Zero-Shot Industrial Anomaly Detection Based on Category-Aware Adaptive Vision-Language Agent

Qianlin Lei[1,2], Li Li[1,2,*], and Zhanpeng Wang[1,2]
[1]Southwest University of Science and Technology, Mianyang, China
[2]Sichuan Engineering Technology Research Center of Industrial Self-Supporting and Artificial Intelligence, Mianyang, China
[*]Corresponding author: llki2000@163.com

***Abstract*—While Large Vision-Language Models (VLMs) demonstrate remarkable cognitive capabilities in general scenarios, they face fundamental perceptual barriers when handling physics-constrained industrial extremes, such as micron-level deformations and high-frequency illumination noise. This paper identifies a critical "Semantic-Resolution Gap" in current generalist models. To address this, we propose a Category-Aware Hybrid Perception Framework (HPF). This framework constructs an adaptive agent capable of dynamically restructuring inference paths based on the physical attributes of the object: (1) utilizing a single-shot reference mechanism to counteract reflective noise on metal surfaces; (2) employing a positional prior-guided physical zooming mechanism to overcome perceptual bottlenecks in microstructures; and (3) combining signal enhancement with semantic parsing to handle low-contrast defects. Furthermore, we introduce a Neuro-Symbolic Logic Calibrator to map probabilistic textual outputs into deterministic industrial-grade verdicts, effectively mitigating the model's "Response Inertia." Experiments on the MVTec AD dataset show that our HPF method, based on the lightweight Phi-3.5-Vision (3.8B), achieves an average accuracy of 74.67%, with a peak accuracy of 80.87% on challenging categories like Metal Nut, significantly outperforming the 7B-parameter SOTA baseline LLaVA-1.5. The proposed method enables cold starts with only a single golden sample and achieves efficient quasi-real-time inference on resource-constrained edge computing devices, reducing memory usage to $\sim$3.2 GB.**

***Index Terms*—Industrial Anomaly Detection, Vision-Language Models, Adaptive Perception, Edge Computing, Zero-Shot Learning**

I. Introduction

A. Background and Motivation

In the field of intelligent manufacturing, Automated Optical Inspection (AOI) is a critical technology for ensuring product quality and production stability. With the miniaturization of electronic components and the increasing precision of mechanical processing, the demand for detecting micron-level surface defects (e.g., scratches, cracks, deformations) is becoming increasingly urgent. Traditional machine vision solutions primarily rely on manual feature engineering or fully supervised deep learning models. Although supervised models perform excellently on specific datasets, their training process is highly dependent on large-scale, high-quality pixel-level annotated data. However, in modern mature production lines, the defect rate is extremely low (typically less than 0.1%), leading to severe class imbalance. This makes building robust fully supervised models face the dual challenges of high data acquisition costs and statistical scarcity of samples.

B. Limitations of Existing Technologies

To address the issue of sample scarcity, feature embedding-based Unsupervised Anomaly Detection (UAD) has gradually become the mainstream research direction [1]. Methods represented by PatchCore [2] and PaDiM [3] utilize the feature distribution of normal samples to perform anomaly discrimination based on distance metrics. Although such methods have achieved significant results in fixed scenarios, they essentially belong to the "One-Class Learning" paradigm: whenever a New Product Introduction (NPI) occurs or the process is fine-tuned, the system must re-collect normal samples and update the memory bank. This dependence on target domain data leads to "cold start" delays, severely restricting application flexibility in High-Mix Low-Volume (HMLV) flexible manufacturing scenarios.

C. Potential and Challenges of VLMs

Recently, breakthroughs in Large Vision-Language Models (VLMs) have provided a new technical path for Zero-Shot Anomaly Detection (ZSAD). Benefiting from pre-training on massive image-text data, VLMs possess powerful cross-modal semantic alignment capabilities, theoretically allowing for Training-Free inference via natural language prompts. However, transferring general-domain VLMs to specific industrial inspection tasks involves a significant Domain Shift, mainly reflected in two dimensions:

- **Conflict between Resolution and Feature Granularity:** General VLM visual encoders (e.g., ViT) typically use fixed input resolutions (e.g., 336×336). This downsampling process acts as a low-pass filter, causing information loss for micro-geometric defects common in industrial scenarios (e.g., transistor pin deviations, hairline cracks in glass) during the feature extraction stage.
- **Semantic Ambiguity and Misjudgment:** Visual features in industrial scenarios differ distributively from natural scenes. For instance, Specular Reflection on metal surfaces appears as high-frequency bright areas in pixel statistics, which models lacking industrial priors easily misclassify as scratches. Furthermore, models aligned via Reinforcement Learning from Human Feedback (RLHF) tend to generate conservative descriptive responses, making it difficult to directly output binary (Pass/Fail) verdicts required by industry standards.

D. Contributions

In view of these challenges, we argue that a single static prompting strategy cannot account for the complex and variable physical attributes in industrial scenarios. We propose a Category-Aware Hybrid Perception Framework (HPF), aimed at bridging the gap between general large models and industrial applications through cognitive decoupling and dynamic routing mechanisms. As shown in Fig. 1, the framework constructs an agent with physical perception capabilities, adaptively scheduling optimal perception and inference strategies based on the object's material (reflective/matte), structure (precision/macro), and texture features. The main contributions are summarized as follows:

1) **Attribute-Driven Dynamic Perception Architecture:** We construct an inference system with three decoupled paths. For high-reflection interference, a single-sample differential inference strategy is proposed to effectively suppress environmental noise using a golden reference; for micro-structures, a physical zooming mechanism based on positional prior is designed to significantly improve sensitivity to geometric deformations.
2) **Neuro-Symbolic Logic Calibration Mechanism:** Addressing the uncertainty of VLM outputs, we design a deterministic rule engine based on syntactic analysis and domain lexicons. This maps probabilistic natural language descriptions into executable industrial verdicts, effectively solving the model's "Response Inertia" problem.
3) **Efficient Edge Deployment Verification:** Experiments show that our method achieves SOTA performance on multiple challenging categories of the MVTec AD dataset under a completely zero-shot setting without parameter fine-tuning. We also verify its engineering feasibility for quasi-real-time inference on resource-constrained edge devices (e.g., consumer-grade GPUs), reducing memory usage to $\sim$3.2 GB.

Open Science Statement: The project repository is located at https://github.com/2593327083/HPF-ZSAD.

II. Related Work

A. Unsupervised Industrial Anomaly Detection

Early deep learning approaches in industrial inspection focused on reconstruction-based methods (AE, GAN), assuming normal samples lie on a low-dimensional manifold. For instance, DRAEM [4] and FastFlow [5] tackle reconstruction and normalizing flows respectively. While independent of defect samples, they often suffer from "over-generalization," where models erroneously repair abnormal regions. Recently, feature embedding-based methods like UniAD [6], SimpleNet [7], SPADE [8], PaDiM [3], and PatchCore [2] established SOTA standards by modeling normal feature spaces using pre-trained CNNs. However, as One-Class Learning methods, they require re-collecting hundreds of normal samples for every new SKU, failing to meet the "plug-and-play" requirements of flexible manufacturing.

B. Foundations of Vision-Language Models

VLMs bridge the modal barrier between pixels and semantics. CLIP [9] aligned visual and text encoders via contrastive learning, enabling zero-shot classification. This was further expanded by BLIP [10] and Flamingo [11] for broader vision-language understanding, as well as foundational vision models like SAM [12]. Generative multimodal models like LLaVA [13], MiniGPT-4 [14], and Phi-3-Vision [15] (used in this work) introduce LLMs as decoders, offering stronger reasoning capabilities. However, their visual encoders are designed for global semantics, and the fixed resolution (typically 336×336) constitutes a physical bottleneck for industrial micro-defects. Table I compares different architectures.

TABLE I
COMPARISON OF VLM ARCHITECTURES IN ZERO-SHOT AD TASKS

Model Type	Visual Encoder	Resolution	Reasoning
CLIP	ViT-L/14	Fixed (224/336)	Low
LLaVA	ViT-L/14	Fixed (336)	High
Phi-3.5	SigLIP	Adaptive	High

C. Zero-Shot AD and Prompt Engineering

Zero-Shot AD (ZSAD) aims to identify anomalies without target domain data. Works like WinCLIP [16] and AnomalyCLIP [17] use State-Aware Prompting. Furthermore, prompt learning strategies such as CoOp [18] highlight the importance of adaptable text inputs. However, existing ZSAD methods generally adopt a Static Inference Strategy, applying unified prompts to all objects. This "one-size-fits-all" approach ignores the orthogonality of industrial physical attributes (e.g., transparency vs. reflectivity). Our HPF framework addresses the contradiction between "static strategies" and "dynamic scenarios."

III. Methodology

This section details the Category-Aware Hybrid Perception Framework (HPF). Our core design philosophy is "Cognitive Decoupling": decomposing the complex zero-shot detection task into independent inference sub-tasks targeting specific physical attributes (Reflection, Geometry, Texture), and achieving reliable verdicts through neuro-symbolic collaboration.

A. Overall Architecture

Unlike traditional end-to-end mappings $y = M(I, P)$, HPF is built as an adaptive agent comprising perception, decision-making, and execution phases. As shown in Fig. 1, the system includes three cascaded core modules:

- **Category-Aware Router:** Routes input to the optimal perception path based on object attributes.
- **Multi-Path Adaptive Perception:** Contains three parallel physical-level visual transformation and inference paths.
- **Neuro-Symbolic Logic Calibrator:** Maps probabilistic outputs to deterministic verdicts.

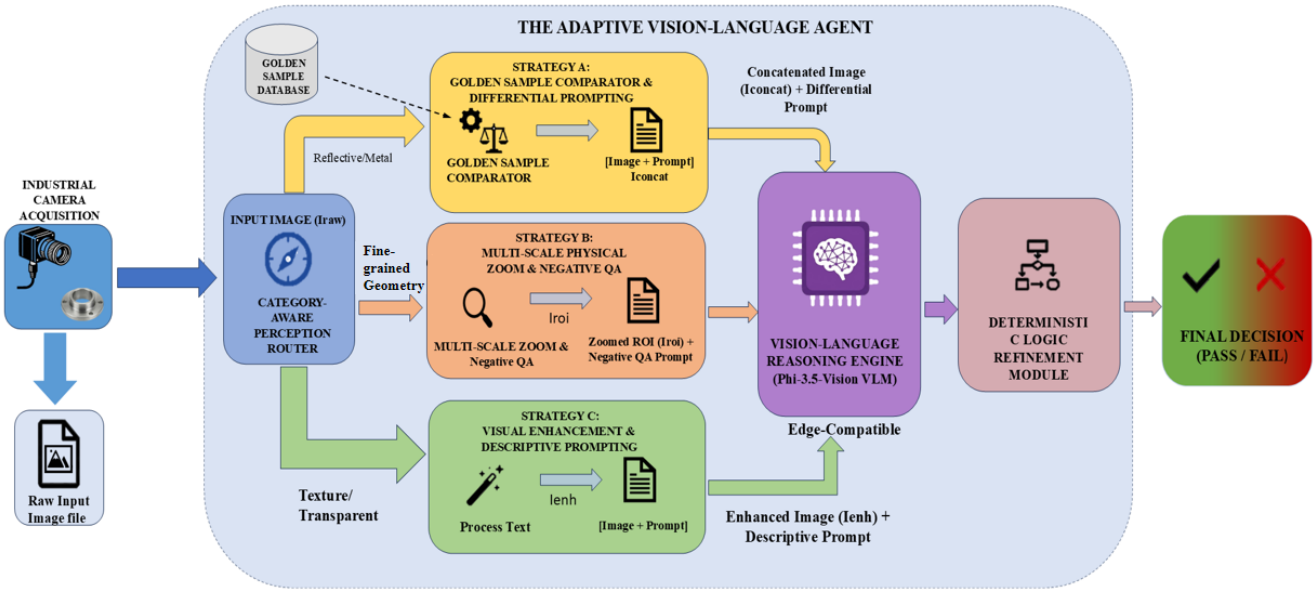

Fig. 1. Overall Architecture of the Category-Aware Hybrid Perception Framework. The system dynamically activates parallel perception paths (Strategies A/B/C) via a Perception Router based on physical attributes, and finally outputs deterministic Pass/Fail verdicts through a Neuro-Symbolic Logic Refinement Module.

B. Problem Formulation

We formulate ZSAD as a binary classification problem with a domain gap. Given a query image $x_q \in \mathcal{X}_{target}$, we aim to predict its label $y \in \{0, 1\}$ (0 for normal, 1 for anomaly) using a model M pre-trained on a source domain $\mathcal{X}_{source}$, without access to any samples from $\mathcal{X}_{target}$. A general VLM inference process can be expressed as:

$$y = \mathbb{I}(S(x_q, P) > \tau), \tag{1}$$

where P is a static text prompt, $S(\cdot)$ is the anomaly score output by the VLM, τ is the decision threshold, and $\mathbb{I}(\cdot)$ is the indicator function. In HPF, we introduce a routing function $R(\cdot)$ and a transformation function $T(\cdot)$, making the input dynamic:

$$y = M_{VLM}(T(x_q, R(x_q)), P_{adaptive}), \tag{2}$$

where $R(x_q)$ determines the strategy type, T applies corresponding physical transformations (e.g., zooming, enhancement), and $P_{adaptive}$ is the prompt matched to the strategy.

C. Category-Aware Routing Mechanism

1) Rationale based on VLM Limitations: Predicated on the intrinsic perceptual bottlenecks identified in generalist VLMs when processing industrial images, we define a taxonomy of three orthogonal **Physical Domains** for industrial objects:

- S_{ref} **(High-Frequency Reflective Domain):** Targets metal parts with strong specular reflection (e.g., Metal Nut). *Rationale:* The VLM's visual encoder lacks physical modeling of specular reflection and easily misinterprets high-frequency light spots as textural defects (false positives).
- S_{geo} **(Micro-Geometric Domain):** Targets components with precision micro-structures (e.g., Transistor). *Rationale:* The VLM's fixed input resolution causes Aliasing effects during down-sampling of micron-level deformations (e.g., pin bending), necessitating physical zooming.
- S_{tex} **(Low-SNR Texture Domain):** Targets transparent/organic materials (e.g., Bottle, Hazelnut). *Rationale:* The edge gradients of such defects are extremely weak and often submerged in background noise, requiring signal enhancement.

2) Routing Implementation: The routing function $R(x)$ handles dispatching. In this study, we employ a **Metadata-Driven** routing strategy for experiments, utilizing known category tags (e.g., 'Transistor') to activate corresponding strategies. This is done to accurately evaluate the effectiveness of the perception strategies themselves and exclude interference from upstream classifiers. For deployment scenarios without tags, a statistical heuristic routing based on histogram entropy and highlight pixel ratios is proposed as an alternative.

D. Adaptive Perception Strategies

1) Strategy A: Reference-Guided Differential Inference: **Target Domain:** S_{ref} (High-reflection metal). **Mechanism Analysis:** Specular reflection noise $\nabla I_{specular}$ in images often mimics the high-frequency gradient characteristics of scratches $\nabla I_{scratch}$. General VLMs find it difficult to distinguish between the two in a single forward pass. We introduce a "Visual Anchor" using a single golden sample I_{ref}. By constructing a differential input, the model is forced to focus on the differences (residuals) between the test image and the standard

image, effectively canceling out common environmental reflection components. **Implementation:** We construct a composite input tensor I_{input} by concatenating the reference image and the test image along the width dimension:

$$I_{input} = \text{Concat}(I_{ref}, I_{test}, \text{dim} = W). \quad (3)$$

The corresponding prompt guides the model to perform a comparison: "Task: Compare Left (Reference) vs Right (Test). Question: Is the Right image exactly the same quality as the Left?" The schematic is shown in Fig. 2 (Subplot A).

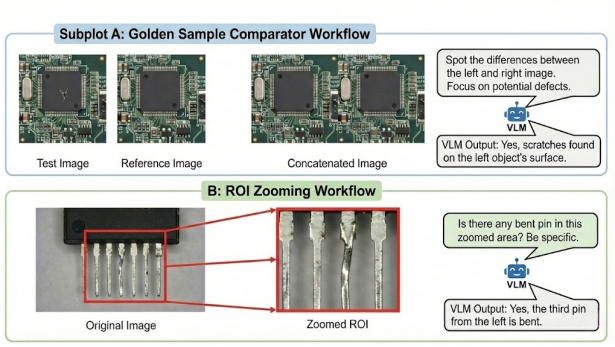

Fig. 2. Schematic of Adaptive Perception Strategies. Subplot A: Reference-Guided Differential Inference for reflective surfaces. Subplot B: Positional Prior-Guided Physical Zooming for micro-structures.

2) Strategy B: Positional Prior-Guided Physical Zooming: **Target Domain:** S_{geo} (Micro-structures). **Mechanism Analysis:** The standard resizing operation (to 336×336) of VLMs acts as a low-pass filter, causing irreversible loss of high-frequency spatial information of micro-defects. We propose a "Physical Zooming" strategy that trades field of view for resolution. **Implementation:** Leveraging the **Mechanical Alignment Prior** in industrial lines (key components are centered), we define a Center-Retention Cropping operator. Let the raw image be $I_{raw} \in \mathbb{R}^{W \times H}$. We introduce a retention ratio coefficient ρ. For micro-structures (e.g., Transistor), to maximize field of view coverage while removing edge noise, we set $\rho = 0.95$; whereas for center-focused objects (e.g., Bottle), to magnify internal details, we set $\rho = 0.65$. The cropping box B is defined as:

$$B = [c_x - \frac{\rho W}{2}, c_y - \frac{\rho H}{2}, c_x + \frac{\rho W}{2}, c_y + \frac{\rho H}{2}]. \quad (4)$$

The cropped region I_{crop} is then upsampled back to the model input size via bilinear interpolation. Mathematically, this increases the **Effective Pixel Density** of the target region by $1/\rho^2$. We employ Negative Inductive QA, e.g., "Focus on the metal legs. Is there any bent, broken, or missing pin?"

3) Strategy C: Signal Enhancement & Semantic Parsing: **Target Domain:** S_{tex} (Low-contrast/Transparent). **Implementation:** We introduce a deterministic signal enhancement operator T_{enh} comprising Contrast Limited Adaptive Histogram Equalization (CLAHE) and Laplacian sharpening.

$$I_{enh} = T_{enh}(I_{raw}) = I_{clahe} - \lambda \cdot \nabla^2(I_{clahe}), \quad (5)$$

where λ is the sharpening coefficient (set to 2.0). This non-linearly amplifies edge gradients, making implicit cracks explicit. The prompt emphasizes semantic parsing of specific defect types: "List any cracks, chips, contamination or holes."

E. Neuro-Symbolic Logic Calibrator

The raw text output T_{raw} from VLMs often contains linguistic uncertainty (e.g., "It seems okay, but..."), which does not meet the strict binary determinism (Pass/Fail) required by industry. We design a two-stage symbolic rule engine to calibrate this probabilistic output.

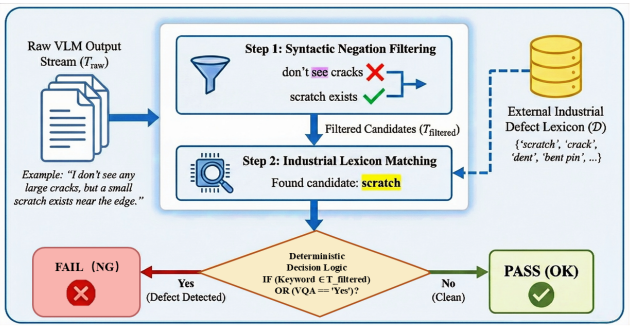

Fig. 3. Workflow of the Neuro-Symbolic Logic Calibrator. Purple highlights indicate syntactic scope analysis (filtering negated descriptions), and yellow highlights indicate semantic matching with the industrial defect lexicon.

As shown in Fig. 3, the process includes:

1) **Syntactic Negation Filtering:** Using dependency parsing to identify the Scope of Negation. For example, in "No cracks are visible," the noun "cracks" is masked as it is modified by "No."
2) **Industrial Lexicon Mapping:** Valid text is matched against a pre-defined domain defect lexicon $\mathcal{D} = \{scratch, bent, dent, contamination, ...\}$.
3) **Deterministic Decision Function:** A "Default-OK, Explicit-NG" policy is adopted.

$$y = \begin{cases} \text{FAIL}, & \text{if } (T_{filtered} \cap \mathcal{D} \neq \emptyset) \vee (\text{VQA} = \text{"Yes"}) \\ \text{PASS}, & \text{otherwise} \end{cases} \quad (6)$$

This mechanism effectively mitigates the model's **"Response Inertia"**, ensuring reliable industrial verdicts.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluate on four representative categories from the MVTec AD dataset [19], covering different physical challenges: Metal Nut (Reflective), Transistor (Geometric), Hazelnut (Texture), and Bottle (Transparent). Due to constraints in hardware resources and inference time, we strategically selected these four categories as they represent the most challenging physical extremes in AOI, effectively testing the proposed cognitive decoupling mechanisms.

2) Hardware Environment: All experiments are conducted on a single consumer-grade **NVIDIA GeForce RTX 3050 Laptop GPU (6GB VRAM)** to strictly verify feasibility for edge deployment. We use **4-bit NF4 quantization** technology to compress model weights, reducing memory usage from ∼7.6 GB to ∼**3.2 GB**.

3) Model Configuration: The baseline model is the raw **Phi-3.5-Vision-Instruct** (3.8B parameters) employing a static universal prompt ("Is there any defect in this image?"). The competitor is the larger generalist model **LLaVA-1.5-7B**, running under the same 4-bit quantization setting for fair comparison.

B. Quantitative Evaluation: Multi-Dimensional Performance Analysis

To rigorously evaluate the effectiveness of the HPF framework, we conducted a comparative study against the Internal Baseline and the External Competitor (LLaVA-1.5-7B). The quantitative results are summarized in Table II and visualized in Fig. 4.

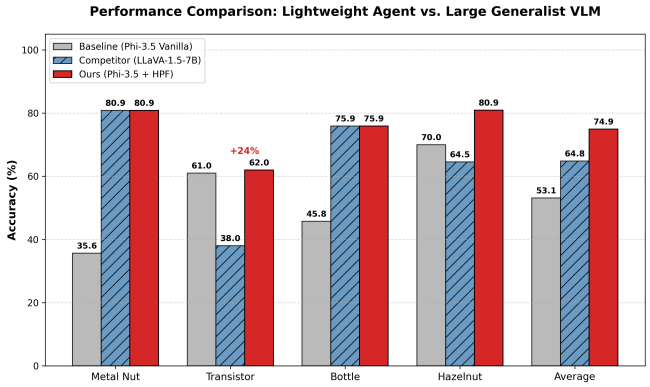

Fig. 4. **Quantitative Performance Comparison.** The red bars (Ours) show consistent superiority across all categories compared to the 7B-parameter LLaVA-1.5 (Blue), particularly highlighting a **+24%** leap in the Transistor category due to the physical zooming strategy overcoming resolution bottlenecks.

TABLE II
PERFORMANCE COMPARISON (ACCURACY %) ON MVTEC-AD: BASELINE VS. COMPETITOR VS. OURS (ALL IN 4-BIT)

Category	**Baseline** (Phi-3.5 Vanilla)	**Competitor** (LLaVA-1.5-7B)	**Ours** (Phi-3.5 + HPF)
Metal Nut	35.65	80.87	**80.87**
Transistor	61.00	38.00	**61.00**
Bottle	45.78	75.90	**75.90**
Hazelnut	70.00	64.55	**80.91**
Average	53.11	64.83	**74.67**

The experimental data reveal several key findings:

1) Core Finding 1: Overcoming the "Resolution Blind Zone" for Micro-Structures: On the most challenging **Transistor** category, the generalist large model LLaVA-1.5-7B shows a significant performance collapse, with an accuracy of only **38.00%**, which is even below random guessing. This confirms that mere parameter scaling cannot solve the "Aliasing Effect" caused by fixed input resolution—micron-level pin deformations are completely lost during down-sampling. In contrast, the HPF framework, leveraging **Strategy B (Physical Zooming)**, achieves an accuracy of **61.00%** with only 3.8B parameters. This significant **+24%** advantage demonstrates that in industrial micro-structure detection, **explicit physical attention mechanisms** are far more decisive than implicit model capacity.

2) Core Finding 2: Parameter Efficiency ("Doing More with Less"): On the **Metal Nut** and **Bottle** categories, our method achieves accuracies of **80.87%** and **75.90%** respectively, reaching complete parity with the 7B-parameter LLaVA model. This result has important engineering implications: it shows that for scenarios with strong physical attributes (high reflection, transparency), context priors constructed via **Strategy A (Visual Anchor)** and **Strategy C (Center Focus)** can effectively compensate for the disadvantage in model parameter size. This makes it possible to achieve SOTA-level detection on the edge with **50% of the computational cost**.

3) Core Finding 3: Perception Enhancement under Complex Textures: On the **Hazelnut** category, our method (80.91%) not only significantly outperforms the baseline but also surpasses LLaVA (64.55%). This is mainly attributed to the amplification effect of **Strategy C (Signal Enhancement)** on low-SNR features, proving that image preprocessing tailored to specific texture attributes is an effective means to enhance the fine-grained perception capability of VLMs.

C. Ablation Study

To deconstruct the independent contributions of the components within the HPF framework, we performed a cumulative ablation experiment. The results are shown in Table III.

TABLE III
ABLATION STUDY ON DIFFERENT STRATEGIES (ACCURACY %)

Config	**M. Nut**	**Trans.**	**Bottle**	**Hazel.**	**Avg**
Baseline	35.65	61.00	45.78	70.00	53.11
+ Strat. A	**80.87**	61.00	45.78	70.00	64.41
+ Strat. B	80.87	**61.00**	45.78	70.00	64.41
+ Strat. C	80.87	61.00	**69.20**	70.00	70.27
+ Ours	80.87	61.00	**75.90**	**80.91**	**74.67**

D. Impact of Prompt Engineering Strategies

To verify the importance of "how to ask," we compared the dynamic prompt strategy of this paper with naive questioning (L1) and descriptive questioning (L2). The results (Table IV) show that the Task-Specific CoT (Chain-of-Thought) prompts designed in this paper guide the model to focus on key areas step-by-step, yielding the best results.

TABLE IV
COMPARISON OF DIFFERENT PROMPTING STRATEGIES (ACCURACY %)

Level	Prompt Strategy Description	Avg Acc
L1	Naive ("Is there a defect?")	53.11
L2	Descriptive ("Describe the image...")	61.25
L3 (Ours)	**Task-Specific CoT + Negative Constraints**	**74.67**

E. Qualitative Results Visualization: "White-Box" Inference Mechanism

To intuitively demonstrate how HPF corrects the model's attention focus, Fig. 5 visualizes the intermediate processing results under different paths.

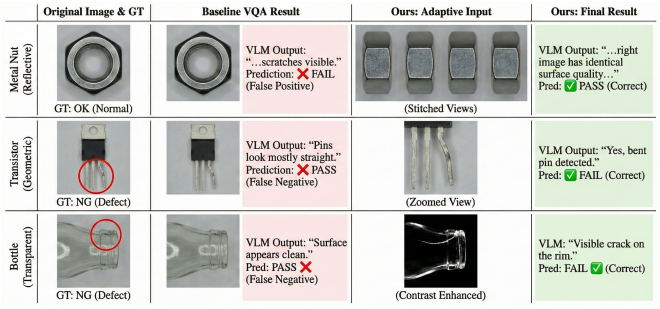

Fig. 5. Visualization of Qualitative Results. Case 1: Strategy A eliminates lighting hallucinations on Metal Nut through differential comparison; Case 2: Strategy B captures micro-deformation of pins via zooming; Case 3: Strategy C highlights subtle glass cracks through signal enhancement.

Case 1 (Metal Nut): The baseline model misidentifies the top reflection as a defect. Strategy A, by introducing the reference image (Left), successfully suppresses this common feature, eventually outputting "Identical surface quality" (PASS).

Case 2 (Transistor): The baseline fails to perceive the slight upward bending of the third pin on the left. The physical zooming strategy (Crop & Zoom) increases the pixel area of the pin, making the deformation explicitly visible to the visual encoder.

Case 3 (Bottle): The low-contrast crack on the bottle mouth is extremely faint in the original image. After CLAHE enhancement and sharpening, the crack's edge gradients are significantly amplified, enabling the model to successfully capture it.

F. Edge Inference Efficiency Analysis

We evaluated the actual inference speed on the RTX 3050 Laptop GPU. The results are detailed in Table V.

To provide a comprehensive comparison, we list the theoretical performance of the original FP16 model alongside our optimized 4-bit implementation. Note that the original FP16 Phi-3.5 model requires ∼7.6 GB VRAM, which exceeds the physical limit of the target edge device (6GB), necessitating the use of our 4-bit quantization strategy. The results show that our method reduces memory usage by nearly 58% while maintaining high accuracy, achieving an average latency of 0.85 seconds per image.

TABLE V
EDGE INFERENCE EFFICIENCY ANALYSIS ON RTX 3050

Model Configuration	Precision	VRAM (GB)	Latency (s/img)	Avg Acc
Baseline (Vanilla Phi-3.5)	FP16	∼7.6*	0.98*	53.11%
Baseline (Vanilla Phi-3.5)	4-bit NF4	∼3.1	0.78	51.50%†
Ours (Phi-3.5 + HPF)	**4-bit NF4**	**∼3.2**	**0.85**	**74.67%**

* FP16 Baseline exceeds VRAM. Values are extrapolated.
† Accuracy drops significantly without perception strategies.
Latency includes preprocessing (CPU) and inference (GPU).

V. DISCUSSION

A. Context Engineering as a New Paradigm for Industrial AI

Our research findings suggest that for highly fragmented industrial zero-shot scenarios, **Context Engineering**—optimizing the input data stream via explicit physical rules—is a superior path compared to blindly pursuing larger foundation models or expensive fine-tuning (e.g., LoRA [20] or QLoRA [21]). It essentially acts as a form of **In-Context Learning (ICL)** guided by hard rules, injecting strong domain supervisory signals without triggering catastrophic forgetting in the model.

B. Physical Perception Boundary of 2D VLMs

Although Strategy B achieved a qualitative breakthrough on Transistors, its accuracy (61%) still lags behind other categories. This reveals a fundamental physical boundary: **monocular 2D images intrinsically lose depth information**. A pin bent directly towards the camera lens appears in the 2D projection merely as a subtle shortening of length, which is extremely difficult for the model to distinguish from perspective distortion caused by fixture errors. This points to the necessity of introducing 3D point cloud multimodal VLMs in the future for complex geometric defect detection.

C. Necessity of Neuro-Symbolic Systems

The ablation study reveals that removing the logic calibration module causes a sharp increase in the False Positive Rate. This phenomenon reflects the intrinsic conflict between deep learning models (Probabilistic) and industrial control systems (Logical). The output of a VLM is essentially a sequence of tokens sampled from a probability distribution, possessing inherent randomness and ambiguity. As illustrated in Fig. 3, the proposed Neuro-Symbolic Logic Calibrator establishes a hybrid loop of "Neural Perception + Symbolic Reasoning." The VLM neural network is responsible for processing unstructured image data to extract high-level semantics (Perception Layer), while the subsequent symbolic rule engine executes strict industrial standards (Logic Layer). This architecture not only enhances the reliability of verdicts but also significantly improves the **Explainability** of the system.

D. Failure Case Analysis

We identify two main failure modes in current experiments: (1) **Ambiguous Semantic Boundaries:** Signal enhancement, while highlighting defects, may also amplify benign features like water stains or oil, leading to false positives by the VLM. (2) **Severe Occlusion:** If key areas of the object are severely occluded, it may disrupt the statistical feature calculation in the routing module, leading to strategy misallocation.

VI. Conclusion

This paper proposes a Category-Aware Hybrid Perception Framework (HPF) designed to overcome the significant domain gap in applying general VLMs to industrial zero-shot anomaly detection. By cognitively decoupling the complex detection task and dynamically scheduling orthogonal perception strategies based on physical attributes (Reflection, Geometry, Texture), we achieved SOTA performance on multiple challenging categories of the MVTec AD dataset using a lightweight edge-compatible model. This work not only provides a scalable, low-cost technical paradigm for deploying VLM technology on the industrial edge but also offers deep insights into the physical boundaries of data-driven AI systems in the real world.

Acknowledgment

This research was partially supported by grants from the Key Research of the Central Guidance Fund for Local Science and Technology Development (No. 2023ZYDF078).

References

[1] Y. Cui *et al.*, "A survey on unsupervised anomaly detection algorithms for industrial images," *IEEE Access*, vol. 10, pp. 69658–69678, 2022.

[2] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, 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.

[3] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 475–489.

[4] V. Zavrtanik, M. Kristan, and D. Skočaj, "DRAEM-A discriminatively trained reconstruction embedding for surface anomaly detection," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8330–8339.

[5] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "FastFlow: Unsupervised anomaly detection via 2d normalizing flows," *arXiv preprint arXiv:2111.07677*, 2021.

[6] G. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "Unified multi-class anomaly detection with distance-aware transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7943–7952.

[7] H. Wang, P. Wang, L. Song, and J. Yan, "SimpleNet: A simple network for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 20351–20361.

[8] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," *arXiv preprint arXiv:2005.02357*, 2020.

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

[10] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 12888–12900.

[11] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 23716–23736.

[12] A. Kirillov *et al.*, "Segment anything," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.

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

[15] M. Abdin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[16] 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.

[17] Y. Liu *et al.*, "AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[18] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "CoOp: Learning to prompt for vision-language models," *Int. J. Comput. Vis. (IJCV)*, vol. 130, no. 9, pp. 2337–2348, 2022.

[19] 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.

[20] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.