DefectGPT: Towards Multi-Class Defect Detection with Limited Electrical Samples

Zhuoyi Lin[1]
Zhuoyi.Lin@outlook.com
Kaixin Xu[1]
Xu_Kaixin@i2r.a-star.edu.sg
Aye Phyu Phyu Aung[1]
Aye_Phyu_Phyu_Aung@i2r.a-star.edu.sg
Wen Qiu[2]
Wen.Qiu@amd.com
Bernice Zee[2]
Bernice.Zee@amd.com
Jiann Min Chin[2]
JM.Chin@amd.com
J. Senthilnath[1,*]
J_Senthilnath@i2r.a-star.edu.sg

[1] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
[2] Device Analysis Lab, Advanced Micro Devices (Singapore) Pte Ltd, Singapore

Abstract

Existing deep learning (DL) approaches for electrical defect detection typically require large quantities of labeled data. However, in real-world scenarios, defect samples are significantly rarer than normal samples, making data collection both challenging and resource-intensive. Meanwhile, the wide variability in anomaly types and the unpredictable nature of defect locations further complicate the labeling process. In this paper, we present DefectGPT, a novel approach for multi-class electrical defect detection with limited samples by leveraging multimodal large language models (MLLMs). Specifically, DefectGPT generates defect descriptions to establish a rich semantic context for subsequent reasoning and classification. To effectively utilize limited training samples and enhance classification accuracy, we introduce a novel in-context learning technique, termed hypothesis-first learning (HFL), which facilitates the DefectGPT to generate an initial hypothesis before refining its knowledge. This approach enables DefectGPT to improve defect classification progressively through self-reflection, enhancing its effectiveness and generalization with limited electrical samples. We further conduct a theoretical analysis to gain deeper insights into the underlying learning mechanism of DefectGPT. Experimental evaluations on two distinct electrical datasets demonstrate the effectiveness and generalizability of our approach, highlighting its potential to reduce data dependency while maintaining high classification accuracy.

*Corresponding author

1 Introduction

The semiconductor industry serves as a foundational pillar of modern technology, driving advancements across diverse domains including computing, telecommunications, healthcare, and automotive systems. Economically, it represents a critical global sector, with annual revenues exceeding $500 billion in 2024 globally, underscoring its central role in the digital economy [1]. As semiconductor devices grow in complexity with emerging packaging methods including flip chip, stacked die, wafer-level packaging, and through-silicon vias, ensuring their reliability becomes increasingly critical. Effective defect detection is fundamental to maintaining high yield, consistent performance, and long-term operational stability. The rise of buried and non-visual defects within multilayer structures presents significant challenges, requiring advanced and nondestructive detection techniques for accurate classification. These advancements are crucial for efficient failure analysis and minimizing production disruptions, emphasizing the need for ongoing innovation in defect detection methodologies to keep pace with industry advancements.

In recent years, deep learning (DL) has proven highly effective in domains like computer vision, natural language processing, and recommender systems, demonstrating its ability to capture underlying patterns and representations from data [9, 12, 21, 26]. However, as illustrated in Figure 1, a key limitation of DL in electrical defect detection is its reliance on large, well-labeled datasets. Labeling microscopic or non-visible defects requires expertise and advanced imaging, making the process time-consuming and costly. The imbalance between defective and non-defective samples further impacts model performance. While data augmentation and synthetic labeling offer partial solutions, they do not fully address the challenges of large-scale manual annotation or generalization to new manufacturing conditions. Inspired by recent advances in multimodal language models (MLLMs) such as CLIP [29], which demonstrate impressive capabilities in visual understanding and language reasoning, there is growing potential to leverage MLLMs for more data-efficient defect detection in the semiconductor manufacturing process. Specifically, MLLMs offer a promising approach to analyzing images while incorporating contextual information, thus reducing dependence on large labeled datasets. By integrating visual features with descriptive annotations, MLLMs could enhance defect classification and localization even with limited labeled data, addressing key limitations of DL in electrical defect detection. Additionally, MLLMs can adapt to new defect types or variations with limited additional data, significantly reducing the time and costs associated with manual labeling [7, 11, 18, 32]. Such generalization capability positions MLLMs as a valuable solution for improving failure analysis in increasingly complex semiconductor technologies.

In this work, we present an early exploration into multi-class defect classification using an MLLM named GPT-4o [14], applied to electrical defect images. We propose a novel approach called DefectGPT, which aligns the GPT-4o model with electrical images and textual defect descriptions, even in scenarios with extremely limited labeled data, demonstrating its potential for data-efficient defect detection. Specifically, we first prompt GPT-4o to generate detailed defect descriptions for various defect types. We then employ in-context learning to align GPT-4o with the specific characteristics of defects, enabling GPT-4o to iteratively update its knowledge base. Specifically, we propose hypothesis-first learning (HFL), which explicitly structures the classification process into a two-stage reasoning workflow: (1) initial hypothesis generation and (2) contextual self-reflection. This structure encourages the model to articulate and justify its prediction before refining its decision based on feedback. By consolidating human-like reasoning and self-reflection, HFL enables the DefectGPT to

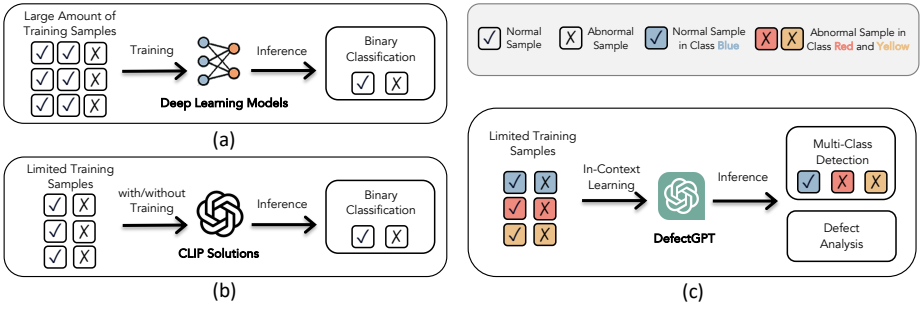

Figure 1: Different paradigms for anomaly detection. (a) Deep learning models (e.g., DRA [9] and DevNet [26]) typically require large training datasets for binary anomaly detection. (b) CLIP solutions reduce this need: AnomalyCLIP [33] requires fewer training samples, while WinCLIP [16] enables training-free anomaly detection. (c) DefectGPT uses in-context learning for multi-class detection and defect analysis with minimal samples.

better capture the defect characteristics and lead to more accurate classification performance, particularly in few-shot settings where data samples are very limited. To demonstrate the generalization capability of DefectGPT, we evaluated it on two distinct datasets: Magnetic Field Imaging (MFI) and WaferMap. Our evaluation across both datasets highlights DefectGPT's effectiveness and generalization, achieving 18.18% and 37.58% improvement against comparing methods in classification accuracy on MFI and WaferMap datasets, respectively. The contributions of this paper are summarized as follows:

- We provide an early attempt to explore multi-class defect classification with limited electrical samples using MLLMs. To this end, we propose DefectGPT, which first generates defect descriptions and then leverages HFL to iteratively refine its understanding of visual defect characteristics.
- To the best of our knowledge, we are the first to perform defect detection on MFI using machine learning methods, which is much faster than traditional X-Ray scanning.
- We provide a theoretical analysis to understand the behavior of DefectGPT and conduct extensive experiments on two distinct datasets to demonstrate its effectiveness.

2 Related Work

2.1 MLLM for Industrial Anomaly Detection

Recent work has highlighted the need for generalization in industrial anomaly detection, leading to zero-shot and few-shot models based on vision-language frameworks like CLIP [29]. For example, WinCLIP [16] is proposed to address binary zero-shot and few-shot anomaly classification and segmentation by extracting and aggregating multi-scale image features. Subsequently, CLIP has been further investigated for detecting building defect images [31]. In addition, AnomalyGPT [11] and AnomalyCLIP [33] propose to fine-tune model parameters for zero-shot anomaly detection and few-shot anomaly detection, respectively. Recent advancements in MLLMs (e.g., GPT4V [3] and Phi3.5 [2]) have enabled the handling of various tasks through the Visual Question Answering (VQA) paradigm without

finetuning. As pioneering efforts, comprehensive experiments are conducted using GPT-4V(ision), evaluating its anomaly detection capabilities under zero-shot and one-shot settings. The evaluations frequently rely on a reference normal image and are performed across a diverse range of datasets and tasks [7, 32]. While most anomaly detection methods operate within a binary classification framework (i.e., determining whether an image is normal or defective), a recent benchmark, MMAD [19], reformulates the industrial anomaly detection task as a multiple-choice VQA problem to evaluate the performance of various MLLMs. Although this approach demonstrates the effectiveness of models such as GPT-4o in defect detection, it is inherently limited to four-class classification, thereby constraining its applicability to more fine-grained or complex scenarios. Despite these promising results, how to further improve the multi-class defect detection performance of MLLMs remains an open and challenging research problem.

2.2 Multi-Modal In-Context Learning

In-context learning (ICL), initially introduced in large language models (LLMs) like GPT-3 [6], enables models to perform new tasks from a few examples without parameter updates, and has recently been extended to MLLMs [10]. Foundational models such as CLIP [29] and ALIGN [17] built robust multimodal representations but were not inherently designed for ICL. More recent approaches like Flamingo [4] and BLIP-2 [20] integrate vision encoders with language models to support flexible, few-shot reasoning across tasks such as visual question answering and image captioning. While ICL shows promise for flexible task adaptation, it remains unclear how in-context learning can effectively enhance electrical defect detection where visual nuances are subtle, task-specific supervision is limited, and robustness to unseen variations is critical. In this work, we bridge this gap by investigating the effectiveness of in-context learning for multi-class electrical defect detection within a VQA framework. Specifically, we introduce a novel multimodal in-context learning approach named HFL, specifically developed to address the challenges of multi-class defect classification with limited samples.

3 Methodology

3.1 Overall Framework

As shown in Figure 2, the proposed DefectGPT consists of two main stages: (1) generating defect descriptions for each defect class and (2) perform HFL based on the generated defect description and limited training samples.

3.2 Defect Description Generation

Traditional DL approaches for defect detection typically require extensive labeled datasets to learn latent representations of normal and abnormal samples [7, 16, 32, 33]. However, collecting such datasets is often costly and time-consuming, particularly within the context of semiconductor manufacturing. Inspired by CuPL [28], which employs LLMs to generate class-specific prompts for image categories and demonstrates improved zero-shot classification performance, we propose a defect description generation strategy tailored to the domain

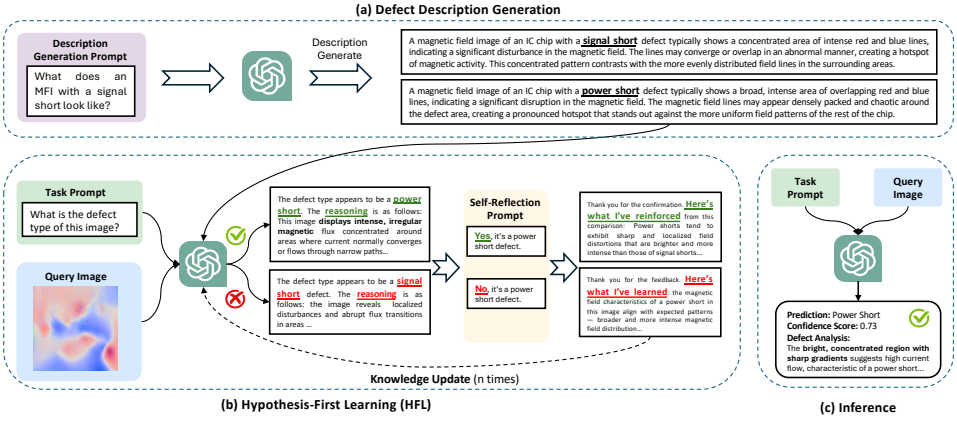

Figure 2: An overview of the DefectGPT. Note that n indicates the number of examples during the HFL process.

of electrical defect detection. As shown in Figure 2(a) and Figure 3, we use GPT-4o to generate defect descriptions for the MFI and WaferMaps datasets, respectively. The generated defect descriptions provide rich semantic context that enhances DefectGPT's ability to distinguish between closely related defect classes, reducing the reliance on extensive task-specific supervision. These prompts are tailored to specific categories and incorporate detailed visual features, embedding class-specific attributes that improve the model's discriminatory capacity. In the following subsection, we present how HFL further leverages the generated defect description for more accurate defect detection.

3.3 Hypothesis-First Learning

Electrical defect detection presents several critical challenges that limit the effectiveness of few-shot learning approaches [11, 16, 18, 33]. For example, defects in MFIs often exhibit subtle visual variations that are difficult to capture with simple class labels or fixed prompts. Moreover, many defect types, including signal shorts and power shorts, exhibit overlapping visual characteristics, which makes them difficult to distinguish without domain-specific reasoning. Standard few-shot classification methods, which rely on direct comparisons with limited demonstrations, often lack the nuanced reasoning capabilities necessary to differentiate fine-grained categories [11, 16].

In this work, we propose HFL, a novel ICL technique to enhance few-shot defect classification using MLLMs. Unlike existing approaches that produce direct classification outputs [2, 18, 32], HFL guides the model to first generate an explicit hypothesis about the defect type, supported by natural language reasoning grounded in both visual evidence and textual descriptions. As shown in Figure 2(b), given a query image, defect descriptions, and a task prompt, DefectGPT initially synthesizes a hypothesis by interpreting the visual input in the context of class-specific descriptions. This hypothesis includes both a predicted class label and an accompanying rationale. Subsequently, the hypothesis undergoes contextual validation through a feedback loop: the label of the query image is introduced into the self-reflection prompt. The model then compares its initial hypothesis with these examples, triggering a self-reflection process. In this process, DefectGPT articulates what it has reinforced or learned from the comparison, enabling the model to update its understanding

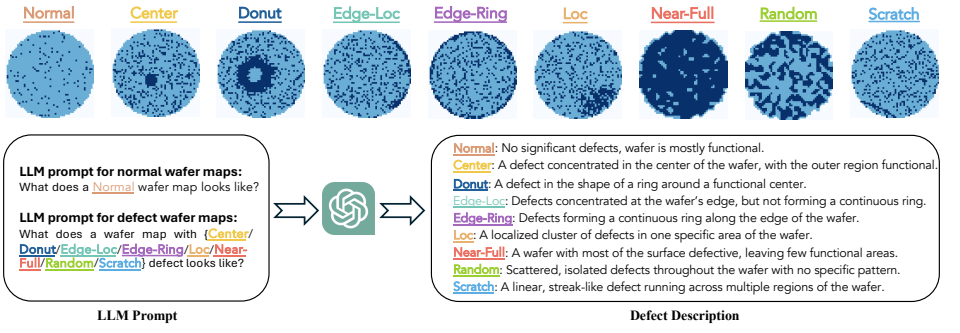

Figure 3: The process of defect description generation on WaferMap dataset.

of key discriminative visual characteristics. This reflective mechanism allows the model to progressively refine both its knowledge of defect features and its reasoning strategy, even in the absence of gradient-based updates. Through a two-stage reasoning process, which involves initial hypothesis generation followed by reflective refinement, HFL enhances the model's ability to distinguish subtle and context-dependent defect types. We investigate how the refined knowledge of defect characteristics benefits MLLM performance in Section 4.5 and Section 4.7.

3.4 Hypothesis-First Learning as Implicit Parameter Adaptation

We begin with the result of Irie et al., who showed that one gradient-descent step on a linear layer yields a *linear-attention kernel* [8, 15]:

$$\mathcal{F}(x) = \big(W_0 + \Delta W\big)x = W_0 x + \sum_i (e_i \otimes x'_i)\, x = W_0 x + \sum_i e_i \left(x'^{\top}_i x\right) = W_0 x + \mathrm{L\_Attn}(E, X', x) \tag{1}$$

where $\mathrm{L\_Attn}(V, K, \mathbf{q})$ represents the linear-attention operation, treating past output-error signals E as the values, earlier inputs X' as the keys, and the current input x as the query. In the context of this work, we let the current token representation be $\mathbf{x} \in \mathbb{R}^d$ and its query vector $\mathbf{q} = W_Q \mathbf{x} \in \mathbb{R}^{d'}$. For a single attention head, the output is formulated as:

$$\mathcal{F}_{\mathrm{HFL}}(\mathbf{q}) = \mathrm{Attn}(V, K, \mathbf{q}) = W_V [D; X'; R]\ \mathrm{Softmax}\left(\frac{(W_K [D; X'; R])^{\top} \mathbf{q}}{\sqrt{d}}\right) \tag{2}$$

where D are class-description tokens, X' are demonstration tokens, R are self-reflection tokens, and $W_Q, W_K, W_V \in \mathbb{R}^{d' \times d}$ are the projection matrices.

Assumption 1 (Linear-attention surrogate)**.** *Throughout the analysis, we approximate dot-product attention by linear attention for simplicity, dropping the Softmax and scale factor:*

$$\tilde{\mathcal{F}}_{\mathrm{HFL}}(\mathbf{q}) = W_V [D; X'; R] \big(W_K [D; X'; R]\big)^{\top} \mathbf{q} \tag{3}$$

Define the *few-shot kernel* $W_{\mathrm{FSL}} = W_V X' (W_K X')^{\top}$, so that $W_{\mathrm{FSL}} \mathbf{q}$ is the linear-attention output in pure few-shot setting.

Proposition 1 (Equivalence to parameter update)**.** *Under Assumption 1, $\tilde{\mathcal{F}}_{\mathrm{HFL}}(\mathbf{q})$ equals a single linear map applied to $\mathbf{q}$:*

$$\tilde{\mathcal{F}}_{\mathrm{HFL}}(\mathbf{q}) = \big(W_{\mathrm{FSL}} + \Delta W_{\mathrm{HFL}}\big)\mathbf{q} \tag{4}$$

Table 1: Performance comparison on two electrical datasets. The best results are highlighted in bold, with the second-best results underlined. Relative improvements of DefectGPT over the second-best methods are also reported for reference.

Methods	MFI				WaferMap			
	Top-1 Acc.	AUPR	Macro-F1	Micro-F1	Top-1 Acc.	AUPR	Macro-F1	Micro-F1
CLIP	50.00	54.43	33.33	50.00	30.33	<u>73.05</u>	19.19	30.33
CuPL	60.00	60.93	52.38	60.00	30.22	71.35	18.75	30.22
LLaVA	52.50	53.37	42.03	52.50	12.90	50.25	8.16	12.90
Phi-3.5 (ZSL)	45.00	48.50	37.32	45.00	11.11	51.86	2.22	11.11
Phi-3.5 (FSL)	60.00	52.19	52.38	60.00	11.11	50.00	2.22	11.11
GPT-4o (ZSL)	61.67	52.00	61.67	59.11	20.70	50.84	12.14	20.70
GPT-4o (FSL)	<u>82.50</u>	<u>88.87</u>	<u>82.21</u>	<u>82.50</u>	<u>48.78</u>	66.75	<u>48.33</u>	<u>48.78</u>
DefectGPT	**97.50**	**95.67**	**97.49**	**97.50**	**67.11**	**92.16**	**62.52**	**67.11**
Improvement	18.18%	7.65%	18.59%	18.18%	37.58%	38.07%	29.36%	37.58%

where the update matrix is:

$$\Delta W_{\mathrm{HFL}} = \sum_i (W_V \mathbf{d}_i) \otimes (W_K \mathbf{d}_i) + \sum_i (W_V \mathbf{r}_i) \otimes (W_K \mathbf{r}_i) \tag{5}$$

with $\{\mathbf{d}_i\}$ *the description tokens and* $\{\mathbf{r}_i\}$ *the self-reflection tokens.*

Proof. From Assumption 1, we derive from Eq. (3) and obtain:

$$\tilde{\mathcal{F}}_{\mathrm{HFL}}(\mathbf{q}) = \sum_{Z \in \{D, X', R\}} W_V Z (W_K Z)^\top \mathbf{q}$$

Note that the term with $Z = X'$ is exactly $W_{\mathrm{FSL}}\mathbf{q}$. For $Z \in \{D, R\}$, we have:

$$W_V Z (W_K Z)^\top \mathbf{q} = \sum_i (W_V \mathbf{z}_i)\big((W_K \mathbf{z}_i)^\top \mathbf{q}\big) = \Big(\sum_i (W_V \mathbf{z}_i) \otimes (W_K \mathbf{z}_i)\Big)\mathbf{q}$$

Combining all components gives the stated result:

$$\tilde{\mathcal{F}}_{\mathrm{HFL}}(\mathbf{q}) = \Big(W_{\mathrm{FSL}} + \sum_i (W_V \mathbf{d}_i) \otimes (W_K \mathbf{d}_i) + \sum_i (W_V \mathbf{r}_i) \otimes (W_K \mathbf{r}_i)\Big)\mathbf{q} \qquad \square$$

Interpretation. Proposition 1 shows that, under the linear-attention surrogate, attending to the defect descriptions tokens D and self-reflections tokens R is algebraically equivalent to updating the few-shot weight matrix W_{FSL} by the additive term ΔW_{HFL}. In other words, HFL performs an *implicit, on-the-fly parameter adaptation* before the query is processed, mirroring gradient steps but executed entirely within the attention mechanism.

4 Experimental Results

4.1 Electrical Dataset

To demonstrate the effectiveness and generalization ability of the proposed DefectGPT in learning defect patterns across various defect detection scenarios, we conducted experiments and analysis using two electrical datasets: (1) **Magnetic Field Image (MFI)**: The MFI samples are synthesized using a physics-informed diffusion model, PI-GenMFI [5], which is

specifically designed for MFI and trained on limited experimental semiconductor data. Unlike conventional 3D imaging methods such as 3D X-ray scanners [25], this approach provides a non-destructive means of visualizing electrical defects in three dimensions, while substantially reducing scan time. For more details, please refer to [5]; (2) **MixedWM38 Dataset (WaferMap)**: a multi-class dataset comprising defect patterns collected from real-world semiconductor fabrication processes [30]. Since this work focuses on single-class classification, we test our DefectGPT and baseline methods on one normal pattern and eight single-defect patterns (i.e., center, donut, loc, edgeloc, edgering, near-full, scratch, and random), representing various manufacturing anomalies. For the training dataset, we randomly select 5 samples per class, yielding a total of 45 images. To maintain computational efficiency, the test set is constructed by randomly selecting another 100 samples from each class, resulting in 900 test images in total.

4.2 Evaluation Protocols

Training Details: To ensure consistency in token length across classes and to facilitate evaluation of the proposed HFL, we construct subsets under 1-shot, 2-shot, 3-shot, and 5-shot settings and report the best results. In each K-shot configuration, the context comprises K examples per label, thereby guaranteeing representation of all unique classes.

Evaluation Metrics: We use 4 metrics to provide a comprehensive understanding of the model's performance. (1) **Top-1 accuracy** reports the proportion of samples for which the model's highest-confidence prediction is correct [13]. (2) **Area under the precision–recall curve (AUPR)** quantifies the precision–recall trade-off across decision thresholds, a standard measure in anomaly detection [16, 34]. (3) **Macro-F1** averages the F1 scores computed independently for each class, giving equal weight to all classes, while (4) **Micro-F1** aggregates true positives, false positives, and false negatives over all classes before calculating the F1 score [23, 24]. The formal equations are given in Supplementary Material A.

4.3 Baseline Methods

We benchmark our method against five advanced MLLMs. (1) **CLIP** [29]: A popular MLLM that learns visual concepts from natural language supervision. (2) **CuPL** [27]: It enhances the CLIP model with customized prompts for each image category. (3) **LLaVA** [22]: It enables strong zero-shot capabilities in visual reasoning tasks (4) **Phi-3.5** [2]: a state-of-the-art MLLM by Microsoft, designed for high-quality vision-language reasoning. (5) **GPT-4o** [14]: is OpenAI's advanced multimodal model, demonstrating the best performance in defect detection tasks in a recent benchmarking paper [18]. Note that Phi-3.5 and GPT-4o support both zero-shot learning (ZSL) and few-shot learning (FSL) settings, while other baselines only support the ZSL setting due to the limit of the context window. The prompt formats are presented in Supplementary Material B.

4.4 Performance Comparison

Table 1 presents a comprehensive evaluation of DefectGPT against representative baselines on the MFI and WaferMap datasets. Across all metrics and datasets, DefectGPT consistently achieves the best performance, demonstrating its effectiveness for multi-class electrical defect detection. In particular, DefectGPT attains 97.50% Top-1 accuracy on the MFI dataset and 67.11% on the WaferMap dataset, yielding relative improvements of 18.18% and 37.58%

Table 2: The effect of the number of demonstrations per class on DefectGPT.

#Demonstrations	MFI				WaferMap			
	Top-1 Acc.	AUPR	Macro-F1	Micro-F1	Top-1 Acc.	AUPR	Macro-F1	Micro-F1
1-shot	83.33	82.89	83.33	79.45	35.44	74.92	31.65	32.65
2-shot	**97.50**	95.67	**97.49**	**97.50**	**67.11**	**92.16**	**62.52**	**67.11**
3-shot	83.33	98.96	82.80	83.33	56.33	74.22	57.42	56.33
4-shot	85.00	**99.00**	84.65	85.00	57.27	77.76	55.29	58.23
5-shot	92.50	96.52	92.49	92.50	59.11	87.61	49.64	59.11

over the strongest baseline GPT-4o (FSL). The consistent performance gap among GPT-4o (ZSL), GPT-4o (FSL), and DefectGPT indicates that while examples provide useful guidance for ICL, they are insufficient on their own. The integration of the defect descriptions with the proposed HFL proves essential for capturing nuanced defect patterns and significantly enhances the accuracy of electrical defect detection. In addition, CLIP, CuPL, LLaVA, and Phi 3.5 perform adequately on the MFI dataset but show substantial performance degradation on the more complex multi-class WaferMap dataset, showing their limited ability to effectively capture subtle variations across multiple defect types.

4.5 Ablation Study

We conduct an ablation study to evaluate the contribution of key components in DefectGPT, particularly focusing on the impact of defect description quality and the HFL mechanism. As shown in Figure 4, we begin with baseline GPT-4o models under ZSL and FSL settings. When directly augmented with the initial generated defect descriptions (D), the resulting GPT-4o (ZSL+D) and GPT-4o (FSL+D) variants exhibit degraded performance compared to their non-augmented counterparts, suggesting that raw descriptive prompts may introduce noise or ambiguity. Importantly, GPT-4o (FSL+D) is equivalent to removing the HFL technique from DefectGPT, as HFL is able to refine the initial defect descriptions into more accurate and discriminative forms (D') through self-reflection. When we introduce these refined descriptions via GPT-4o (FSL+D'), performance improves significantly. The model achieves 78.00% Top-1 Accuracy and 89.94 AUPR, results that are comparable to those of the full DefectGPT. This demonstrates the effectiveness of HFL in enhancing the utility of defect descriptions, providing rich textual supervision for achieving more accurate performance in multi-class defect detection.

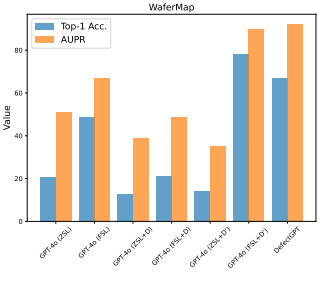

Figure 4: Ablation study.

4.6 Impact of Supervision Scale

Table 2 presents the impact of varying the number of demonstrations on DefectGPT's performance across the MFI and WaferMap datasets. Performance improves markedly from 1-shot to 2-shot supervision, with 2-shot achieving the highest Macro-F1 and Micro-F1 scores on both MFI (97.49 and 97.50) and WaferMap (62.52 and 67.11). Although increasing to 3–5 shots yields improvements in some metrics (e.g., MFI AUPR peaks at 99.00 with 4-shot), the gains are inconsistent and occasionally decline. These findings suggest that a small number of well-selected demonstrations is sufficient for optimal performance, while additional supervision offers limited benefit.

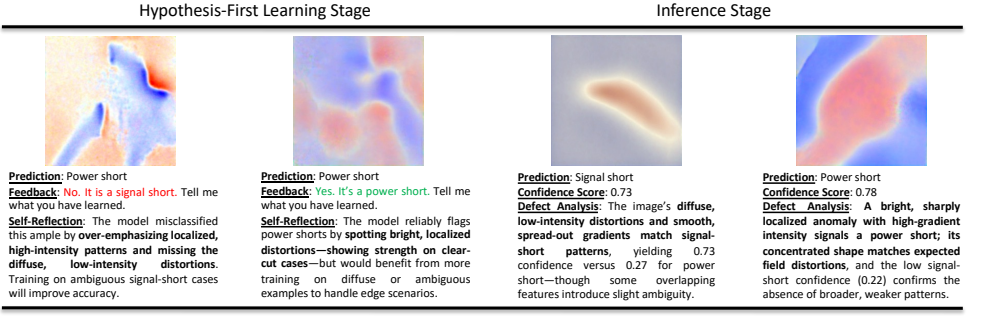

Figure 5: Illustrative outputs of DefectGPT for MFI samples.

4.7 Model Reasoning Across Learning and Inference Stages

Figure 5 illustrates DefectGPT's reasoning process across the HFL and inference stages using MFI samples. During the learning stage, the model engages in self-reflection based on external feedback, analyzing misclassifications, recognizing over-reliance on localized features, and suggesting targeted refinements. In the inference stage, DefectGPT performs defect analysis by distinguishing between power and signal shorts using spatial patterns and intensity gradients, accompanied by confidence scores for each prediction.

5 Conclusion and Future Work

In this paper, we propose DefectGPT, a novel MLLM method for multi-class defect detection with limited electrical imaging data. Leveraging generated defect descriptions in conjunction with the HFL technique, DefectGPT is capable of iteratively refining its understanding of defect characteristics, enabling more accurate multi-class electrical defect detection. We provide a theoretical analysis of the DefectGPT and empirically validate its effectiveness on two electrical datasets. While this study focuses on single-label classification, future work will explore the extension of DefectGPT to multi-label anomaly detection scenarios.

6 Acknowledgment:

This study is supported by the Machine Learning Guided Failure Analysis & Diagnostic Capability Development for Next Generation 3D-IC Packaging at A*STAR via the IAF-PP by the Agency for Science, Technology and Research under Grant No. M23K8a0050.

References

[1] 2024 state of the u.s. semiconductor industry. *https://www.semiconductors.org/*, 2024.

[2] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[5] Aye Phyu Phyu Aung, Lucas Lum, Zhansen Shi, Wen Qiu, Bernice Zee, JM Chin, Yeow Kheng Lim, and J.Senthilnath. Physics informed generative models for magnetic field images. *arXiv preprint arXiv:2508.20612*, 2025.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[7] Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards Generic Anomaly Detection and Understanding: Large-scale Visual-linguistic Model (GPT-4v) Takes the Lead. *arXiv.org*, abs/2311.02782, 2023.

[8] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

[9] Choubo Ding, Guansong Pang, and Chunhua Shen. Catching Both Gray and Black Swans: Open-set Supervised Anomaly Detection. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7378–7388, 2022.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A Survey on In-context Learning. In *Conference on Empirical Methods in Natural Language Processing*, 2022.

[11] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting Industrial Anomalies Using Large Vision-Language Models. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1932–1940, 2024.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 558–567, 2019.

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[15] Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. In *International Conference on Machine Learning*, pages 9639–9659. PMLR, 2022.

[16] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/Few-Shot Anomaly Classification and Segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 19606–19616, 2023.

[17] Chao Jia, Yinfei Yang, Ye Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. *ICML*, 2021.

[18] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: The First-Ever Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. *arXiv*, 2024.

[19] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. In *The Thirteenth International Conference on Learning Representations*, volume abs/2410.09453, 2025.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[21] Zhuoyi Lin, Sheng Zang, Rundong Wang, Zhu Sun, J Senthilnath, Chi Xu, and Chee Keong Kwoh. Attention over self-attention: Intention-aware re-ranking with dynamic transformer encoders for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):7782–7795, 2022.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[23] Yunbo Long, Zhengyang Ling, Sam Brook, Duncan McFarlane, and Alexandra Brintrup. Leveraging unsupervised learning for cost-effective visual anomaly detection. *IET Conference Proceedings*, 2024(11):95–101, 2024.

[24] Mingrui Ma, Lansheng Han, and Chunjie Zhou. Research and application of transformer based anomaly detection model: A literature review. *arXiv preprint arXiv:2402.08975*, 2024.

[25] ZEISS Microscopy. Non-destructive failure analysis with zeiss x-ray microscopy, 2023. URL https://www.zeiss.co.jp/content/dam/rms/countries/singapore/apac/electronics-ebook-nondestructivefa.pdf.

[26] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep Anomaly Detection with Deviation Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 353–362. ACM, 2019.

[27] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023.

[28] Sarah M. Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 15645–15655, 2023.

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[30] Junliang Wang, Chuqiao Xu, Zhengliang Yang, Jie Zhang, and Xiaoou Li. Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition. *IEEE Transactions on Semiconductor Manufacturing*, 33(4):587–596, 2020.

[31] Gunwoo Yong, Kahyun Jeon, Daeyoung Gil, and Ghang Lee. Prompt engineering for zeroshot and fewshot defect detection and classification using a visuallanguage pretrained model. *Computer-Aided Civil and Infrastructure Engineering*, 38(11):1536–1554, 2023.

[32] Jiangning Zhang, Haoyang He, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. *GPT-4V-AD: Exploring Grounding Potential of VQA-Oriented GPT-4V for Zero-Shot Anomaly Detection*, volume abs/2311.02612. Springer Nature Singapore, 2025.

[33] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection. In *International Conference on Learning Representations (ICLR)*, 2024.

[34] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022.