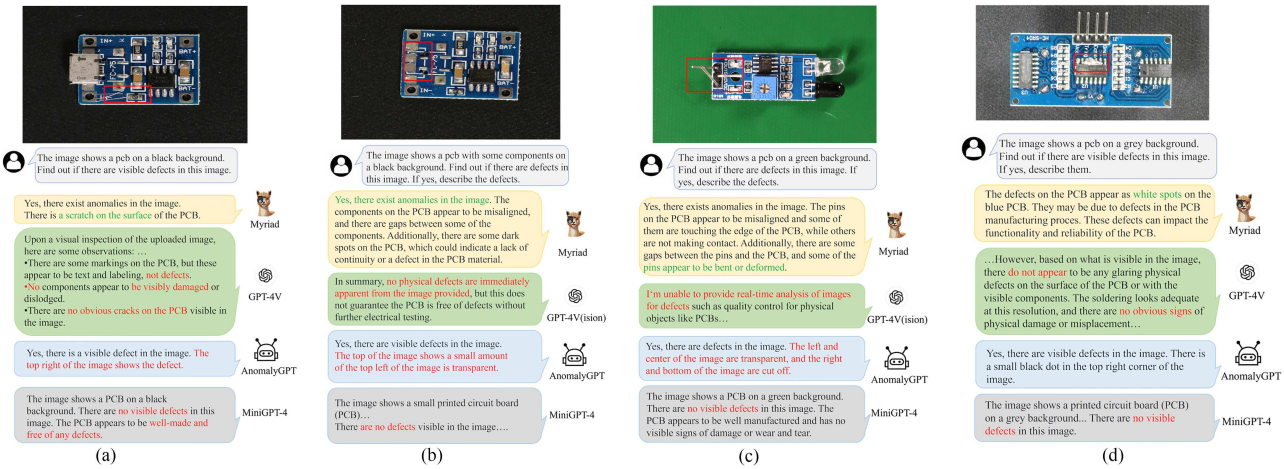

Figure 4 (Color online) The qualitative comparison between Myriad and the state-of-the-art LMMs. (a) Myriad detects a scratch and provides a brief description; (b) Myriad recognizes subtle holes appearing as a dark spot; (c) Myriad identifies noticeable pin deformation; (d) Myriad flags the presence of excess solder as white spots. Correct details are highlighted in green, while incorrect details are marked in red. The ground truth defects are highlighted with red bounding boxes in the image. Best viewed in color.

tokens and resamples them using the Q-Former and expert prompts, resulting in more comprehensive anomaly detection compared to AnomalyGPT.

Few-shot IAD. In the few-shot setting, we mainly compare with SPADE [18], PaDiM [7], PatchCore [8], WinCLIP [28], and AnomalyGPT [43]. These methods all have a memory bank but have different matching strategies or augmentations with few references. We follow [43] to adopt ImageBind [32] to generate few-shot anomaly maps by computing cosine similarity features of test images and those of few-shot normal samples. This expert is denoted as AnomalyGPT because it keeps the same I-AUROC and P-AUROC as AnomalyGPT.

Results are reported in Table 3. Myriad combined with AnomalyGPT achieves the highest one-shot accuracy among both conventional IAD methods and LMM-based approaches, attaining 80.0% on VisA and 87.4% on MVTec-AD. In the two-shot and four-shot settings, Myriad's performance decreases compared to the one-shot scenario due to internal variations among normal samples. We adopt AnomalyGPT's selection of few-shot examples, and a similar performance decline is observed in AnomalyGPT. Nevertheless, Myriad still maintains the best accuracy in both two-shot and four-shot settings across MVTec-AD and VisA.

4.2 Qualitative examples

We investigate the qualitative performance of Myriad in this subsection. First, we qualitatively compare Myriad with recent LMMs, including MiniGPT-4 [14], AnomalyGPT [43], and ChatGPT-4V (ision) [52]. Specifically, we use the November 21, 2023 version for ChatGPT-4V (ision). As shown in Figure 4, GPT-4V provides more detailed image descriptions compared to MiniGPT-4. However, it still fails to detect existing anomalies in complex production environments. While AnomalyGPT can identify the presence of anomalies, it often inaccurately describes them in

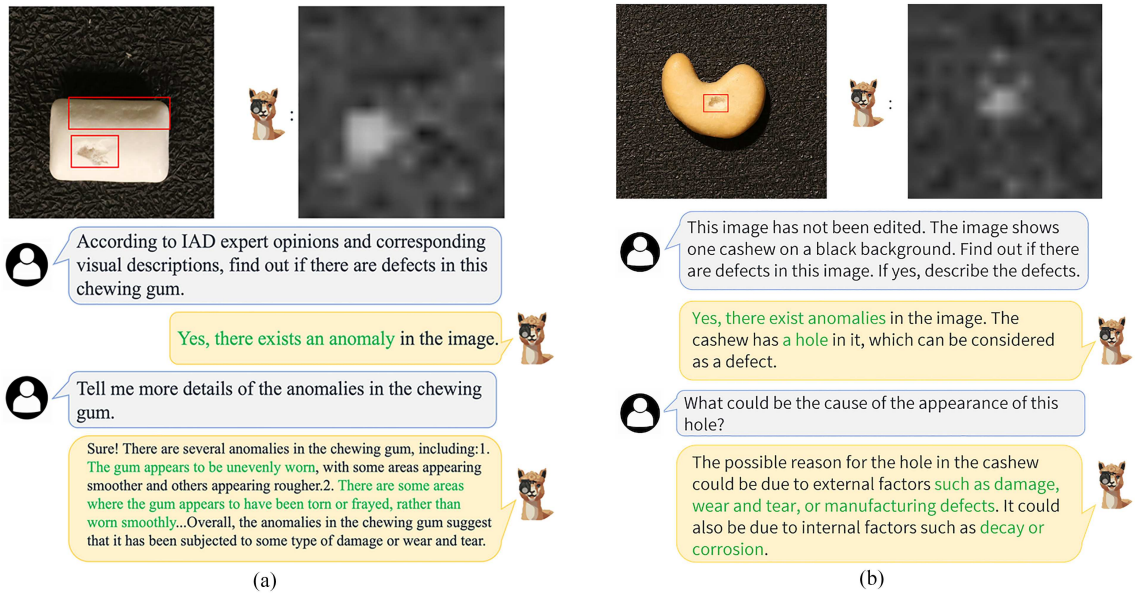

Figure 5 (Color online) The illustration of instruction-following ability. (a) Myriad provides user-instructed descriptions of anomalous regions; (b) Myriad outputs its internal abnormality knowledge according to user instructions. The ground truth defects are highlighted with red bounding boxes in the image. Best viewed in color.

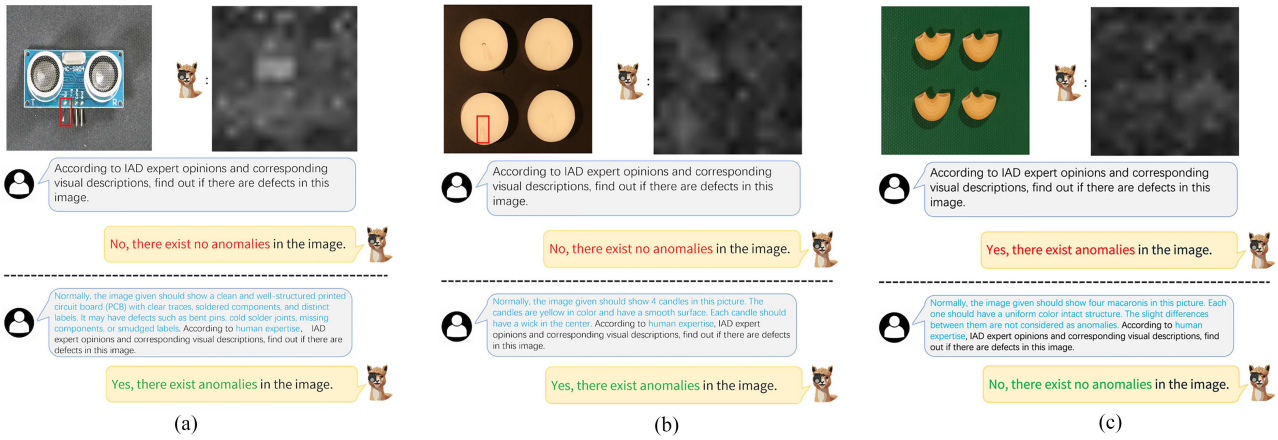

Figure 6 (Color online) The illustration of flexibility. (a) Given possible anomaly categories, Myriad can refine its assessment; (b) given only a description of normal candles' appearance, Myriad can likewise revise its judgement by comparison; (c) given criteria, Myriad can adjust its decision threshold to some extent. The ground truth defects are highlighted with red bounding boxes in the image.

detail. In contrast, our method not only detects the existence of anomalies but also provides additional categorical information.

Compared to existing learning-based IAD approaches, Myriad demonstrates remarkable flexibility and instruction-following capabilities inherited from general-domain LMMs. Specifically, Myriad can follow instructions to provide detailed descriptions of anomalies (see Figure 5(a)). Moreover, users can leverage Myriad to identify the likely causes of these anomalies (see Figure 5(b)).

Myriad further illustrates its flexibility by correcting previously inaccurate responses when provided with additional human expertise, such as information on product categories, process manuals, and common anomaly types (see Figure 6). The flexibility and instruction-following capabilities address the key limitations of current industrial anomaly detection methods, which incur high retraining costs when upgrading production lines or introducing new manufacturing technologies. Consequently, constructing large multimodal models such as Myriad holds significant research value for industrial anomaly detection.

Table 5 Ablation studies on architecture. The best-performing method is in bold.

TPG	VE-guided vision encoder		Accuracy (%)	
	LoRRA	VPG	MVTec-AD (one-class)	VisA (1-shot)
✓			92.0	75.6
	✓	✓	92.5	77.4
✓	✓		93.5	75.8
✓		✓	93.0	76.2
✓	✓	✓	**94.2**	**80.0**

Table 6 Comparison on computation and memory efficiency. Reported results are evaluated on one NVIDIA A800 80 G.

	AprilGAN	MuSc	AnomalyGPT	Myriad
Training time (h)	0.3	–	≈ 24	≈ 24
Inference time (s)	0.1	0.9	1.0	0.5
Inference memory consumption (GB)	2.4	8.8	16.2	21.0

4.3 Ablation studies

Effectiveness of the VE-guided vision encoder. As shown in Table 5, the VE-guided vision encoder plays a pivotal role in Myriad, as removing it causes a 2.2% accuracy drop under the one-class setting and a 4.4% drop under the one-shot setting. Specifically, using LoRRA or the VPG alone in the one-shot setting provides only marginal improvements of 0.2% and 0.6%, respectively. However, combining these two modules yields a performance gain of at least 3.8% (80.0% vs. 76.2%), indicating that both expert guidance and IAD-specific visual features are crucial for detecting unseen product anomalies.

Under the one-class setting, improvements are largely smooth. LoRRA outperforms VPG, suggesting that LoRRA effectively learns product visual patterns for anomaly detection, while VPG can concentrate on the regions highlighted by external IAD models as potential defect areas.

Effectiveness of the textual prompt generator. The textual prompt generator offers comparable performance in both the one-class and one-shot settings. When using only expert guidance, AnomalyGPT outperforms Myriad (93.3% vs. 92.0%) in the one-class setting and 77.4% vs. 75.6% in the one-shot setting, likely due to its ImageBind vision encoder, which has three times as many parameters as Myriad (1.2B vs. 0.4B). When relying solely on the VE-guided vision encoder, Myriad achieves only 92.5% and 77.4% in the one-class and one-shot settings, respectively. This limitation arises from the expert guidance loss introduced by the Q-Former. By contrast, directly translating expert knowledge through the textual prompt generator allows Myriad to recover this lost information, yielding performance gains of 1.7% and 2.6% in the one-class and one-shot settings, respectively.

Comparison on computation cost and memory consumption. In Table 6, we report the computational cost and memory efficiency of typical zero- and few-shot methods, including AprilGAN, MuSc, AnomalyGPT, and Myriad. In practice, Myriad can run on a single NVIDIA RTX 3090, which has 24 GB of memory. Moreover, Myriad can process up to 1.73 million samples per day, while AnomalyGPT handles only 860000 images per day. As a result, Myriad is suitable for many industrial inspection scenarios that do not require strict real-time performance, such as final inspection stages.

5 Conclusion

In this paper, we introduce a novel large multimodal model, Myriad, designed to address industrial anomaly detection in a flexible and data-efficient manner. By leveraging existing industrial anomaly detection methods as "vision experts", Myriad incorporates anomaly maps into large multimodal models to highlight critical regions, enhance the learning of anomalous features, and stimulate the underlying cognition of the experts. In this way, our approach benefits from both the specialized knowledge of IAD methods and the powerful generalization and instruction-following abilities of large multimodal models. Extensive experiments on MVTec-AD, VisA, and PCB Bank benchmarks demonstrate that Myriad achieves state-of-the-art performance under both one-class and few-shot settings. Moreover, thanks to the modularity of its architecture and the integration of vision experts, Myriad can be naturally extended to various real-world industrial scenarios without requiring substantial model redesign. We believe that this work opens promising avenues for bridging specialized expert models with large multimodal backbones, paving the way for more robust, adaptable, and efficient solutions in industrial anomaly detection.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2023YFA1008500) and Heilongjiang Postdoctoral Fund (Grant No. LBH-Z25148).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Bergmann P, Fauser M, Sattlegger D, et al. MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 9592–9600
2 Zou Y, Jeong J, Pemula L, et al. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: Proceedings of European Conference on Computer Vision, 2022. 392–408
3 Mishra P, Verk R, Fornasier D, et al. VT-ADL: a vision transformer network for image anomaly detection and localization. In: Proceedings of the 30th International Symposium on Industrial Electronics, 2021. 1–6
4 Jezek S, Jonak M, Burget R, et al. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In: Proceedings of the 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, 2021. 66–71
5 Bao T, Chen J, Li W, et al. Miad: a maintenance inspection dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 993–1002
6 Liu Z, Zhou Y, Xu Y, et al. SimpleNet: a simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 20402–20411
7 Defard T, Setkov A, Loesch A, et al. Padim: a patch distribution modeling framework for anomaly detection and localization. In: Proceedings of International Conference on Pattern Recognition, 2021. 475–489
8 Roth K, Pemula L, Zepeda J, et al. Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 14318–14328
9 Chen X, Han Y, Zhang J. 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
10 Zhang X, Li S, Li X, et al. DeSTSeg: segmentation guided denoising student-teacher for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 3914–3923
11 Zhang X, Li N, Li J, et al. Unsupervised surface anomaly detection with diffusion probabilistic model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 6782–6791
12 Zhang H, Wang Z, Wu Z, et al. DiffusionAD: norm-guided one-step denoising diffusion for anomaly detection. 2023. ArXiv:2303.08730
13 Yin H, Jiao G, Wu Q, et al. LafitE: latent diffusion model with feature editing for unsupervised multi-class anomaly detection. 2023. ArXiv:2307.08059
14 Zhu D, Chen J, Shen X, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models. In: Proceedings of International Conference on Learning Representations, 2024. 18378–18394
15 Fang Y, Wang W, Xie B, et al. Eva: exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19358–19369
16 Li J, Li D, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of International Conference on Machine Learning, 2023. 19730–19742
17 Yao H, Liu M, Wang H, et al. GLAD: towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In: Proceedings of European Conference on Computer Vision, 2024. 1–17
18 Cohen N, Hoshen Y. Sub-image anomaly detection with deep pyramid correspondences. 2020. ArXiv:2005.02357
19 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations, 2015. 1–14
20 Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In: Proceedings of International Conference on Machine Learning, 2019. 6105–6114
21 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016. 770–778
22 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, 2022. 9737–9746
23 Tien T D, Nguyen A T, Tran N H, et al. Revisiting reverse distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 24511–24520
24 You Z, Cui L, Shen Y, et al. A unified model for multi-class anomaly detection. In: Proceedings of Advances in Neural Information Processing Systems, 2022. 4571–4584
25 Yao X, Li R, Qian Z, et al. Focus the discrepancy: intra-and inter-correlation learning for image anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 6803–6813
26 Zhao Y. OmniAL: a unified CNN framework for unsupervised anomaly localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 3924–3933
27 Lu F, Yao X, Fu C W, et al. Removing anomalies as noises for industrial defect localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 16166–16175
28 Jeong J, Zou Y, Kim T, et al. Winclip: zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19606–19616
29 Deng H, Zhang Z, Bao J, et al. AnoVL: adapting vision-language models for unified zero-shot anomaly localization. 2023. ArXiv:2308.15939
30 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of International Conference on Machine Learning, 2021. 8748–8763
31 Zhong Y, Yang J, Zhang P, et al. Regionclip: region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16793–16803
32 Girdhar R, El-Nouby A, Liu Z, et al. ImageBind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15180–15190
33 Peng Z, Wang W, Dong L, et al. Grounding multimodal large language models to the world. In: Proceedings of International Conference on Learning Representations, 2024. 51575–51598
34 Chen K, Zhang Z, Zeng W, et al. Shikra: unleashing multimodal LLM's referential dialogue magic. 2023. ArXiv:2306.15195
35 Zhou Q, Yu C, Zhang S, et al. RegionBLIP: a unified multi-modal pre-training framework for holistic and regional comprehension. 2023. ArXiv:2308.02299
36 Liu H, Li C, Wu Q, et al. Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 34892–34916
37 Wang W, Shi M, Li Q, et al. The all-seeing project: towards panoptic visual recognition and understanding of the open world. In: Proceedings of International Conference on Learning Representations, 2024. 24025–24057

38 Wang W, Chen Z, Chen X, et al. VisionLLM: large language model is also an open-ended decoder for vision-centric tasks. In: Proceedings of Advances in Neural Information Processing Systems, 2023. 61501–61513
39 Ding H, Tang S, He S, et al. Multimodal referring segmentation: a survey. 2025. ArXiv:2508.00265
40 Ding H, Ying K, Liu C, et al. MOSEv2: a more challenging dataset for video object segmentation in complex scenes. 2025. ArXiv:2508.05630
41 Ding H, Liu C, He S, et al. MOSE: a new dataset for video object segmentation in complex scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 20224–20234
42 Ding H, Liu C, He S, et al. MeViS: a large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 2694–2703
43 Gu Z, Zhu B, Zhu G, et al. AnomalyGPT: detecting industrial anomalies using large vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 1932–1940
44 Chiang W L, Li Z, Lin Z, et al. Vicuna: an open-source chatbot impressing GPT-4 with 90
45 Hu E J, Wallis P, Allen-Zhu Z, et al. LoRA: low-rank adaptation of large language models. 2021
46 Schlüter H M, Tan J, Hou B, et al. Natural synthetic anomalies for self-supervised anomaly detection and localization. In: Proceedings of European Conference on Computer Vision, 2022. 474–489
47 Changpinyo S, Sharma P, Ding N, et al. Conceptual 12 M: pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 3558–3568
48 Sharma P, Ding N, Goodman S, et al. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. 2556–2565
49 Ordonez V, Kulkarni G, Berg T. Im2text: describing images using 1 million captioned photographs. In: Proceedings of Advances in Neural Information Processing Systems, 2011. 1143–1151
50 Schuhmann C, Vencu R, Beaumont R, et al. Laion-400M: open dataset of clip-filtered 400 million image-text pairs. 2021. ArXiv:2111.02114
51 Li X, Huang Z, Xue F, et al. MuSc: zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. 2024. ArXiv:2401.16753
52 Yang Z, Li L, Lin K, et al. The dawn of LMMs: preliminary explorations with GPT-4V (ision). 2023. ArXiv:2309.17421