INSIGHT-X: Multimodal VLM Assistant for Artifact Localization, and Mitigation Guidance in X-ray Inspection

Antika Roy, Pratyush Shukla, Junfu Cheng, Shajib Ghosh, Patrick Craig, Akash Balaji, Maitrey Vivek Phatak, Hafez Bahrami and Navid Asadizanzani
Dept. of Electrical and Computer Engineering, University of Florida,
5000 Malachowsky Hall, 1889 Museum Road
Gainesville, FL 32611 USA
Email: antika.roy@ufl.edu, hafezbahrami@ufl.edu, nasadi@ece.ufl.edu

Abstract
Advanced semiconductor packages rely on X-ray CT and laminography for non-destructive inspection of buried interconnects and internal structures. However, dense metal features and heterogeneous materials often introduce artifacts, including ring artifacts, beam-hardening streaks, geometric distortions, and missing-cone artifacts. These artifacts obscure critical features during inspection, reduce reconstruction fidelity, and complicate defect interpretation, requiring expert-driven and iterative troubleshooting. To address these challenges, we propose INSIGHT-X, a vision–language model framework for artifact diagnosis, localization, root-cause and mitigation in X-ray inspection workflows. INSIGHT-X formulates artifact analysis as a multimodal question-answering task by integrating reconstructed X-ray images, artifact annotations, and expert knowledge into an artifact-centric dataset. Through parameter-efficient fine-tuning, the model learns to identify artifact types, localize affected regions, and generate interpretable explanations. The system also provides artifact-aware cause and mitigation recommendations, including general acquisition and reconstruction adjustments, and calibration guidance. By combining visual understanding, spatial localization with explainable answers, INSIGHT-X transforms expert knowledge into a scalable AI assistant for improving inspection consistency, accelerating troubleshooting, and supporting decision-making in advanced semiconductor manufacturing.

Key words
Artifact-centric dataset, CT, Laminography, Parameter-efficient fine-tuning, Vision-language model.

I. Introduction
Non-destructive inspection (NDI) using X-ray imaging has become a cornerstone of modern semiconductor manufacturing, particularly as device architectures evolve toward highly integrated 2.5D and 3D packaging technologies. Advanced packages, such as chiplet-based architectures, stacked dies, high-bandwidth memory (HBM), and heterogeneous integration platforms contain dense interconnect networks, multilayer substrates, and deeply embedded structures that cannot be accessed using conventional optical inspection methods [1,2]. X-ray computed tomography (CT) and X-ray laminography therefore play a critical role in enabling volumetric visualization of internal features, allowing engineers to inspect solder joints, micro-bumps, through-silicon vias (TSVs), redistribution layers (RDLs), and buried interconnects without damaging the device. Reliable X-ray inspection is essential for ensuring manufacturing yield, reliability, and supply-chain assurance in advanced semiconductor systems [3].

However, the very structures that make modern packages dense also exacerbate X-ray artifacts due to the complex interaction between X-ray radiation and heterogeneous device materials. For example, uneven X-ray penetration in flat samples produces cupping and streak artifacts that distort voxel values and obscure small features [4,5]. Modern packages contain high-density metal structures such as copper pillars, solder bumps, and TSV arrays embedded within multilayer dielectric and substrate materials [2]. These dense structures produce severe attenuation variations that generate reconstruction artifacts, including ring artifacts, beam-hardening streaks, geometric distortions, and missing-cone artifacts in laminography resulting from incomplete angular sampling [6]. Collectively, these artifacts obscure fine structural details, mimic defect signatures, or reduce effective spatial resolution, making it difficult to reliably distinguish true defects from imaging artifacts [5,6]. In

practice, diagnosing the source of X-ray artifacts and determining appropriate mitigation strategies requires substantial domain expertise in X-ray physics, imaging geometry, and reconstruction algorithms. Inspection engineers often rely on iterative adjustments of acquisition parameters, reconstruction settings, and calibration procedures to improve image quality which is a time-consuming process that hinders throughput and consistency.

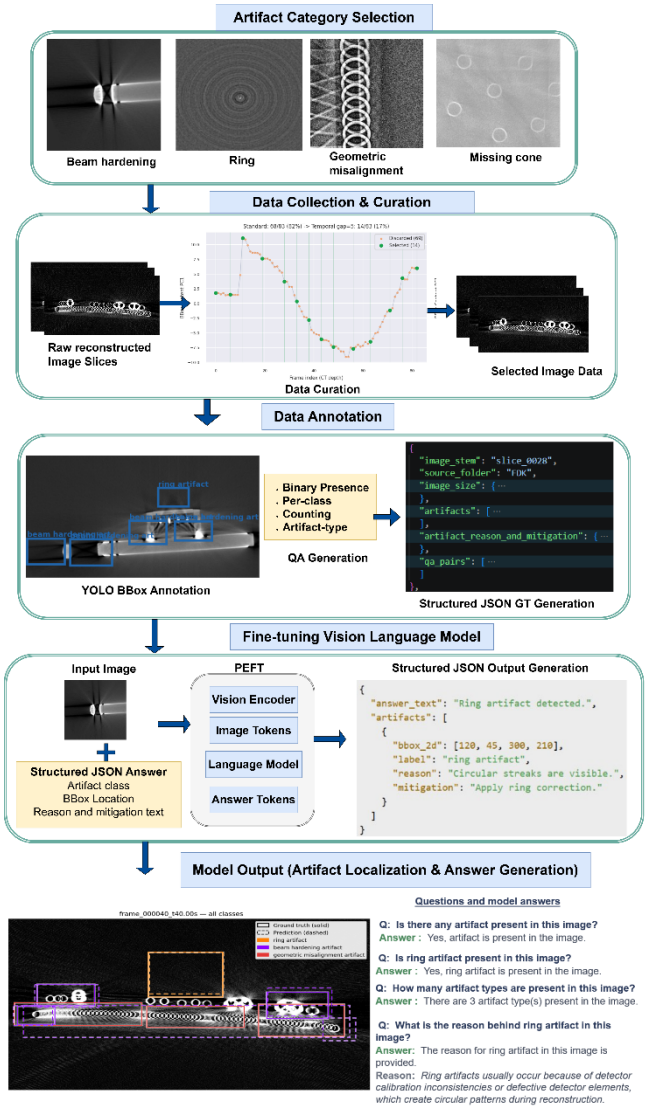


Figure 1: Overview of our proposed INSIGHT-X framework. The framework integrates artifact selection, data curation, annotation, structured JSON/QA generation, and vision-language model fine-tuning to perform automated artifact localization, reasoning and question answering from X-ray reconstruction images.

At the same time, the rapid advancement of large language models (LLMs) and multimodal AI assistants has demonstrated the potential for AI systems to support complex technical workflows by integrating reasoning, knowledge retrieval, and decision support. These developments have led to the emergence of intelligent AI assistants capable of interacting with domain-specific data and guiding users through complex diagnostic tasks [7,8]. Within this broader AI landscape, vision–language models (VLMs) represent a particularly promising technology for inspection and analysis tasks because they can jointly process visual information and textual knowledge. By combining image understanding with language-based reasoning, VLM systems can interpret visual patterns, associate them with domain knowledge, and provide contextual explanations [9,10]. Within semiconductor inspection, recent studies have begun applying VLMs for defect analysis. For example, WaferMapVQA [25] reformulated wafer-map analysis as a structured visual question answering task encompassing anomaly detection, defect classification, localization, appearance description, and process-informed analysis achieving over 92.43% question level accuracy. Similarly, Jiang et al. [26] introduced FabGPT, a wafer-focused multimodal model that detects microscopic defects in SEM images and performs root-cause Q&A; FabGPT demonstrated significant improvements in wafer defect detection and knowledge querying compared to prior methods. These examples highlight the potential of VLMs to fuse visual and textual information for complex semiconductor defect reasoning tasks. In this paper, we introduce **INSIGHT-X**, INtelligent System for Image-based Guidance and Heuristic Troubleshooting in X-ray. Our INSIGHT-X is a VLM system that is designed to support artifact diagnosis, localization, and explainable answer with root-cause and mitigation for each artifact type in text in X-ray inspection workflows. INSIGHT-X structures artifact handling as an end-to-end question-answering pipeline that integrates imaging data, expert knowledge, and multimodal AI reasoning.

The main contributions of our work are as follows:

a) We introduce INSIGHT-X, a multimodal vision–language framework for automated X-ray artifact analysis in semiconductor CT and laminography inspection. We formulate artifact diagnosis as a vision-language question-answering task, enabling the model to identify artifact types, localize affected regions, explain underlying physical causes, and interpretable mitigation recommendations in a structured JSON format.

b) We develop an artifact-centric multimodal dataset consisting of curated X-ray reconstruction images, expert-defined artifact taxonomies, bounding-box annotations, and structured question–answer pairs that connect visual evidence with artifact causes and recommended corrective actions.

c) We propose a localization-aware fine-tuning strategy for Qwen2.5-VL that combines semantic supervision with geometric constraints to improve artifact localization and structured response generation.
d) We demonstrate through extensive experiments and ablation studies that domain-specific VLM fine-tuning substantially improves artifact classification and localization accuracy compared with zero-shot inference, establishing the feasibility of VLM-based assistants for semiconductor X-ray inspection and troubleshooting.

II. Background and Related Works

A. X-ray Artifacts in Microelectronics

In X-ray tomography and laminography of PCB and chip assemblies, several well-known artifacts can degrade image quality. We are going to discuss most commonly seen artifacts such as beam hardening, ring artifacts, geometric misalignment artifacts and missing cone here. Beam-hardening artifacts arise from the polychromatic nature of X-ray sources: lower-energy photons are preferentially absorbed during penetration of dense metallic structures, producing cupping effects and streak distortions that corrupt voxel intensities and obscure fine features [6]. There are different techniques that can be applied to get rid of this artifact. Recent deep-learning methods have begun to address this type of artifact by learning a nonlinear mapping between artifact-corrupted and artifact-free images in both the image and sinogram domains [11]. Ring artifacts are caused by detector pixel non-uniformities or source instabilities. It manifests as concentric circular patterns in reconstructed slices that can mimic true structural defects. Geometric misalignment artifacts result from mechanical drift or miscalibration during scanning, introducing projection inconsistencies that degrade spatial accuracy throughout the reconstructed volume. In laminographic imaging systems which is designed for flat, high-aspect-ratio samples such as printed circuit boards and advanced packages, incomplete angular sampling can also introduce missing-cone artifacts that further distort reconstructed volumes [5]. The tilted acquisition geometry restricts the angular sampling range, causing missing-cone artifacts that produce anisotropic blurring and inter-layer aliasing.

B. Need for Automated Artifact Detection and Remedy

Conventional artifact correction methods rely on detector calibration, beam-hardening correction, projection normalization, and iterative reconstruction algorithms [6]. More recently, deep learning has enabled artifact suppression by learning mappings between artifact-corrupted and artifact-free images. CNN-based methods have shown promising results for beam-hardening correction [11], residual networks have been applied to ring artifact removal [12], and diffusion-based frameworks have been proposed to mitigate missing-cone artifacts in laminography while preserving structural fidelity [13]. However, these methods primarily focus on post-reconstruction artifact suppression and do not provide automated artifact localization, explain artifact causes, or recommend corrective acquisition and reconstruction strategies. Current artifact mitigaiton in X-ray inspection relies on expert-driven, iterative workflows. Inspection engineers often rely on iterative adjustments of acquisition parameters, reconstruction settings, and calibration procedures to improve image quality, which is a time-consuming process that hinders throughput and consistency [3,4]. The absence of systematic automated diagnostic tools has motivated the broader application of large VLMs to imaging-heavy inspection domains. In industrial anomaly detection, models such as AnomalyGPT demonstrate that LVLMs can simultaneously localize defects and generate natural-language explanations without requiring manually specified thresholds [14]; contrastive cross-modal training strategies further improve anomaly localization by aligning visual and textual features in a shared embedding space [15].

C. Vision–language models for Object Detection

Object detection has evolved from conventional closed-set detectors to multimodal VLMs capable of open-vocabulary understanding and explainable reasoning. Early deep learning methods, including Faster R-CNN[16], YOLO[17], and DETR[18], significantly improved object localization accuracy and inference efficiency by introducing region proposal networks, single-stage regression, and transformer-based end-to-end detection, respectively [16-18]. However, these approaches are limited to predefined object categories learned during training and typically produce only bounding boxes and class labels without contextual reasoning or natural-language explanations.
Vision–language models have seen rapid progress in grounded spatial understanding and open-vocabulary perception. The introduction of CLIP demonstrated that aligning images and text within a shared embedding space enables zero-shot recognition of unseen categories through language supervision, laying the foundation for open-vocabulary visual understanding. Building upon this concept, GLIP reformulated object detection as phrase grounding, allowing localization of arbitrary objects specified by natural-language prompts. More recently, large VLMs such as LLaVA have extended this capability through visual instruction tuning, enabling multimodal dialogue, visual reasoning, and interpretable explanation generation. Grounding DINO further integrated open-vocabulary detection with transformer-based localization, while SpatialVLM [9] enhanced spatial reasoning and object localization from natural-language queries.

These advances demonstrate that modern VLMs combine visual understanding, localization, and language reasoning within a unified framework, making them well suited for X-ray artifact analysis [19-22].

III. Experimental Design of INISGHT-X

INSIGHT-X workflow begins with X-ray scan acquisition and reconstruction of electronics samples, collecting artifact-affected images and then annotating the artifacts with bounding box coordinates that are paired with question answer sets to build an artifact-centric question answer (QA) dataset (shown in Fig.1). The QA dataset links visual evidence to artifact identification. A key innovation of INSIGHT-X is the integration of VLM capable of not only recognizing artifact types but also spatially localizing artifact regions within inspection images. Through parameter-efficient fine-tuning on the factual QA dataset and annotated artifact images, the VLM learns to associate visual patterns with artifact categories while generating localization outputs such as bounding regions that highlight the affected areas. This capability allows the system to provide interpretable visual explanations alongside textual responses which will enable operators to quickly identify problematic regions within reconstructed X-ray images.

D. VLM Dataset Generation

Artifact Categories: Our dataset was prepared to support multimodal learning for artifact understanding in semiconductor X-ray inspection workflows. The dataset consists of reconstructed X-ray images acquired from both CT and X-ray laminography systems used for the inspection of PCBs and advanced semiconductor packages. Artifact categories were selected based on their prevalence in industrial inspection environments and their impact on reconstruction quality and defect interpretation. As illustrated in Fig. 1, the selected four artifacts were: beam-hardening, ring artifacts, geometric misalignment artifacts, and limited-angle/missing-cone artifacts. Overall, the selected categories provide a physically meaningful representation of artifact formation mechanisms and enable the model to learn relationships between visual appearance, root causes, and corrective actions.

Data Curation: The curated dataset (shown in Figure 2) was selected from 6,199 raw images using a bounding-box-centric diversity strategy, rather than random sampling. The visual diversity was measured from cropped artifact regions inside bounding boxes, ignoring background and whole-image variation. Each crop was represented using content features such as intensity statistics, contrast, sharpness, Sobel edge response, LBP texture, and HOG gradient descriptors. These 36-dimensional features were standardized and compressed using PCA [23] before applying Maximin farthest-point sampling [24] to retain visually diverse but non-redundant images. For sequential CT slice stacks, a temporal Maximin strategy with a minimum frame gap was used to avoid selecting many nearly identical neighboring slices. This process reduced the dataset to 1,339 representative images, preserving artifact diversity while removing redundant slices from the original 6,199-image collection.

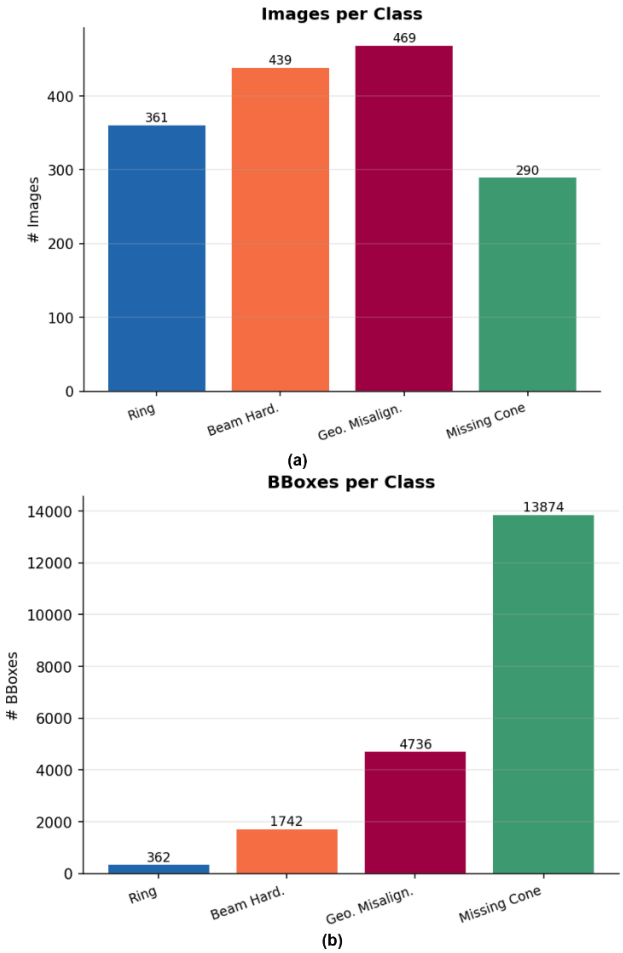


Figure 2: Distribution of the dataset across the four artifact categories, showing (a) the number of images containing each artifact class and (b) the total number of annotated bounding boxes per class.

Artifact bbox annotation: For artifact localization, bounding-box annotations were performed in YOLO format, where each artifact instance was assigned a class identifier and a normalized bounding box. The bounding boxes were subsequently converted into pixel coordinates during dataset generation and stored together with artifact labels. This annotation strategy enables visual grounding and localization tasks, allowing the VLM to identify both the presence and spatial location of artifacts. Each reconstructed image was manually reviewed by domain experts and assigned one or more artifact labels for each image.

QA dataset generation for VLM training: We generated structured question–answer (QA) dataset to train multimodal reasoning models. Each image was linked to a set of task-oriented questions covering artifact presence detection, artifact-specific classification, artifact counting, and type of artifact identification. For every QA pair, the corresponding image, artifact labels, bounding boxes, mitigation and reasoning metadata were preserved within a unified JSON structure. The generated QA pairs cover four question types to train diverse visual reasoning capabilities. Binary presence questions determine whether any artifact exists in an image, while per-class presence questions verify the presence or absence of individual artifact categories. Counting questions estimate the number of unique artifact types present, and artifact type identification questions require the model to recognize and list all artifact categories visible in that category (example shown in Table 1).

Table 1: Examples of the four question types generated for the VLM dataset.

Question Type	Example
Binary Presence (whether the image contains any artifact by answering a simple yes/no)	Is there any artifact present in this image? Answer: yes/no. If yes then localizes the artifacts with bbox and provides reason and mitigation in answer text.
Per-Class Presence (whether a specific artifact class is present in the image and localizes them.)	Is ring artifact present in this image? Answer: yes/no. If yes then localizes all the ring artifacts with bbox and provides reason and mitigation in answer text.
Counting (Counts the total number of unique artifact categories present in the image.)	How many artifact types are present in this image? Answer: 2 Localizes all the artifacts with bbox and provides reason and mitigation for each type.
Artifact Type (Identifies and lists all artifact classes present in the image.)	What type of artifact is present in the image? Answer: Beam hardening artifact, and geometric misalignment artifact. Localizes all the artifacts with bbox and provides reason for each type.

Although the curated dataset was selected to maximize visual diversity, the final dataset still shows class imbalance at both the image level and bounding-box level, as shown in Fig. 2. To reduce the influence of this imbalance during model training, we used weighted sampling so that underrepresented artifact categories contributed more effectively to the fine-tuning process. This strategy was used to improve model reliability and reduce bias toward artifact classes with more training samples. In addition, we report class-wise performance to evaluate whether the model generalizes consistently across artifact categories.

E. Fine Tuning VLM Model

For our INSIGHT-X system, we adapted the Qwen2.5-VL-7B-Instruct vision–language model for X-ray artifact understanding in X-ray images. The model was trained using a parameter-efficient fine-tuning strategy based on Low-Rank Adaptation (LoRA), allowing domain-specific knowledge to be incorporated while updating only a small fraction of the total model parameters. LoRA adapters were inserted into both the language backbone and vision encoder modules, enabling the model to learn artifact-specific visual features and reasoning patterns while preserving the general multimodal capabilities acquired during large-scale pretraining. The adopted configuration used a LoRA rank of 16, scaling factor of 32, and dropout rate of 0.05, resulting in approximately 97.6 million trainable parameters, corresponding to only 1.16% of the full model size. The model was trained to generate structured JSON responses containing artifact presence indicators, artifact categories, bounding-box locations, artifact counts, and expert-level explanations of artifact causes and mitigation strategies. Training samples consisted of reconstructed CT or laminography images paired with instruction prompts and ground-truth JSON outputs. Two fine-tuning strategies were investigated. The first strategy employed a bounding-box-aware weighted cross-entropy loss objective that emphasized localization coordinates and artifact-class tokens during autoregressive generation. In our second strategy, we introduced a composite loss objective combining cross-entropy loss with geometric localization constraints and counting supervision to improve spatial accuracy and structured prediction consistency. We will discuss these two loss functions in detail in the next subsection.

F. Loss Function Formulation

Fine-tuning is initially performed using a conventional token-level cross-entropy loss with position-dependent weighting, defined as in the following equation (1):

$$L_{CE} = -\Sigma\, w_t \cdot \log p(y_t \mid y_{<t}, x) \quad (1)$$

where w_t denotes the scalar weight assigned to token t and $p(y_t \mid y_{<t}, x)$ is the predicted probability of the ground-truth token given image x and all preceding tokens. Bounding box coordinate tokens receive a base weight $w_{coord} = 12$ for bounding-box-annotated samples via a 1.5× localization multiplier. Artifact class name tokens are assigned $w_{class} = 3$ to reinforce multi-class discrimination, whereas all remaining tokens retain the default weight $w_{other} = 1$.

To explore the benefits of incorporating spatial information, we then extend the training objective by introducing a composite loss that includes geometry-aware auxiliary terms for improved localization fidelity. The composite loss is defined as:

$$L_{total} = \alpha_1 L_{CE} + \alpha_2 L_{L1} + \alpha_3 L_{count} + \alpha_4(1 - GIoU) + \alpha_5(1 - DIoU) \quad (2)$$

where L_{L1} penalizes absolute coordinate error between predicted and ground-truth box vertices, and L_{count} penalizes per-class instance count discrepancies. The Generalized Intersection-over-Union term (1 − GIoU) regularizes box overlap and spatial enclosure, while the Distance Intersection-over-Union term (1 − DIoU) constrains centroid proximity. The coefficients α_1, α_2, α_3, α_4, and α_5 control the relative contribution of each loss component. Based on experimental tuning, we found the optimal coefficients to be $\alpha_1 = 1.0$, $\alpha_2 = 0.5$, $\alpha_3 = 0.3$, $\alpha_4 = 0.3$, and $\alpha_5 = 0.3$. The cross-entropy component retains unit weight as the primary sequence-generation objective. An auxiliary weight of 0.5 is assigned to L_{L1} to penalize gross coordinate errors without overshadowing token-level learning, while equal weights of 0.3 for the geometric terms ensure balanced spatial supervision across both overlap quality and centroid alignment.

G. Experimental Setup

The curated dataset contained 1,339 annotated X-ray reconstruction images collected from independent scan folders. To ensure robust performance evaluation and avoid data leakage, a five-fold stratified cross-validation strategy was adopted. Rather than stratifying only on artifact class, we used a compound label consisting of the inspection folder and primary artifact category during fold generation. This approach preserved both class distribution and inspection-source diversity across training, validation, and testing partitions. The default split assigned approximately 60% of the data for training with 804 images, 20% for validation with 268 images, and 20% for testing with 267 images.

Different artifact classes were highly imbalanced in our dataset. For instance, images with ring artifacts appeared substantially less frequently than geometric misalignment artifacts. To address this issue, inverse-frequency class weights were computed from the training partition and incorporated into both data sampling and loss computation. A Weighted Random Sampler was used to oversample minority artifact classes during mini-batch formation, while an additional class-specific loss weight amplified the optimization signal from underrepresented categories. This dual strategy ensured that rare artifacts contributed proportionally during training and reduced bias toward dominant classes.

Overall, three experimental configurations were evaluated. The first was a zero-shot baseline in which the pretrained Qwen2.5-VL model was directly applied to the test set without adaptation. The second configuration employed supervised fine-tuning using the weighted cross-entropy objective. The third configuration utilized composite-loss fine-tuning, combining classification, localization, and counting objectives within a unified optimization framework. Model optimization was performed using AdamW with a learning rate of 2×10^{-4}, cosine learning-rate scheduling, 5% warmup, gradient accumulation, and bfloat16 mixed-precision training. Early stopping and checkpoint selection were based on validation loss.

H. Model Evaluation

Model performance was evaluated on the held-out test partition using metrics designed to assess visual recognition, localization accuracy and structured response generation capability. Artifact classification performance was measured using presence accuracy, exact artifact-label matching accuracy, macro-averaged F1 score, and artifact-count accuracy. These metrics quantify the model's ability to correctly identify artifact categories and determine the number of artifact types present within an image. Localization performance was evaluated using class-aware bounding-box matching at Intersection-over-Union (IoU) thresholds of 0.25, 0.50, and 0.75. Precision, recall, F1-score, and mean IoU were computed to assess spatial agreement between predicted and annotated artifact regions. To evaluate the quality of generated textual responses, token-level F1 score and Average Normalized Levenshtein Similarity (ANLS) [25] were calculated between predicted and ground-truth explanations. Additionally, JSON validity and schema-conformance metrics were used to verify whether generated outputs adhered to the required structured-response format.

The comparative evaluation between zero-shot inference, weighted cross-entropy fine-tuning, and composite-loss fine-tuning enabled quantification of the benefits of domain adaptation and task-specific supervision. Beyond conventional classification performance, the evaluation framework measured the model's ability to generate interpretable, spatially grounded, and machine-readable outputs, reflecting the practical requirements of semiconductor X-ray inspection and artifact troubleshooting workflows.

I. Ablation Studies

We conducted ablation studies to evaluate the contribution of domain-specific fine-tuning and different loss components in the proposed VLM framework. Following the fine-tuning setup described in Section E, we compared the zero-shot Qwen2.5-VL baseline with LoRA fine-tuned variants trained

using different loss objectives, including weighted cross-entropy loss, weighted cross-entropy with artifact-count supervision, weighted cross-entropy with GIoU- or DIoU-based localization supervision, weighted cross-entropy with L1 bounding-box regression, and the full composite objective that jointly combines semantic prediction, counting supervision, and bounding-box localization. All variants used the same model backbone, dataset split, prompt format, LoRA configuration, optimizer, and evaluation protocol, so that differences in performance could be attributed to the loss design. Model performance was evaluated from two perspectives. First, we assessed artifact localization and structured artifact prediction as described in Session H. Second, we evaluated question-answering performance across closed-form and free-text functions. Closed-form QA tasks were evaluated using accuracy. Free-text rationale reason and mitigation responses were evaluated using TF-IDF cosine similarity. In addition, overall text-generation quality and output structure were assessed using Text Token F1, ANLS.

Label-preserving Data Augmentation: To address the severe class imbalance in the artifact detection dataset, particularly the under-representation of geometric misalignment artifacts relative to dominant classes such as beam hardening, we adopted a label-preserving and bounding-box-aware augmentation strategy rather than immediately relying on generative synthesis. The augmentation pipeline was applied only to the training split and combined controlled geometric transformations with photometric perturbations while preserving the original artifact semantics. Geometric operations included horizontal and vertical flipping, 90°-based rotations, small translations, and moderate scaling, with bounding boxes remapped exactly after each transformation and discarded only when less than 25% of the box remained visible. Photometric operations included brightness, contrast, gamma adjustment, mild Gaussian blur, and Gaussian noise to simulate realistic X-ray acquisition and reconstruction variability without changing the artifact label. Transformations likely to corrupt labels, such as hue/saturation changes, elastic deformation, perspective warping, and large arbitrary rotations, were deliberately excluded because they could introduce non-physical distortions or fabricate misalignment-like patterns. This strategy increased minority-class exposure during training while maintaining clean supervision through geometrically consistent annotations.

IV. Experimental Results

In this section we are going to discuss all different experiments and their results with ablation studies.

J. Experimental Results

We evaluated the VLM model performance using three experimental configurations: (1) zero-shot inference using the pretrained Qwen2.5-VL-7B model, (2) supervised fine-tuning using the proposed weighted CE objective, and (3) fine-tuning using the composite loss function that incorporates classification, localization, and counting constraints. All evaluation was performed on a held-out test set consisting of 267 images spanning four artifact categories: ring artifacts, beam-hardening artifacts, geometric misalignment artifacts, and missing-cone artifacts.

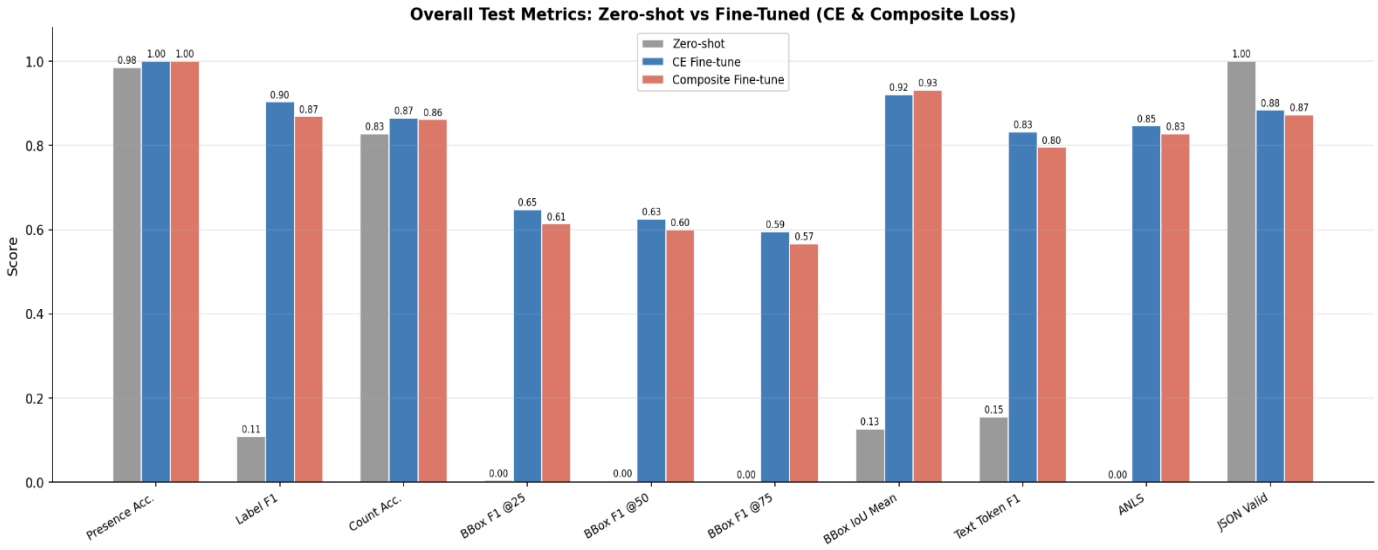


Figure 3: Overall performance comparison of the zero-shot, weighted CE fine-tuned, and composite loss fine-tuned models on test images across classification, localization, text generation, and JSON validity metrics, demonstrating substantial performance improvements after fine-tuning.

Table 2: Overall test metrics for each experiment: zero-shot model, weighted CE fine-tuned model, and composite loss fine-tuned model

Metric	Zero-Shot	CE Loss	Composite Loss
Presence Accuracy	0.985	1.000	1.000
Label Set Exact Acc	0.116	0.843	0.802
Label F1 (macro)	0.109	**0.904**	0.869
Count Accuracy	0.828	0.865	0.861
BBox F1 at IoU 0.25	0.004	0.680	0.656
BBox F1 at IoU 0.50	0.003	0.625	0.600
BBox F1 at IoU 0.75	0.001	0.595	0.566
BBox IoU Mean	0.127	0.921	0.932
BBox Prec at IoU 0.5	0.025	0.778	0.744
BBox Rec at IoU 0.5	0.002	0.522	0.502
Text Token F1	0.155	0.833	0.795
ANLS	0.000	0.847	0.828
JSON Validity	1.000	0.884	0.873
Schema Valid	1.000	0.884	0.873

The quantitative results on test images are summarized in Table 2 and Figure 3 and the class-wise performance is reported in Table 3.

All three models achieved high presence accuracy because both our curated train and test set contains only artifact-positive images, so the ground-truth for artifact presence is almost always true; therefore, any model that predicts artifact presence for most images can achieve near-perfect presence accuracy even if it fails to identify the correct artifact class or bounding boxes. Although the pretrained zero-shot model achieved a high artifact presence accuracy of 98.5%(indicating that the model could generally detect artifacts from images, shown in Table 2), its capability for artifact classification, localization, and structured answer generation was extremely limited. The zero-shot model achieved only 0.116 label exact-match accuracy and a macro label F1-score of 0.109, while localization performance remained poor with bounding-box F1 scores below 0.001 at all IoU thresholds. Although the zero-shot model achieves 100% JSON validity, its predictions are largely semantically incorrect, yielding an ANLS of 0.0 and a bounding-box F1 score of approximately 0.003. Conversely, the fine-tuned models exhibit a slight reduction in JSON validity due to occasional output truncation but achieve substantially higher performance across all task-specific evaluation metrics. This behavior represents an acceptable trade-off, reflecting the commonly observed alignment tax associated with task-specific fine-tuning, where minor degradation in general instruction-following behavior is exchanged for significant gains in task-specific accuracy.

Table 3: Performance comparison of the zero-shot model, weighted CE fine-tuned model, and composite loss fine-tuned model across four artifact classes using classification, localization, and artifact presence detection metrics.

Class	Metric	Zero-shot	CE Fine-tune	Composite Fine-tune
Ring	Label F1	0.4345	0.7130	0.6718
Ring	BBox IoU	0.1269	0.6517	0.6815
Ring	BBox F1@50	0.0456	0.5203	0.5412
Ring	Presence Acc	1.0000	1.0000	1.0000
Beam Hard.	Label F1	0.0000	0.9770	0.9773
Beam Hard.	BBox IoU	0.0000	0.8568	0.7979
Beam Hard.	BBox F1@50	0.0000	0.8184	0.7560
Beam Hard.	Presence Acc	1.0000	1.0000	1.0000
Geo. Misalign.	Label F1	0.0000	0.9667	0.9492
Geo. Misalign.	BBox IoU	0.0000	0.7303	0.8730
Geo. Misalign.	BBox F1@50	0.0000	0.2898	0.3427
Geo. Misalign.	Presence Acc	0.9890	1.0000	1.0000
Missing Cone	Label F1	0.0000	0.9580	0.8772
Missing Cone	BBox IoU	0.0000	0.9853	0.9827
Missing Cone	BBox F1@50	0.0000	0.7273	0.6787
Missing Cone	Presence Acc	0.9474	1.0000	1.0000

The test results demonstrate that domain-specific fine-tuning substantially improves performance over the zero-shot baseline. After fine-tuning with the CE loss objective, label classification performance improved dramatically, achieving a label exact-match accuracy of 84.3%, macro label F1-score of 0.904, and count accuracy of 86.5%. Overall artifact presence accuracy also increased from 98.5% to 100%. Localization performance also improved significantly, reaching bounding-box F1-scores of 0.648, 0.625, and 0.595 at IoU thresholds of 0.25, 0.50, and 0.75, respectively. Furthermore, the model produced high-quality textual explanations, achieving a token-level F1-score of 0.833 and an ANLS score of 0.847.

Our composite-loss model produced slightly lower classification performance than the weighted CE model but achieved the highest localization accuracy. As shown in Table 1, the composite model achieved the highest mean bounding-box IoU of 0.932 compared with 0.921 for weighted CE fine-tuning. This improvement can be attributed to the additional localization-aware loss terms (L1, GIoU, and DIoU), which explicitly optimize spatial agreement between predicted and ground-truth artifact regions. Therefore, the weighted CE model is preferable for artifact classification and reasoning tasks, whereas the composite model offers advantages when precise artifact localization is required.

The experimental results from per-class analysis presented in Table 3 and Figure 3 further demonstrate the effectiveness of INSIGHT-X across diverse artifact categories. Beam-hardening, geometric misalignment, and missing-cone artifacts achieved label F1-scores exceeding 0.95 under weighted CE fine-tuning, indicating strong class discrimination capability. Ring artifacts remained the most challenging category where we achieved a label F1-score of 0.713 and a localization IoU of 0.652. This behavior is expected because ring artifacts often exhibit subtle circular intensity variations that may overlap with background reconstruction patterns. Interestingly, the composite-loss model improved localization for ring artifacts and geometric misalignment artifacts, achieving higher IoU and bounding-box F1 values than the weighted CE model. Across all artifact categories, both fine-tuned models achieved near-perfect presence accuracy, demonstrating robust detection performance regardless of artifact type. The observed per-class performance differences may partly reflect both artifact-specific visual difficulty and the remaining imbalance in the curated dataset. To mitigate this issue, weighted sampling was used during fine-tuning, and class-wise metrics were reported to evaluate model reliability across artifact categories. The results show that fine-tuning substantially improves performance for all artifact classes, although ring artifacts remain the most challenging category. This suggests that future improvement should focus on targeted data expansion for underrepresented and visually difficult artifact classes, together with continued evaluation using stratified per-class metrics.

K. Results from Ablation Studies

The ablation results show that fine-tuning greatly improved both localization and structured VLM response quality over the zero-shot baseline. As shown in Table 4, the zero-shot model had near-zero bounding-box F1 and weak label recognition, indicating that domain adaptation is necessary for spatially grounded artifact prediction. Among the fine-tuned models, weighted CE Loss, Composite Loss, and weighted CE + L1 achieved the strongest overall performance. Weighted CE Loss produced the best bounding-box F1, Composite Loss achieved the highest mean IoU, and CE + L1 performed competitively in label-set prediction. In contrast, the GIoU and DIoU variants showed weaker localization, suggesting that these geometric losses were less effective in this autoregressive VLM setting.

Table 4. Ablation comparison of artifact localization and label-set prediction across VLM training objectives.

Variant	Zero-shot baseline	Composite	CE only	CE + GIoU	CE + DIoU	CE + Count	CE + L1
BBox F1 @ 0.25	0.004	0.614	**0.648**	0.016	0.018	0.107	0.617
BBox F1 @ 0.50	0.003	0.600	**0.625**	0.003	0.005	0.082	0.583
BBox F1 @ 0.75	0.001	0.566	**0.595**	0.001	0.003	0.063	0.529
BBox IoU Mean	0.127	**0.932**	0.921	0.150	0.169	0.633	0.869
BBox Prec @ 0.5	0.025	0.744	**0.778**	0.005	0.007	0.180	0.734
BBox Rec @ 0.5	0.002	0.502	**0.522**	0.002	0.004	0.053	0.483
Label Set Exact Acc	0.116	0.802	0.843	0.614	0.018	0.607	**0.846**
Label F1 (macro)	0.109	0.869	**0.904**	0.675	0.005	0.629	0.901

For question-answering performance, Table 5 and Figure 4 show that fine-tuning substantially improved structured VLM responses across question types, but the remaining errors are concentrated in harder recognition tasks. Binary artifact presence was nearly saturated after fine-tuning, suggesting that simple artifact detection is not the main bottleneck. Instead, artifact-type identification and counting remain more challenging, especially in multi-artifact images. The per-class QA reports show that ring artifact is consistently the weakest category: the zero-shot model achieved only 28.8% ring-artifact accuracy, and ring remained lower after fine-tuning, with 84.1% for CE + L1, 81.5% for Composite Loss, and 71.8% for CE + DIoU. In contrast, beam-hardening, geometric misalignment, and missing-cone artifacts generally achieved higher per-class QA accuracy, such as 98.3%, 99.6%, and 100.0% under CE + L1. These results indicate that fine-tuning improves artifact-specific QA performance, but the remaining errors are more likely to occur in visually subtle or multi-artifact cases. Because the current dataset and held-out test set are

still relatively limited in size, the near-perfect performance on some tasks should be interpreted cautiously and may decrease as more diverse unseen data are collected. Future work will therefore expand the dataset, especially for difficult and underrepresented artifact categories, and evaluate the model on larger and more diverse test sets. Overall, weighted CE-based fine-tuning provides the strongest general QA performance, while Composite Loss offers a balanced trade-off between QA and artifact localization performance.

Table 5. Ablation comparison of QA text generation and structured-output validity across VLM training objectives.

Variant	Zero-shot baseline	Composite	CE Loss	CE + GIoU	CE + DIoU	CE + Count	CE + L1
Text Token F1	0.155	0.795	**0.833**	0.738	0.727	0.775	0.825
ANLS	0.000	0.828	0.847	0.841	0.832	**0.879**	0.835
JSON Validity	**1.000**	0.873	0.884	0.918	0.918	0.985	0.873
Schema Valid	**1.000**	0.873	0.884	0.918	0.918	0.985	0.873

For the question-answering evaluation, the ablation tables show that fine-tuned models maintained strong response quality across the five answer functions: binary artifact presence, per-class artifact identification, artifact counting, artifact-type recognition, and reason/mitigation explanation. Counting questions achieved consistently high performance across variants, indicating that structured artifact-count generation was comparatively stable. For binary presence, per-class identification, artifact type, and reason/mitigation responses, the fine-tuned models generally preserved or improved semantic answer quality compared with the zero-shot baseline. These results indicate that the composite-loss design improved spatial grounding without compromising the VLM's structured question-answering ability.

Impact of Label-Preserving Data Augmentation on Model Performance: The augmented training strategy improved the model's overall detection and localization behavior compared with the baseline (as illustrated in Figure 5 below). The aggregate detection performance increased most clearly at the relaxed localization threshold, where F1 improved from approximately 0.60 to 0.78 at IoU ≥ 0.10, while performance at stricter thresholds remained comparable, with F1 around 0.62 at IoU ≥ 0.25 and 0.58 at IoU ≥ 0.50.

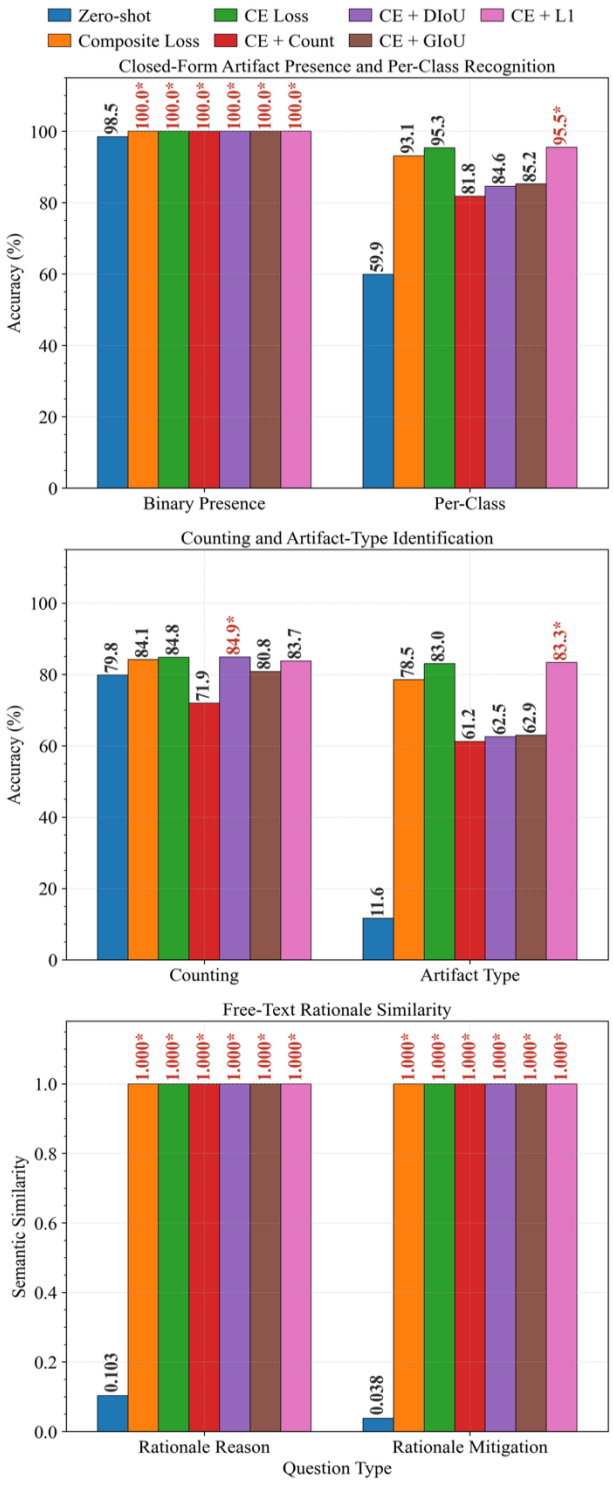


Figure 4. Function-specific QA performance across VLM ablation variants. counting and localization-oriented supervision for artifact-aware structured prediction.

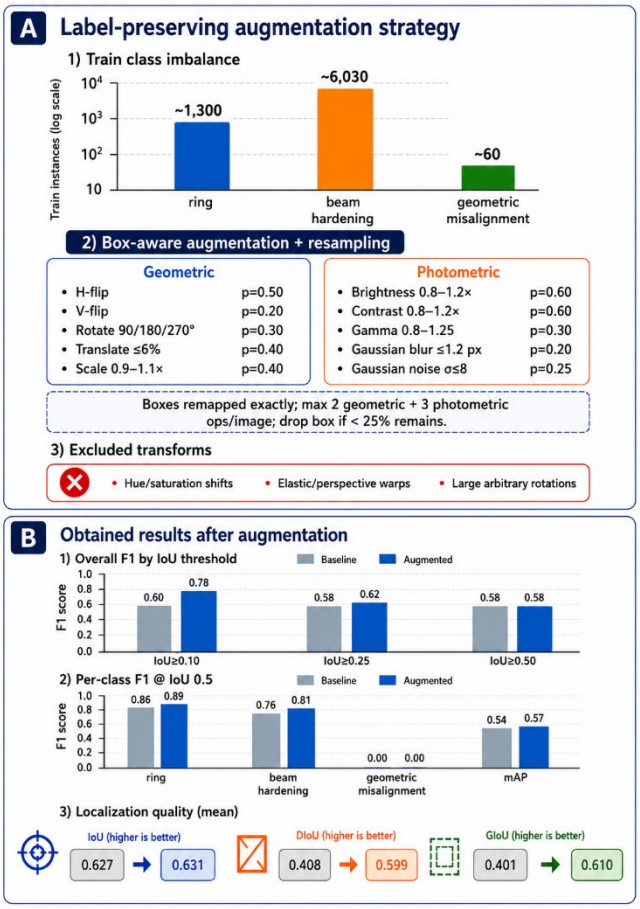


Figure 5. Overview of the proposed label-preserving, box-aware augmentation strategy and its effect on artifact detection. (a) Class imbalance is mitigated using geometric and photometric augmentations with exact bounding-box remapping, while label-corrupting transforms are excluded. (b) Augmentation improves overall F1, per-class detection for ring and beam-hardening artifacts, and localization quality measured by mean IoU, DIoU, and GIoU; geometric misalignment remains the most challenging class.

The localization distributions also showed better spatial agreement after augmentation: the mean IoU remained stable at approximately 0.63, while the mean DIoU and GIoU increased from 0.408 and 0.401 to 0.599 and 0.610, respectively, indicating improved center-distance and generalized overlap quality. Per-class analysis showed consistent gains for the dominant detectable classes, with ring artifact F1 improving from 0.86 to 0.89 and beam hardening F1 improving from 0.76 to 0.81 at IoU ≥ 0.5. The label-confusion results also became more balanced after augmentation, with stronger diagonal responses across artifact categories. However, geometric misalignment still showed weak detection-level performance at IoU ≥ 0.5, suggesting that although augmentation improved label recognition and localization quality overall, this class remains the primary bottleneck and may require additional real samples, targeted hard-example mining, or a later-stage validated synthesis strategy.

V. Discussion

The experimental evaluation demonstrates that domain-specific fine-tuning is essential for adapting general-purpose vision-language models to X-ray artifact analysis. Although the pretrained Qwen2.5-VL model achieved high artifact presence accuracy (98.5%), it performed poorly in artifact classification, localization, and structured reasoning, indicating that generic visual representations alone are insufficient for semiconductor inspection tasks. Fine-tuning substantially improved all downstream tasks scoring 100% in artifact presence accuracy, with the weighted CE model achieving the highest classification performance (macro F1 = 0.904) and strong question-answering quality, while the composite-loss model achieved the highest mean localization accuracy (IoU = 0.932). These results suggest that semantic supervision primarily benefits artifact recognition and reasoning, whereas explicit geometric constraints improve spatial grounding. The class-wise evaluation further reveals that beam-hardening, geometric misalignment, and missing-cone artifacts are recognized reliably after fine-tuning, whereas ring artifacts remain comparatively challenging because of their subtle circular intensity variations and visual similarity to reconstruction background patterns. The ablation studies further confirm that integrating localization-aware losses with counting supervision improves bounding-box quality without substantially degrading question-answering performance, highlighting the effectiveness of jointly optimizing semantic and geometric objectives for multimodal inspection assistants.

Despite these promising results, our approach has several limitations. The current dataset contains only four artifact categories and approximately 1,339 curated images collected from a limited number of CT and laminography inspection scenarios. Consequently, the model has not yet been evaluated on a broader range of acquisition systems, reconstruction algorithms, semiconductor package types, or previously unseen artifact distributions. Furthermore, localization accuracy decreases at stricter IoU thresholds, indicating that while the model can reliably identify artifact regions, precise boundary prediction remains difficult. The present framework also relies on manually annotated bounding boxes and predefined reasoning and mitigation knowledge, which limits scalability to newly emerging artifact types.

As the current dataset is limited to 1,339 curated images

spanning four artifact categories, our primary focus of future work will be expanding the dataset through the collection of additional industrial X-ray CT and laminography data, incorporating a wider range of reconstruction and acquisition artifacts, and increasing variability across devices and materials. Future versions of INSIGHT-X will also integrate retrieval-augmented expert knowledge, adaptive reasoning, and interactive troubleshooting capabilities to provide scan-specific mitigation recommendations. Finally, coupling the VLM with acquisition parameter optimization and real-time reconstruction feedback may enable closed-loop AI-assisted X-ray inspection that proactively minimizes artifact formation during image acquisition rather than only diagnosing artifacts after reconstruction.

VI. Conclusion

This paper presented INSIGHT-X, a multimodal vision-language framework for automated artifact diagnosis, localization, and mitigation guidance in X-ray computed tomography and laminography for advanced semiconductor inspection. By integrating an artifact-centric dataset, structured question–answer generation, expert knowledge, and parameter-efficient fine-tuning of a vision-language model, INSIGHT-X unifies artifact recognition, spatial localization, and explainable reasoning within a single framework. Experimental results demonstrate that domain-specific fine-tuning substantially improves artifact classification, localization, and structured response generation compared with zero-shot inference, while the proposed composite loss further enhances localization accuracy through geometry-aware supervision. These findings demonstrate the feasibility of employing vision-language models as intelligent assistants for semiconductor X-ray inspection, reducing dependence on manual expert interpretation while improving inspection consistency and decision support. The proposed framework also provides a foundation for future AI-assisted scan optimization and adaptive X-ray acquisition, where imaging parameters and reconstruction strategies can be guided automatically to minimize artifacts and improve reconstruction quality, enabling more reliable and autonomous non-destructive inspection workflows for next-generation semiconductor manufacturing.

Acknowledgment

We humbly thank all the co-authors who helped in supervising, writing, experimentation and extensive data annotation for this paper.

References

[1] B. Peterson, M. Kwan, F. Duewer, A. Reid and R. Brooks, "Optimizing X-Ray Inspection for Advanced Packaging Applications," in *International Symposium on Microelectronics*, Sept. 2020.

[2] C. Schmidt, "3-D X-Ray Imaging With Nanometer Resolution for Advanced Semiconductor Packaging FA," in *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 8, no. 5, pp. 745–749, May 2018.

[3] K. Yahyaei, M. S. M. Khan, G. Cadima, H. R. Kottur and N. Asadizanjani, "Design Guidelines for In-line X-ray Inspection in Advanced Packaging Technology: A CoWoS Case Study," in *International Symposium on Microelectronics*, pp. 338–344, 2025.

[4] H. Villarraga-Gómez, K. Crosby, M. Terada et al., "Assessing Electronics with Advanced 3D X-ray Imaging Techniques, Nanoscale Tomography, and Deep Learning," in *Journal of Failure Analysis and Prevention*, vol. 24, pp. 2113–2128, 2024.

[5] L. Helfen, T. Baumbach, P. Mikulík, J. Pernot and M. Cloetens, "High-Resolution Three-Dimensional Imaging of Flat Objects by Synchrotron-Radiation Computed Laminography," in *Applied Physics Letters*, vol. 86, no. 7, 2005.

[6] J. F. Barrett and N. Keat, "Artifacts in CT: Recognition and Avoidance," in Radiographics, vol. 24, no. 6, pp. 1679–1691, 2004.

[7] S. Yao, J. Zhao, D. Yu et al., "ReAct: Synergizing Reasoning and Acting in Language Models," in *Advances in Neural Information Processing Systems (NeurIPS),* 2023.

[8] Y. Liang, C. Wu, T. Song et al., "TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs," in *Intelligent Computing*, 2024.

[9] B. Chen, X. Zhang, Y. Li, J. Sun and L. Wang, "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[10] H. Laurençon, A. Touvron, L. Martin et al., "What Matters When Building Vision-Language Models?" in Advances in Neural Information Processing Systems (NeurIPS), 2024.

[11] X. Ji, D. Gao, Y. Gan, Y. Zhang, Y. Xi, G. Quan, Z. Lu, and Y. Chen, "A Deep-Learning-Based Method for Correction of Bone-Induced CT Beam-Hardening Artifacts," in *IEEE Transactions on Instrumentation and Measurement*, vol. 72, art. 4504012, 2023.

[12] T. Fu et al., "Deep-Learning-Based Ring Artifact Correction for Tomographic Reconstruction," *Journal of Synchrotron Radiation*, vol. 30, no. 3, pp. 620–626, 2023.

[13] T. Liu, "LaminoDiff: Artifact-Free Computed Laminography in Non-Destructive Testing via Diffusion Model," *arXiv preprint* arXiv:2601.07254, 2025.

[14] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang, "AnomalyGPT: Detecting Industrial Anomalies Using Large Vision-Language Models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 1932–1940, 2024.

[15] K. Qian, Z. Li, Y. Wang, H. Xu, J. Wu, X. Chen, Y. Luo, and H. Li, "Exploring Large Vision-Language Models for Robust and Efficient Industrial Anomaly Detection," *arXiv* preprint arXiv:2412.00890, 2024.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 779–788, 2016.

[18] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers (DETR)," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 213–229, 2020.

[19] A. Radford, J. W. Kim, C. Hallacy et al., "Learning Transferable Visual Models from Natural Language Supervision (CLIP)," in Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763, 2021.

[20] L. H. Li, P. Zhang, H. Zhang et al., "Grounded Language-Image Pre-Training (GLIP)," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10965–10975, 2022.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning (LLaVA),” in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
[22] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” in Proceedings of the European Conference on Computer Vision (ECCV), pp. 38–55, 2024.
[23] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 498–520, 1933.
[24] T. F. Gonzalez, "Clustering to Minimize the Maximum Intercluster Distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985.
[25] A. F. Biten, et al., “Scene Text Visual Question Answering,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4291–4301.
[26] S. C. Lin, et al. “A Multimodal Vision–Language Framework for Yield-Oriented and Explainable Wafer Defect Reasoning.” 2026 37th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC) (2026): 1-6.
[27] Y. Jiang, et al. “FabGPT: An Efficient Large Multimodal Model for Complex Wafer Defect Knowledge Queries.” 2024 ACM/IEEE International Conference On Computer Aided Design (ICCAD) (2024): 1-8.