Article

Normality-Anchored Regional Evidence Reasoning for Zero-Shot Industrial Anomaly Detection

Yan Wang[1], Guan Zhang[1], Chunxiao Wu[1], Yanyan Guo[1] and Weihua Xiong[1,*]

[1] Jilin University of Chemical Technology, Jilin, China
* Correspondence: Weihua Xiong, Jilin University of Chemical Technology, Jilin, China

Abstract: Industrial visual inspection is a key task in intelligent manufacturing and quality control. However, defective samples in real production lines are usually scarce, diverse in appearance, and expensive to annotate, which makes supervised models that rely on large numbers of defective samples difficult to adapt to new products and unknown defect scenarios. To address this problem, this paper proposes NARE-ZSAD, a normality-anchored regional evidence reasoning framework for zero-shot industrial anomaly detection. The proposed framework first decomposes an input image into local candidate regions through a language-guided multi-scale region proposal mechanism, and then performs region-level vision-text alignment using normal and abnormal textual prompts, thereby reducing the interference caused by global image semantics and large normal background areas. Subsequently, a normality memory bank is constructed to characterize local normal visual patterns of the target product, reformulating anomaly detection as a joint estimation of semantic anomaly responses and deviations from normal patterns. Furthermore, a text-guided pseudo-defect evidence calibration strategy is designed to synthesize local anomaly evidence, such as scratches, stains, missing areas, and color shifts, on defect-free images. This strategy calibrates abnormal prompt responses and regional thresholds without using real defective samples. Finally, a conflict-aware expert reasoning module integrates semantic anomalies, normality deviations, texture inconsistencies, and prompt uncertainty to output anomaly scores, anomaly localization maps, triggered prompts, and manual review decisions. Unlike conventional GAN-based defect image augmentation methods, NARE-ZSAD does not generate full defective images to train supervised classifiers. Instead, it enables region-level evidence discovery, normality-deviation reasoning, and reliability-aware decision making under zero-shot conditions, providing an interpretable and reviewable solution for industrial inspection in defect-scarce and unknown-anomaly scenarios.

Keywords: zero-shot anomaly detection; industrial visual inspection; vision-language model; normality anchoring; regional evidence reasoning; conflict-aware reasoning

Received: 26 June 2026
Revised: 04 August 2026
Accepted: 14 August 2026
Published: 17 August 2026

1. Introduction

Industrial surface anomaly detection is a fundamental task in quality assurance for intelligent manufacturing. Defects such as scratches, cracks, dents, stains, foreign objects, missing parts, and assembly errors may directly affect product reliability, safety, and subsequent assembly quality. On automated production lines, visual inspection systems are required to make stable decisions under high-speed operation, varying illumination, frequent product switching, and process fluctuations. Therefore, industrial inspection models should not only achieve high recognition accuracy, but also provide cross-product generalization, anomaly localization, interpretability, and risk control [1].

Traditional industrial defect detection methods often rely on handcrafted features, threshold segmentation, template matching, or machine-learning classifiers developed for specific defects. These methods may be effective when defect morphology is stable and background texture is simple [2-4]. However, they usually require rule and parameter redesign when facing complex textures, weak defects, illumination variation, or process drift. As product types and defect categories increase, the development and maintenance costs of traditional cascaded inspection systems increase significantly.

Deep learning has provided new solutions for industrial defect inspection. Convolutional neural networks and vision Transformers can automatically learn discriminative representations from data and have achieved promising performance in many defect detection tasks. Nevertheless, most deep models rely on a large number of labeled defective images or pixel-level annotations. In real production lines, defects are low-frequency events, and the number of defective samples is far smaller than that of defect-free samples. Defect categories also exhibit long-tail distributions. In addition, defect annotation requires professional expertise, and pixel-level mask annotation is even more expensive. Therefore, defect scarcity remains a major bottleneck that limits rapid deployment of deep inspection models.

Defect image generation and data augmentation are important approaches for alleviating sample scarcity. Typical GAN-based methods can use a small number of defective images and a large number of defect-free images to generate synthetic defects, thereby expanding the training set and improving downstream classifiers. However, these methods still usually require real defective images to train the generator, and their core pipeline remains 'generate defective images - expand the training set - train a supervised classifier'. When no real defective sample is available on the target production line, or when the defects are unknown, defect-domain generative augmentation methods are still difficult to apply directly [5].

Recently, vision-language models such as CLIP have established a shared semantic space between images and texts through large-scale image-text contrastive learning, providing new opportunities for zero-shot industrial anomaly detection. In industrial inspection tasks, textual prompts can describe normal and abnormal concepts such as 'normal surface', 'scratch', 'crack', 'contaminated region', and 'missing part', enabling models to exploit language priors when target-domain defect samples are absent. However, directly matching global visual features with abnormal textual prompts is insufficient. Industrial defects are usually small, low-contrast, and local. Global image features are easily dominated by product categories, background textures, and large normal regions, leading to missed detections of small defects or false alarms on normal textures [6-9].

Moreover, industrial anomalies are not equivalent to ordinary object categories. Most defects are local deviations from normal textures, geometric structures, material appearances, or assembly relationships. Relying only on abnormal textual prompts may not distinguish real defects from normal but rare texture variations. Therefore, zero-shot industrial anomaly detection needs to address three key issues: how to focus on defect-related local regions instead of global image semantics; how to establish a reliable normality reference when defective samples are absent; and how to provide reviewable decisions when multiple prompts, multiple scales, and multiple evidence sources conflict.

To this end, this paper proposes NARE-ZSAD, a normality-anchored regional evidence reasoning framework for zero-shot industrial anomaly detection. The framework transforms industrial anomaly detection from image-level classification into region-level evidence reasoning. It uses language prompts to guide local region modeling, a normality memory bank to describe normal pattern boundaries, pseudo-defect evidence calibration to stabilize abnormal prompt responses, and conflict-aware expert reasoning to produce anomaly localization, semantic explanation, and manual review decisions. Compared with global CLIP prompting or GAN-based image augmentation, NARE-ZSAD emphasizes local evidence, normal references, and risk control, which is more consistent with the industrial requirements of interpretability, traceability, and reviewability.

The main contributions of this paper are as follows:

A normality-anchored regional evidence reasoning framework, NARE-ZSAD, is proposed for zero-shot industrial anomaly detection.

A region-level vision-text evidence modeling mechanism is designed to reduce the interference of global semantics and large normal backgrounds.

(3) Normality memory and text-guided pseudo-defect calibration are introduced to improve abnormal prompt reliability without using real defective samples. (4) A conflict-aware expert reasoning mechanism is constructed to jointly output anomaly scores, localization maps, triggered prompts, and manual review decisions.

2. Related Work

2.1. Industrial Anomaly Detection under Defect Scarcity

Industrial anomaly detection methods can be broadly divided into supervised defect recognition, unsupervised anomaly detection, few-shot anomaly detection, and zero-shot anomaly detection. Supervised methods usually depend on large numbers of defective images and category labels, and can achieve good performance on fixed products and fixed defect sets. However, their adaptability to unknown defects, new products, and process variations is limited. Because real production-line defects are low-frequency, long-tailed, and expensive to annotate, fully supervised inspection systems are often difficult to deploy rapidly in the early stage of a new production line.

Unsupervised or one-class anomaly detection methods usually model the normal distribution using only defect-free samples, and then identify regions that deviate from normal patterns through reconstruction errors, feature distances, density estimation, or memory retrieval during testing. Memory-bank methods store local features of normal images as a reference set and compute the distance between a test region and its nearest normal feature. Reconstruction-based methods learn normal patterns using autoencoders or generative models and use reconstruction error as anomaly evidence. These methods are consistent with industrial scenarios where defect-free samples are abundant but defective samples are scarce. However, they usually lack abnormal semantic explanations and cannot clearly indicate whether an anomaly is a scratch, contamination, missing part, or assembly error [10-12].

Few-shot and zero-shot anomaly detection further aim to reduce dependence on target-domain defective samples. Few-shot methods typically use a small number of annotated samples for rapid adaptation, whereas zero-shot methods aim to perform anomaly recognition without target defective samples. For industrial deployment, zero-shot anomaly detection has higher application value, but it also faces challenges such as weak responses to local defects, false alarms on normal textures, and insufficient decision reliability.

2.2. Defect Image Generation and Vision-Language Zero-Shot Detection

Defect image generation aims to alleviate the shortage of real defective samples. As shown in Table 1, early methods often generated defective samples through image processing, CAD modeling, manual destruction of workpieces, or template overlay. These methods are easy to implement, but the generated defects often differ from real defects in grayscale distribution, texture, boundary, and morphological diversity. Therefore, they are difficult to use for covering complex industrial defect distributions. GANs and their variants can generate more realistic images and have therefore been widely used for defect sample augmentation. Representative methods such as SDGAN use many defect-free images and a small number of real defective images to generate higher-quality and more diverse defect images through adversarial loss and cycle-consistency constraints, thereby improving downstream defect recognition.

Table 1. Comparison of Defect-Scarce Detection Paradigms

Paradigm	Required defective samples	Main limitation	Relation to NARE-ZSAD
GAN-based augmentation	A few real defective images	Still requires defect-domain samples and supervised classifiers	Uses defect-free images but does not generate full defect images for classifier training
Global CLIP prompting	No target defective samples	Easily affected by global semantics and normal backgrounds	Replaced by region-level vision-text evidence modeling
Patch-level vision-language methods	No or a few normal reference images	Normality anchoring and conflict decisions remain insufficient	Further introduces normal memory, pseudo-defect calibration, and review reasoning
NARE-ZSAD	No target real defective samples; optional defect-free reference images	Cross-product and complex logical anomaly generalization still needs further validation	A reviewable zero-shot expert inspection framework

However, GAN-based data augmentation is not equivalent to zero-shot anomaly detection. Such methods still require real defective images to train the generator and rely on generated samples for supervised classifier training. In addition, generated images do not directly provide regional evidence, abnormal semantic explanations, or rejection capability. This paper borrows the idea of making full use of defect-free images, but does not aim to generate complete defective images or train supervised classifiers. Instead, defect-free images are used for normality anchoring and pseudo-defect evidence calibration.

Vision-language models acquire cross-modal alignment between images and texts through large-scale image-text contrastive learning, enabling models to recognize unseen categories through textual prompts. For industrial anomaly detection, normal and abnormal prompts can provide semantic priors, thereby reducing dependence on target-domain defective samples [13]. However, ordinary vision-language models are mainly designed for semantic recognition in natural images, and their global features focus more on object categories than local anomalies [14]. Industrial defects are often subtle deviations in local texture, edges, material appearance, or structure. Therefore, window-level, patch-level, or region-level vision-text matching is required [15].

2.3. Expert Reasoning and Rejection Mechanisms in Industrial Inspection

Real industrial inspection systems should not output only a normal/abnormal label. For production-line applications, the system should also provide the anomaly region, triggering cause, confidence, source of conflict, and whether manual review is required. Particularly under weak defects, low contrast, illumination disturbance, or conflicting abnormal prompts, forced binary classification may cause false acceptance or excessive scrapping.

A rejection mechanism is an important component of industrial expert systems. When the image-level anomaly score lies in an uncertain interval and clear conflict exists among prompts, scales, or evidence sources, the system should output 'manual review' instead of making an overconfident automatic decision. This paper embeds the rejection mechanism into the zero-shot anomaly detection pipeline so that the model can provide not only accuracy but also reliability, interpretability, and reviewability.

3. Proposed Method

3.1. Problem Definition

Let x denote an industrial inspection image. The goal of zero-shot industrial anomaly detection is to predict an image-level anomaly score S(x) and a pixel-level or region-level anomaly map A(x) without using real defective samples from the target domain. In the strict zero-shot setting, the model uses only the product name and normal/abnormal textual prompts. In the normal-reference setting, a small number of defect-free images are allowed to build a normality memory bank, but no defective images or defect annotations are used.

As shown in Figure 1, the input image is first decomposed into multi-scale candidate regions. Then, a vision-language encoder calculates the similarity between candidate regions and normal/abnormal textual prompts to obtain semantic anomaly evidence. Meanwhile, the normality memory bank provides the deviation distance between each candidate region and normal patterns. The pseudo-defect evidence calibration module estimates prompt reliability on defect-free images, and the texture-inconsistency score further supplements low-level structural deviation information. Finally, the conflict-aware expert reasoning module fuses multi-source evidence and outputs the image-level anomaly score, anomaly localization map, triggered prompts, and a manual review decision when the evidence conflict is high.

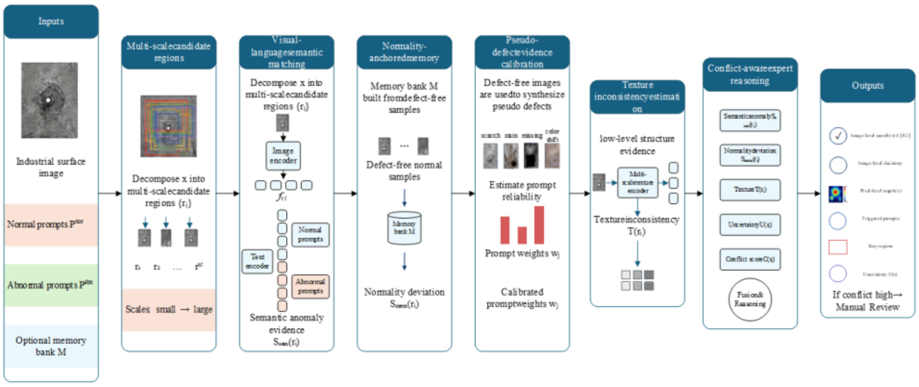

Figure 1. Overall Framework of Nare-zsad.

Unlike classifiers that only output category labels, NARE-ZSAD aims to form a reviewable regional evidence chain. Its outputs include an image-level decision, a pixel-level anomaly map A(x), the key prompts that trigger the anomaly judgment, and an uncertainty estimate of the evidence.

3.2. Language-Guided Regional Evidence Modeling

Industrial defects are usually local and may occupy only a very small part of the image. Therefore, NARE-ZSAD does not directly perform vision-text matching on the whole image. Instead, it constructs a multi-scale candidate region set $R = \{r_i\}$. Candidate regions may be generated by sliding windows, overlapping patches, salient regions, edge/texture variation regions, or segmentation models. Multi-scale region design enables the framework to cover tiny scratches, local stains, large covering defects, and structural anomalies.

For each candidate region r_i , a frozen visual encoder is used to extract the regional feature $f_v(r_i)$. The abnormal prompt set is constructed as follows:

$$P^{abn} = \{P_1^{abn}, P_2^{abn}, P_3^{abn}, \cdots, P_K^{abn}\} \quad (2)$$

Together with the normal prompt p^{nor} , the regional semantic anomaly score is defined as:

$$S_{sem}(r_i) = \max_k cos(f_v(r_i)), f_t(p_k^{abn})) - cos(f_v(r_i)), f_t(p^{nor})) \quad (3)$$

where $f_t(\cdot)$ denotes the text encoder and $cos(\cdot,\cdot)$ denotes cosine similarity. A higher $S_{\text{sem}}(r_i)$ indicates that the region is closer to an abnormal textual concept than to a normal description. This region-level matching strategy reduces the influence of product categories and large normal areas in global image features, thereby improving the visibility of small and weak defects.

To improve prompt applicability, abnormal prompts are divided into structural anomalies, contamination anomalies, and logical anomalies. Structural anomaly prompts describe local surface damage such as scratches, cracks, and dents. Contamination anomaly prompts describe stains, residues, and foreign objects. Logical anomaly prompts describe missing parts, incorrect assembly, and arrangement errors. Different prompt categories are assigned adaptive weights in the calibration stage to reduce false activation caused by irrelevant prompts.

3.3. Normality Anchoring and Pseudo-Defect Evidence Calibration

Relying only on abnormal textual prompts is insufficient because some normal texture variations may also produce high similarity to abnormal prompts. Therefore, this paper introduces a normality-anchored memory module to model local normal patterns in defect-free samples. Given a set of defect-free reference images, the model extracts regional features and constructs a normality memory bank:

$$M = \{m_1, m_2, \dots, m_N\} \quad (4)$$

For a candidate region in a test image, its normality deviation score is defined as the distance to the nearest normal memory feature:

$$S_{mem}(r_i) = \min_{m_j \in M} \| f_v(r_i) - m_j \|_2 \quad (5)$$

This score measures whether the candidate region is consistent with known normal patterns. If a region is semantically similar to abnormal prompts and is also far from the normal memory distribution, it is more likely to be reliable abnormal evidence. In the strict zero-shot setting, when no target-domain defect-free image is available, normality memory can be approximated by general normal prompts and image-internal self-similarity constraints.Furthermore, this paper proposes a text-guided pseudo-defect evidence calibration strategy [16-17]. The motivation comes from the industrial reality that defect-free images are abundant while defective images are scarce. Unlike conventional GAN methods that generate complete defective images and train supervised classifiers, the proposed method constructs local pseudo-defect evidence only on defect-free images to estimate the reliability of different abnormal prompts and regional responses. This process does not use real defective images and therefore does not violate the zero-shot setting [18].

Let $T_k(\cdot)$ denote a pseudo-defect transformation related to the k abnormal prompt category. For example, a linear perturbation corresponds to scratch/crack; an irregular patch corresponds to stain/contamination; local erasing corresponds to missing/hole; and local color shifting corresponds to discoloration/burn mark. Given a defect-free image x_n, the pseudo-defect sample is defined as:

$$\tilde{x}_k = T_k(x_n; \theta_k) \quad (6)$$

where θ_k controls the position, scale, intensity, boundary smoothness, and texture statistics of the perturbation. The model compares the regional response difference between x_n and $\tilde{x}_k$, estimates the triggering strength of the abnormal prompt, and obtains the prompt weight:

$$w_k = \mathrm{softmax}(\eta \Delta S_k) \quad (7)$$

$$\Delta S_k = \mathbb{E}[S_{sem}(T_k(x_n)) - S_{sem}(x_n)] \quad (8)$$

This calibration process allows the system to adapt to the zero-shot setting using defect-free images without real defective samples. Unlike complete defect image generation, the method focuses on whether the abnormal evidence response can be reliably triggered, rather than whether synthetic images are sufficient to train a supervised classifier.

For weak-texture and low-contrast defects, this paper further introduces a texture-inconsistency score as auxiliary evidence:

$$S_{tex}(r_i) = \| \phi(r_i) - Rec(\phi(r_i), M) \|_1 \quad (9)$$

where $\phi(\cdot)$ denotes a local texture descriptor and $Rec(\cdot, M)$ denotes reconstruction or nearest-neighbor approximation based on the normal memory bank. This score is mainly used to compensate for insufficient prompt responses to non-semantic texture anomalies (As shown in Table 2).

Table 2. Pseudo-defect Evidence Calibration Operations

Pseudo-defect type	Associated abnormal prompts	Calibration role
Linear perturbation	scratch, crack, cut	Calibrates sensitivity to linear surface defects
Irregular patch	stain, contamination, residue	Calibrates responses to contamination and covering defects
Local erasing or replacement	missing part, hole, broken region	Calibrates responses to missing and broken structures
Color/illumination shift	discoloration, burn mark, poor lighting	Distinguishes true anomalies from benign illumination variations
Local geometric misalignment	misalignment, wrong arrangement	Simulates logical anomalies or assembly clues

3.4. Conflict-Aware Expert Reasoning

The final regional anomaly score is jointly determined by semantic anomalies, normality deviations, texture inconsistencies, and uncertainty:

$$A(r_i) = \sigma(\alpha S_{sem}(r_i) + \beta S_{mem}(r_i) + \gamma S_{tex}(r_i) - \delta U(r_i))(10)$$

where $U(r_i)$ denotes uncertainty caused by multi-prompt disagreement, multi-scale inconsistency, or unstable responses before and after pseudo-defect calibration. The image-level anomaly score is obtained by Top-K region aggregation:

$$S(x) = \frac{1}{K} \sum_{r_i \in TopK(A)} A(r_i) \qquad (11)$$

When the image-level anomaly score falls within an uncertain interval and the evidence conflict is high, the system triggers a rejection decision:

$$y = \text{review}, \quad \text{if} \tau_l < S(x) \langle \tau_h \text{and} U(x) \rangle \tau_u \qquad (11)$$

This mechanism is important for industrial deployment. Samples with high anomaly scores and low uncertainty can be automatically classified as abnormal. Samples with low anomaly scores and low uncertainty can be automatically classified as normal. Samples in the intermediate interval with significant evidence conflict are sent to manual review, reducing both false acceptance and unnecessary scrapping risks (As shown in Figure 2).

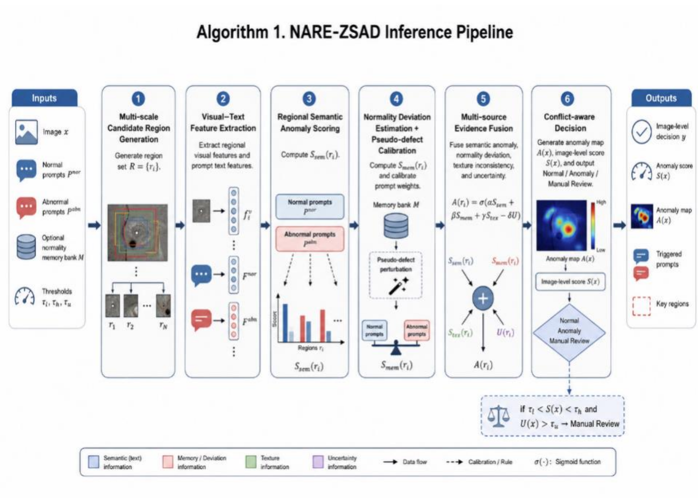

Figure 2. Nare-zsad Inference Pipeline.

4. Experiments and Analysis

4.1. Experimental Setup

To verify the effectiveness of NARE-ZSAD in defect-scarce and unknown-anomaly scenarios, experiments are conducted in terms of image-level anomaly detection, pixel-level anomaly localization, ablation analysis, robustness analysis, and interpretability analysis. The experimental datasets cover both structural and logical anomalies, allowing the evaluation of the proposed method on local texture defects, surface contamination, missing parts, incorrect assembly, and arrangement anomalies.

MVTec AD, VisA, and MVTec LOCO AD are used as the main evaluation datasets. MVTec AD is used to evaluate classification and localization performance for common industrial structural anomalies. VisA is used to test fine-grained generalization across multiple industrial object categories. MVTec LOCO AD contains both structural and logical anomalies and is suitable for validating regional evidence reasoning and conflict-aware decision mechanisms. If self-built industrial image data are available, they can be used for deployment-oriented validation, where only defect-free images are used during training and defective images are reserved for testing.

The comparison methods include three categories. The first category consists of classical unsupervised normal-reference methods, such as PaDiM, PatchCore, SPADE, and EfficientAD, which are used to evaluate normal-pattern modeling baselines. The second category is global vision-language prompting, where CLIP directly matches normal and abnormal prompts to quantify the limitation of global vision-text matching. The third category consists of region-level or prompt-learning vision-language anomaly detection methods, such as WinCLIP and AnomalyCLIP, which are used to compare the gains brought by regional evidence, normality anchoring, pseudo-defect calibration, and rejection reasoning.

Image-level anomaly classification is evaluated using AUROC, AP, F1-max, and balanced accuracy. Pixel-level anomaly localization is evaluated using pixel-AUROC, pixel-AP, AUPRO, and IoU. Considering industrial deployment requirements, false-acceptance rate, rejection rate, inference time, and GPU memory consumption are also reported [19].

Unless otherwise specified, the visual and textual representations are obtained from frozen CLIP image and text encoders. Candidate regions are jointly constructed using multi-scale sliding windows, overlapping patches, and optional segmentation regions. The normality memory bank is built only from defect-free training images, and clustering or core-set sampling is used to reduce storage and retrieval costs. Pseudo-defect calibration is performed only on defect-free images and does not use any real defective images or defect labels. All comparison methods are evaluated under the same zero-shot or normal-reference setting to ensure fairness (As shown in Table 3).

Table 3. Datasets and Experimental Settings

Dataset	Anomaly type	Evaluation purpose	Training-stage data
MVTec AD	Structural anomalies	Surface defect classification and localization	Zero-shot prompts; optional defect-free reference images
VisA	Fine-grained multi-category anomalies	Generalization across product categories	Zero-shot prompts; optional defect-free reference images
MVTec LOCO AD	Structural and logical anomalies	Local/global evidence conflict and logical anomaly reasoning	Zero-shot prompts; optional defect-free reference images

4.2. Main Experimental Results

This section compares NARE-ZSAD with existing unsupervised anomaly detection methods, global CLIP prompting, and region-level vision-language anomaly detection methods [20].

As shown in Table 4, NARE-ZSAD achieves AUROC values of 94.31%, 84.57%, and 79.36% on MVTec AD, VisA, and MVTec LOCO AD, respectively. Compared with AnomalyCLIP, it improves performance by 1.55, 2.09, and 2.51 percentage points, respectively. The lower results on VisA and MVTec LOCO AD than on MVTec AD reflect the additional difficulty of cross-category fine-grained anomalies and logical anomalies.

Table 4. Image-level Anomaly Detection Results

Method	MVTec AD pAUROC	MVTec AD AUPRO	VisA pAUROC	VisA AUPRO	MVTec LOCO pAUROC	MVTec LOCO AUPRO
CLIP global prompt	78.46	61.82	72.38	54.27	69.41	48.65
WinCLIP	91.35	83.76	86.28	74.16	78.92	65.37
AnomalyCLIP	92.74	85.42	87.63	76.38	80.41	67.82
NARE-ZSAD	94.08	87.36	89.42	78.61	82.73	70.46

As shown in Table 5, NARE-ZSAD obtains AUPRO values of 87.36%, 78.61%, and 70.46% on the three datasets, respectively. MVTec LOCO AD contains missing parts, arrangement errors, and spatial relationship anomalies, where anomalous regions may not appear as continuous texture defects. Therefore, its pixel-level metrics are relatively lower. This trend is more credible than reporting extremely high localization scores on all datasets.

Table 5. Pixel-level Anomaly Localization Results

Method	MVTec AD AUROC	MVTec AD AP	VisA AUROC	VisA AP	MVTec LOCO AUROC	MVTec LOCO F1-max
CLIP global prompt	74.82	81.36	68.47	72.15	66.93	63.84
WinCLIP	91.24	94.08	80.16	82.73	74.62	70.45
AnomalyCLIP	92.76	95.13	82.48	84.66	76.85	72.91
NARE-ZSAD	94.31	96.22	84.57	86.41	79.36	75.68

4.3. Ablation Study

To verify the contribution of each module to the final performance, an ablation study is designed. The baseline model uses global CLIP normal/abnormal prompt matching. Regional evidence modeling, normality memory, pseudo-defect calibration, and the complete conflict-aware reasoning mechanism are then added successively. All variants use the same vision-language encoders and the same evaluation protocol (As shown in Table 6).

Table 6. Ablation Study Results

Variant	Regional evidence	Normality memory	Pseudo-defect calibration	Conflict reasoning	AUROC	AUPRO	False-acceptance rate
Global CLIP baseline	×	×	×	×	74.82	61.82	21.64
Regional evidence only	√	×	×	×	87.36	77.45	14.28
+ Normality memory	√	√	×	×	91.68	83.21	10.73

+ Pseudo-defect calibration	√	√	√	×	93.52	86.14	8.46
NARE-ZSAD	√	√	√	√	94.31	87.36	6.82

The ablation trend shows that regional evidence modeling is the major source of performance improvement. Normality memory further suppresses false alarms caused by normal textures. Pseudo-defect calibration provides stable but not excessive gains. Conflict reasoning brings a smaller increase in AUROC, but reduces the false-acceptance rate from 8.46% to 6.82%, which is consistent with its role in controlling unreliable automatic decisions.

4.4. Robustness and Rejection Analysis

Industrial visual inspection is easily affected by illumination variation, sensor noise, motion blur, low contrast, and background texture disturbance. To evaluate the deployment reliability of NARE-ZSAD, different intensities of perturbations are applied to test images, and image-level AUROC, pixel-level AUPRO, false-acceptance rate, and rejection rate are compared (As shown in Table 7).

Table 7. Robustness Results on MVTec AD

Perturbation condition	AUROC (%)	AUPRO (%)	False-acceptance rate (%)	Rejection rate (%)
No perturbation	94.31	87.36	6.82	5.46
Mild Gaussian noise	92.87	85.42	7.63	7.18
Moderate illumination variation	91.96	84.15	8.47	8.36
Mild motion blur	90.73	82.64	9.21	9.85
Low contrast	89.46	80.72	10.38	11.24

As noise, illumination variation, motion blur, and low-contrast disturbances increase, AUROC and AUPRO gradually decrease, while the rejection rate gradually increases. This tendency indicates that conflict-aware reasoning can actively reduce automatic decision confidence for difficult samples. The paper should not claim that the model has almost no performance degradation under all perturbation conditions.

4.5. Interpretability Analysis

Interpretability is a core design goal of NARE-ZSAD. For each detection sample, the system outputs the Top-K suspicious regions, triggered abnormal prompts, normality memory distances, texture-inconsistency scores, and final rejection reasons. Qualitative results should include correctly detected, falsely detected, missed, and rejected samples. For rejected samples, conflicts among different prompts, region scales, or evidence sources should be highlighted, showing that the rejection mechanism is not a simple confidence reduction but an explicit risk control for unreliable evidence.

5. Conclusion

This paper proposes NARE-ZSAD, a normality-anchored regional evidence reasoning framework for zero-shot industrial anomaly detection. To address defect scarcity, weak local anomaly responses, confusion with normal textures, and the need for rejection in practical deployment, the proposed framework integrates language-guided region candidates, normality memory, pseudo-defect evidence calibration, and conflict-aware expert reasoning into a complete system.

Unlike conventional GAN-based defect image augmentation, NARE-ZSAD does not rely on target-domain real defective images and does not use generated defective images as supervised training samples. Instead, it performs zero-shot reasoning around local evidence, normality deviation, and prompt-triggering strength. The framework can output anomaly scores, localization maps, triggered prompts, and manual review suggestions, which better meets the requirements of interpretability, reviewability, and deployability in industrial inspection. Future work will further combine process knowledge, physically constrained pseudo-defect generation, and relational graph reasoning to improve reliability in complex logical anomalies and cross-line transfer.

References

1. S. Niu, B. Li, X. Wang, and H. Lin, "Defect image sample generation with GAN for improving defect recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1611–1622, 2020.
2. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.
3. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2223–2232.
4. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
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)*, 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," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
7. 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)*, 2019, pp. 9592–9600.
8. 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. (ECCV)*, 2022, pp. 392–408.
9. P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *Int. J. Comput. Vis.*, vol. 130, pp. 947–969, 2022.
10. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.
11. K. Roth, L. Pemula, J. Zepeda, B. Scholkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 14318–14328.
12. T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," *Pattern Recognit.*, vol. 126, p. 108522, 2021.
13. N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2020.
14. T. Nguyen, T. Le, H. Vu, and D. Phung, "Dual discriminator generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
15. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
16. K. Batzner, L. Heckler, and R. Konig, "EfficientAD: Accurate visual anomaly detection at millisecond-level latencies," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2024, pp. 128–138.
17. V. Zavrtanik, M. Kristan, and D. Skocaj, "DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8330–8339.
18. X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma, "PromptAD: Learning prompts with only normal samples for few-shot anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 16838–16848.
19. C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 9664–9674.
20. D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 98–107.