READING ANOMALIES AS TYPES: ADAPTING FROZEN MODELS FOR INDUSTRIAL ANOMALY DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scarce and diverse defect examples make product-specific inspection models costly to train. Detecting a deviation from normal is only part of the task: diagnosis also requires identifying which product-specific inspection criterion it violates. We study defect diagnosis with frozen visual and vision-language models, combining normal-reference localization, targeted views, and textual inspection criteria. The target setting requires no defect-image exemplars at diagnosis time. Our experiments use descriptions derived from held-out defect images as a proxy for inspector knowledge; direct image access is tested only in separate controls. On public industrial benchmarks and semiconductor data, descriptions improve defect typing with model weights and visual inputs fixed. Correctly aligned descriptions outperform descriptions assigned to the wrong labels, supporting the importance of matching inspection criteria to defect types. Across separate controlled comparisons, visual marking primarily recovers missed defects, while descriptions can also correct type errors. Detector misses and uneven per-type performance limit deployment claims. These results identify textual inspection criteria as a means of adapting frozen models to product-specific diagnoses.

1 INTRODUCTION

Vision-language models (VLMs) have been studied extensively on everyday images and general-purpose tasks, including captioning, visual question answering, and visual reasoning (Laurençon et al., 2024; Zhao et al., 2024; Yang et al., 2024). Their practical utility under industrial inspection constraints remains less well understood. General benchmark performance does not establish whether a model can apply product-specific criteria to subtle defects. NaturalBench shows that strong VLMs can fail when answers depend on visual differences rather than language priors (Li et al., 2024). Industrial diagnosis therefore requires direct evaluation under the information constraints of inspection.

Defects occur infrequently, vary in appearance, and change with products and manufacturing conditions. Collecting representative examples for every inspection criterion can therefore make dedicated model development costly, particularly when a new product must be inspected before a comprehensive defect dataset exists. Industrial anomaly-detection research addresses limited target data and variation across products (Lv et al., 2025; Zhou et al., 2024). The practical question is how much inspection capability can be obtained without repeatedly learning from product-specific defect datasets.

Pretrained visual features (Oquab et al., 2024) support anomaly detection with limited examples (Zhai et al., 2026), and AnomalyDINO compares them with normal references while keeping the encoder fixed (Damm et al., 2025). VLMs can complement this evidence with textual instructions. Research on interpretable prompts and adaptation without labeled examples motivates examining how supplied context makes pretrained capabilities useful for a particular task (Du et al., 2024; Gadetsky et al., 2025).

Detecting a deviation is only part of inspection. A missing component, a bridge, and surface damage require different diagnoses. Visual distance from normal does not specify which criterion was

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

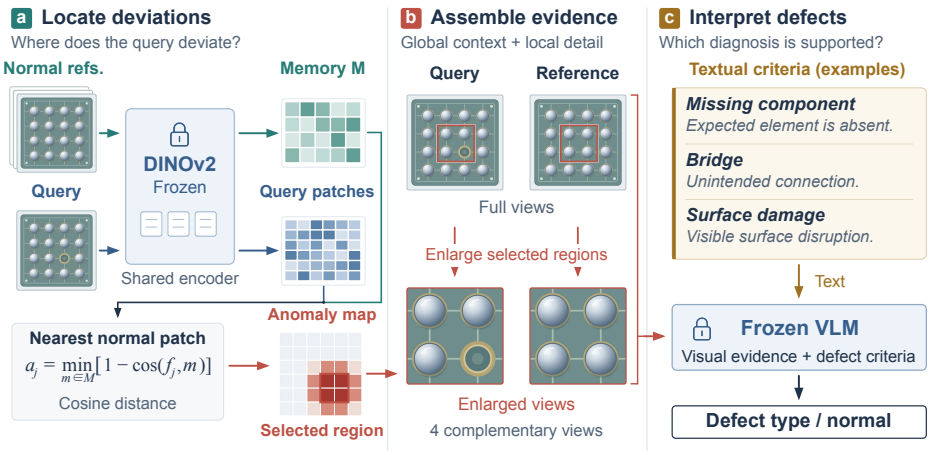

Figure 1: **Knowledge-guided diagnosis with frozen models.** **(a)** DINOv2 compares query patches with normal references to localize anomalies. **(b)** Full and enlarged query/reference views provide context and detail. **(c)** A VLM uses textual defect criteria to predict a defect type or NORMAL. The primary workflow requires neither defect-image exemplars at diagnosis time nor task-specific weight updates.

violated. Industrial understanding evaluations such as MMAD reflect this distinction (Jiang et al., 2025). Practical diagnosis requires applying product-specific information to identify the defect type.

Inspection knowledge provides task-specific information for this adaptation. Specifications and inspection experience describe expected structure, unacceptable deviations, and observations that distinguish defect types. Our target setting supplies this knowledge as text, alongside normal references, without defect-image exemplars at diagnosis time. In the experiments, descriptions derived from held-out defect images simulate the availability of inspector knowledge; their authoring is an experimental setup, not a proposed deployment requirement. We ask whether these criteria and visual evidence can help fixed models identify defects, distinguish their types, and avoid false alarms.

We address this validation gap through a controlled empirical study of industrial defect diagnosis. The configuration connects normal-reference patch comparison, targeted views, and a VLM supplied with descriptions of the allowed defect types (Figure 1). Experiments on public industrial benchmarks and production-line semiconductor images vary knowledge and visual inputs within fixed models. Our contribution concerns the practical value and limitations of existing capabilities; we introduce no new model architecture, learning algorithm, or training procedure. The comparisons identify which information improves diagnosis and how it changes inspection errors, providing evidence for applying pretrained models when product-specific defect data are scarce.

Our contributions are as follows:

1. **Practical evaluation of inspection knowledge.** We evaluate description effects with diagnostic weights and visual inputs fixed, alongside content and modality controls. Correctly aligned descriptions improve defect typing relative to deranged descriptions on public and semiconductor data.

2. **Error changes associated with text and spatial evidence.** Across their respective controlled comparisons, anomaly marking primarily recovers missed defects, while descriptions can also correct defect-type errors. We report these distinct patterns and the costs of detector gating.

3. **A normal-reference diagnosis configuration.** We connect normal-reference localization, targeted views, and textual criteria without task-specific weight updates. The primary diagnosis setting requires no defect-image exemplars and permits NORMAL predictions rather than forcing a defect label.

2 Related Work

General-purpose capabilities and context adaptation. VLM research examines recognition, captioning, and reasoning through controlled comparisons of backbones, architecture, and training (Laurençon et al., 2024). MMICL develops multimodal context schemes and training data (Zhao et al., 2024), while Lever-LM learns to select and order demonstrations (Yang et al., 2024). Joint inference over unlabeled inputs supports both unsupervised fine-tuning and in-context learning (Gadetsky et al., 2025). IPO optimizes human-readable prompts using image descriptions and performance history (Du et al., 2024); Frolic estimates distributions around textual prototypes and corrects label bias without labeled target data or parameter training (Zhu et al., 2024). These studies establish the value of context and descriptions for general-purpose tasks. We examine their practical role when a fixed diagnostic model must apply product-specific inspection criteria.

Visual evidence and reliability. General benchmark performance does not establish reliable use of visual evidence. NaturalBench exposes failures on paired natural images (Li et al., 2024), while diagram evaluations reveal reliance on background knowledge despite weak relationship understanding (Hou et al., 2025). Visual Sketchpad uses drawing and vision tools to support reasoning (Hu et al., 2024); SECOND combines multiscale evidence selection with contrastive decoding to reduce hallucination (Park et al., 2025). Machine Vision Therapy uses denoising in-context examples to obtain corrected predictions that supervise vision models under distribution shift (Huang et al., 2024). These works motivate assessing how supplied evidence changes errors. We distinguish missed defects, incorrect types, and false alarms when evaluating textual criteria and spatial evidence.

Industrial diagnosis and the validation gap. PatchCore models normal patch features (Roth et al., 2022), and AnomalyDINO uses frozen features for reference comparison (Damm et al., 2025). AnomalyCLIP learns object-agnostic prompts (Zhou et al., 2024), while instance-induced prompts support adaptation with few normal samples (Lv et al., 2025). MMAD evaluates industrial understanding and training-free enhancements (Jiang et al., 2025); JUDO combines inspection knowledge and normal comparison with supervised and reinforcement learning (Kang et al., 2026). AnomalyNCD studies anomaly-class discovery (Huang et al., 2025). Despite this progress, the practical utility of general-purpose VLM capabilities remains less established when diagnosis requires supplied criteria but cannot access defect-image exemplars. We address this gap through controlled evaluation on public industrial data and production-line semiconductor images, assuming a supplied defect taxonomy. Our contribution is empirical evidence about the diagnostic benefits and error trade-offs of existing components, rather than a new algorithm or training method. Broader industrial deployment requires further validation.

3 Knowledge-Guided Diagnosis with Frozen Models

Our method adapts pretrained visual and vision-language models to a target inspection task by providing product-specific inspection knowledge in text, without updating model weights. As illustrated in Figure 1, the pipeline separates three operations: *locating where the query deviates from normal*, *assembling visual evidence for that deviation*, and *interpreting the evidence using text-defined defect criteria*.

For product p, we assume an allowed defect taxonomy T_p and textual inspection knowledge $K_p = \{k_c : c \in T_p\}$, where k_c describes the observable criteria for defect type c. The descriptions are treated as given prior knowledge during diagnosis. A frozen diagnostic VLM combines this knowledge with visual evidence from the query and a small set of normal references R_p, and predicts either a defect type in T_p or NORMAL.

3.1 Locate deviations from normal

We first use a frozen DINOv2 encoder to identify regions that differ from normal appearance. The same encoder extracts patch features from the query x and the normal-reference images R_p. Let M denote the memory of normal-reference patch features. For each query patch f_j, we compute its distance to the nearest normal-reference patch as

$$a_j = \min_{m \in M} \left[1 - \cos(f_j, m)\right]. \tag{1}$$

The resulting patch-wise anomaly scores form an anomaly map. The highest-scoring region is selected as the candidate region for subsequent diagnosis. This stage intentionally provides only spatial anomaly evidence: it determines *where* the query differs from the available normal references without assigning a semantic defect label.

3.2 Assemble complementary visual evidence

A localized anomaly alone may be insufficient for distinguishing visually similar defect types. We therefore preserve both local detail and global product context.

For each query, the diagnostic input contains four complementary views: (i) the full query image, (ii) an enlarged view of the selected query region, (iii) the corresponding region from a normal reference, and (iv) the full normal-reference image.

The enlarged views expose fine-grained differences around the candidate region, while the full views retain structural context needed to determine whether the local appearance constitutes a meaningful defect.

3.3 Interpret evidence with textual inspection knowledge

The final stage connects visual anomaly evidence to product-specific defect semantics. Let V denote the frozen diagnostic VLM and $\mathrm{views}(x, R_p)$ the assembled visual evidence. Diagnosis is performed as

$$\hat{y} = V(\mathrm{views}(x, R_p), K_p, T_p), \qquad \hat{y} \in T_p \cup \{\textsc{normal}\}. \tag{2}$$

The textual descriptions specify what visual observations support each defect type and therefore provide the task-specific information that is absent from generic pretrained representations. Importantly, adaptation occurs through the diagnostic context rather than through parameter updates: the visual encoder and VLM remain frozen. In the primary diagnosis setting, the VLM receives normal references and supplied textual criteria, but no defect-image exemplars. Held-out defect images construct a controlled proxy for inspector knowledge in our experiments; direct access to them is evaluated only in separate modality controls (Section 4).

This decomposition gives the two sources of information distinct roles. Normal references provide product-specific evidence about expected appearance and anomalous regions, whereas textual inspection knowledge provides the semantic criteria required to convert those deviations into named defect types. The resulting pipeline therefore adapts pretrained models to an inspection taxonomy by composing visual reference evidence with text-defined prior knowledge rather than by training a new product-specific classifier.

4 Experimental Design

Experimental scope. We use two frozen diagnostic models, Gemma-4-26B-A4B (Gemma4) and Qwen3.6-35B-A3B (Qwen3.6), throughout the experiments. We evaluate the proposed components in three settings summarized in Table 1: a public four-view knowledge study, a public single-image follow-up, and a semiconductor study. These experiments share methodological components but differ in visual inputs, gating, and normal-data construction. We therefore interpret comparisons within each setting rather than across them.

Public benchmarks. The public evaluation uses selected categories from MVTec AD (Bergmann et al., 2019) and Real-IAD (Wang et al., 2024), containing 349 and 300 evaluation defects, respectively. Every experiment in this paper uses the same two frozen diagnostic models, Gemma4 and Qwen3.6. The public studies query a hosted API; the semiconductor study serves the same two checkpoints locally with vLLM, because the partner's images cannot be transmitted off site.

Table 1: **Three experimental settings.** Separate comparisons using frozen Gemma4 and Qwen3.6, with different visual inputs, text, gating, and normal-data construction.

	Public four-view	**Public single-image**	**Semiconductor four-view**
Visual input	Full query, query crop, reference crop, full reference	Query only; original or marked with DINO boxes; no reference image in the VLM input	Full query, query crop, reference crop, full reference; matched crop controls
Text conditions	Names, definitions, region observations, combined	Names or generic criteria	Seven recorded knowledge variants; the same descriptions for both models
Gate	None in the primary knowledge comparison	Calibration-quantile gate; $q = .90$ in the main marking comparison	Ungated primary diagnosis; gated results reported separately
Normal data	Retrospective reference bank; disjoint-bank sensitivity analysis	64 reference + 64 calibration normals per product, disjoint from evaluation	Three folds; 4 reference + 4 calibration + 4 evaluation normals per fold
Scope of separation	Normal separation checked by regeneration	MVTec split images; Real-IAD disjoint normal object IDs	Disjoint normal lots within fold; defect lot independence unavailable

For the single-image follow-up, reference, calibration, and evaluation normals are disjoint. The earlier four-view knowledge study retained the normal references and derived views of the original setup. The disjoint-bank sensitivity analysis preserves the direction of the combined-versus-names gains, with cell-wise differences of at most 4.6 percentage points. Results from the two studies remain separate comparisons; the single-image factorial uses generic criteria rather than the concrete descriptions of the four-view study.

Semiconductor evaluation. We additionally evaluate on 133 top-view BGA package images collected from a production inspection line, covering 11 defect types and `good`. After holding out 31 defect images for knowledge construction and separate controls, the evaluation set contains 90 defects and 12 normals. The 12 normal images are rotated across three folds as reference, calibration, and evaluation normals, with four images assigned to each role per fold. Reference normals define the visual comparison inputs, while calibration normals determine the anomaly threshold. Each evaluation image is tested once. Defect-lot independence cannot be established because the corresponding lot metadata are unavailable.

Target setting. We consider a setting in which no defect images are available to the inspection system. Instead, the system has access to a small set of normal references and textual inspection knowledge supplied by human inspectors.

Experimental proxy for human inspection knowledge. Description authoring is an experimental setup for simulating the availability of human inspection knowledge. The diagnostic task takes the resulting criteria as supplied context; it does not require a description-authoring system at deployment. We combine generic textual criteria with observations derived from held-out defect images, which remain excluded from evaluation. The public and semiconductor studies use different construction procedures. Each description set is fixed throughout its corresponding evaluation, with separately authored sets used only in sensitivity controls. These descriptions are a controlled proxy, not independently authored expert specifications. Separate modality controls provide held-out defect images directly. They compare the tested input configurations, without isolating a pure effect of text versus images.

Metrics. Type accuracy is the fraction of all true defects assigned the correct defect type, including defects rejected by the gate or predicted as NORMAL. Defect TPR measures the fraction of defects assigned any defect label, and normal FPR measures the fraction of normal images assigned a defect label. For the semiconductor study, we additionally report macro-F1 over the 11 defect types.

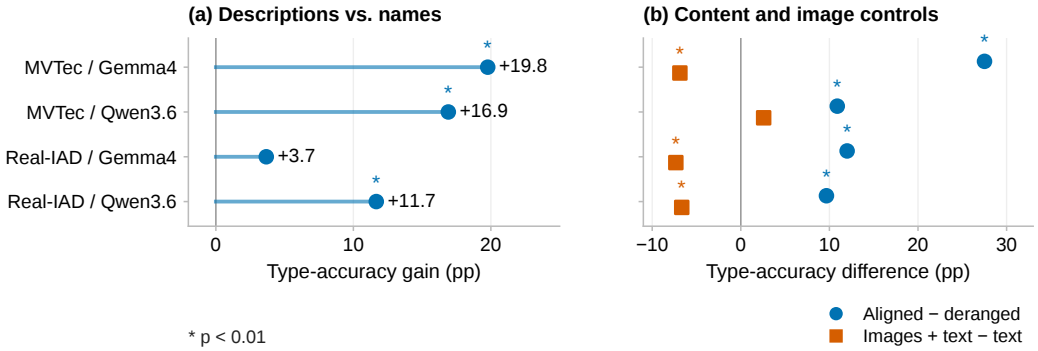

Figure 2: **Public diagnosis contrasts with retrospective references.** (a) Combined descriptions versus names. (b) Label alignment and adding image exemplars to combined descriptions. Positive differences favor the first condition. Asterisks denote $p < 0.01$ (Holm-adjusted in a; exploratory in b); pp: percentage points.

Statistical comparisons. Paired differences are tested with two-sided exact McNemar tests using a $p < 0.01$ threshold. For the twelve public knowledge-versus-names contrasts, p denotes the Holm-adjusted value; other reported comparisons use unadjusted values and are exploratory.

5 Experimental Results

We examine how textual inspection knowledge changes diagnosis with frozen models, then assess the roles of visual localization and detector gating. Comparisons within each setting hold the relevant visual inputs and model weights fixed.

5.1 Textual inspection knowledge improves defect diagnosis

Concrete descriptions help frozen models distinguish defect types. Combined descriptions improve type accuracy over names for both models on MVTec and for Qwen3.6 on Real-IAD (Figure 2a). Improvements extend across most product/model comparisons, but are not universal.

The benefit depends on what the descriptions convey. Generic definitions yield smaller, less consistent changes and can reduce accuracy. Concrete region observations are more effective and often exceed the combined descriptions in point estimates. This pattern suggests that inspection-specific visual criteria, rather than longer label definitions alone, help the VLM interpret deviations. The errors corrected also depend on the model: in the MVTec names-to-combined comparison, Gemma4 mainly corrects wrong defect types, whereas Qwen3.6 mainly recovers defects previously predicted as NORMAL. Descriptions can therefore support both defect recognition and type discrimination.

Matching descriptions to labels matters more than simply supplying text. Correctly aligned descriptions outperform deranged descriptions under identical visual inputs in every dataset/model setting (Figure 2b). Deranged descriptions also score below names for Gemma4. These comparisons support the importance of label–description alignment, although they do not separate format and content effects.

Descriptions also outperform direct defect-image exemplars for both models on MVTec and for Qwen3.6 on Real-IAD. On MVTec, images alone improve Qwen3.6 accuracy over names. Adding images to descriptions reduces accuracy for Gemma4 on both datasets and for Qwen3.6 on Real-IAD. All of these differences were statistically significant ($p < 0.01$).

The mostly negative image-addition contrasts in Figure 2b summarize this inconsistency. Thus, under the tested prompt format, direct access to the authoring images does not provide a consistent benefit over their textual descriptions. The direction of the combined-versus-names gains is retained in the disjoint-bank sensitivity analysis.

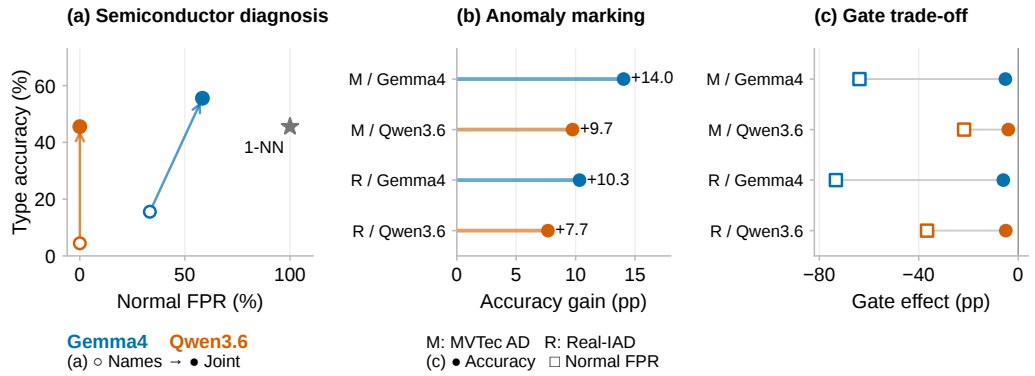

Figure 3: **Diagnosis and evidence trade-offs.** (a) Semiconductor accuracy and normal FPR, from names to joint descriptions; star: DINO 1-NN. (b) Public marking gains with generic criteria and gate fixed. (c) Gate-induced changes on the same disjoint-bank combined runs; negative values indicate reductions.

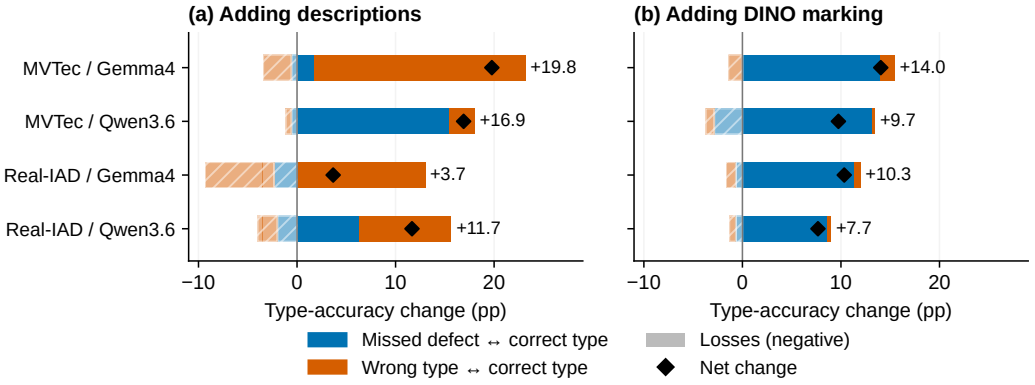

Figure 4: **Which errors are corrected?** (a) Names to combined descriptions, with four views fixed and no gate. (b) Original images to DINO marking, with generic criteria and gate fixed. Bars show corrections and reverse losses as percentages of all true defects; diamonds give net accuracy changes.

5.2 The knowledge effect extends to semiconductor inspection

Descriptions improve diagnosis on industrial data, but their effectiveness depends on the authoring model. Joint descriptions improve type accuracy over names for both Gemma4 and Qwen3.6 with the same visual inputs (Figure 3a). Correctly aligned descriptions also outperform deranged descriptions. Joint descriptions outperform direct image exemplars for both models. The importance of aligning criteria with labels therefore extends beyond the public benchmarks. All of these diagnosis gains were statistically significant ($p < 0.01$).

A gain over names remains with every evaluated authoring model, but its magnitude varies substantially. Gemma4's advantage over Qwen3.6 with the primary descriptions narrows when Gemma4 authors the text. Descriptions can be shared across diagnostic models, but their effectiveness and the relative performance of those models depend on the supplied criteria.

Feature-baseline comparisons depend on the supplied criteria. On both public benchmarks, Gemma4 with combined descriptions exceeds the evaluated DINO variants. With names alone, Qwen3.6 scores below the MVTec DINO prototype. These comparisons support the value of inspection criteria without establishing a uniform advantage over feature retrieval across cohorts and models. All of these differences were statistically significant ($p < 0.01$).

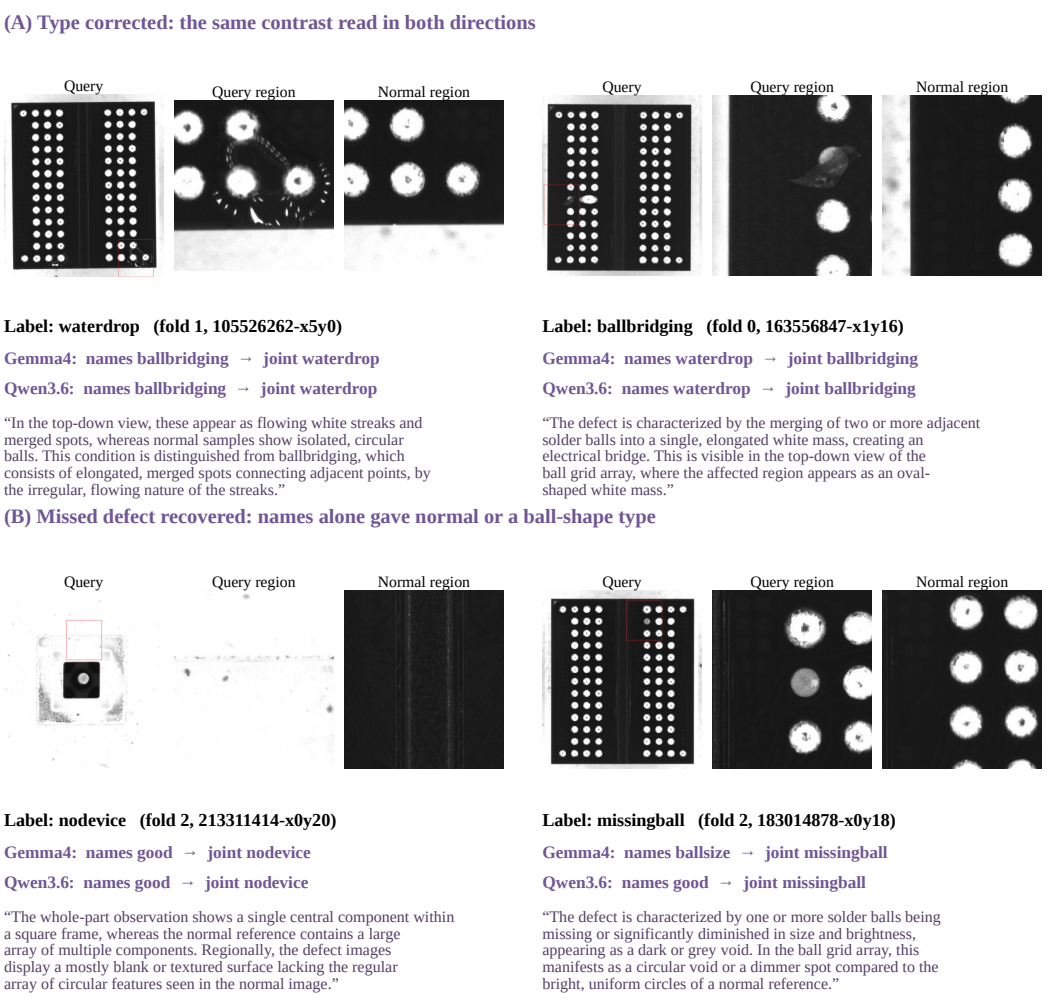

Figure 5: **Selected successful diagnoses with joint descriptions.** With identical views, Gemma4 and Qwen3.6 correct defect types (A) and recover missed defects (B). Red boxes mark detector proposals.

Unlike these feature classifiers, the VLM can predict NORMAL and use criteria that an inspector can read and revise. This flexibility has a cost: under joint descriptions, Qwen3.6 avoids normal false alarms but also misses more defects, while Gemma4 has higher recall and more false alarms. The two models exhibit different recall–false-positive trade-offs: in Figure 3a, both models move upward, but only Gemma4 moves toward higher normal FPR. Type accuracy alone therefore does not characterize their suitability for inspection.

The knowledge gain persists after excluding evaluation images unusually similar to held-out images and among cases that a retrospective strip-position lookup misclassifies. It is therefore not confined to cases explained by these identified shortcuts. However, position carries label information, and the lookup uses evaluation labels; these sensitivity checks do not establish generalization to independent manufacturing lots.

5.3 ERROR CHANGES WITH ANOMALY-GUIDED EVIDENCE AND GATING

Anomaly marking primarily helps the VLM notice missed defects. Normal-reference DINOv2 features detect most evaluated public defects under the disjoint-normal protocol. With generic criteria and the detector gate fixed, showing DINO-localized regions improves type accuracy for both models on both datasets (Figure 3b). Most corrected cases were previously predicted as NORMAL; fewer corrections concern wrong defect types.

Descriptions and marking change different error patterns. Figure 4 separates recovery of missed defects from correction of wrong defect types. Marking gains arise predominantly from missed defects becoming correctly typed in every dataset/model setting. Descriptions also resolve type confusions: this is the main source of Gemma4's corrected diagnoses on both datasets, whereas Qwen3.6's MVTec improvement mainly recovers missed defects. The effect of inspection criteria therefore depends on the diagnostic model's initial errors. The negative bars show correct diagnoses lost after each change; for Gemma4 on Real-IAD, these losses offset much of the description benefit. Spatial evidence and semantic criteria thus exhibit distinct error-correction patterns. Because the two studies use different visual inputs, text, and gating, these patterns do not establish synergy between their final configurations.

Detector gating reduces false alarms at the cost of diagnostic coverage. On the semiconductor cohort, the gate removes defects before the VLM can diagnose them, reducing correct diagnoses under both joint and joint-plus-region conditions. The post-hoc sweep suggests that conservative calibration contributes to these misses: alternative thresholds pass more defects without increasing observed detector false alarms. The pre-specified maximum rule remains the primary result.

The same trade-off appears in the disjoint-bank public four-view study: gating the combined condition lowers normal FPR while reducing type accuracy relative to the corresponding ungated run (Figure 3c). The paired reductions show that fewer false alarms come with lost diagnoses. Better ungated diagnosis therefore does not by itself ensure reliable performance after detector gating.

When the same description helps and when it does not. Figure 5 shows selected successful cases under names and joint descriptions. In row A, both models correct confusion between waterdrop and ballbridging with criteria distinguishing flowing streaks from an elongated merged mass. In row B, descriptions recover a missing device and a missing ball that names alone fail to identify. Cases were selected after scoring, excluding held-out, quality-excluded, and near-held-out candidates; they illustrate observed changes without establishing their cause or frequency.

Success is not uniform across the cohort. Neither model correctly diagnoses sawoffset under joint descriptions, and model-dependent errors remain even with identical views and text. Explicit criteria therefore help diagnosis without ensuring correct application in every case.

6 Limitations

The evaluation covers selected public benchmarks and a semiconductor cohort from one production line. Broader validation across industrial sectors, manufacturing sites, independent production lots, imaging conditions, and defect taxonomies is needed to assess transfer to varied inspection settings. Positional and near-duplicate cues remain possible sources of bias in the semiconductor data. The experimental descriptions are a proxy for inspector knowledge, their effects depend on the authoring model, and independently authored expert criteria remain untested. Per-type failures and model-dependent recall–false-positive trade-offs also limit deployment readiness, while localization and final diagnosis depend on detector calibration. Some analyses are exploratory, including the post-hoc threshold sweep, and require confirmation under independently specified evaluation protocols.

7 Conclusion

This study examines textual inspection knowledge for frozen-model diagnosis, targeting settings with normal references and supplied criteria but no required defect-image exemplars. In the experiments, descriptions constructed from held-out defect images serve as a controlled proxy for inspector knowledge. Detector misses, label–position associations, and the absence of independent-lot and expert-authored evaluations limit the deployment claim. The results show that inspection-specific descriptions can improve defect typing with fixed model weights and visual inputs. Across separate controlled comparisons, descriptions can correct type errors as well as missed defects, while anomaly marking primarily recovers missed defects and gating trades recall for fewer false alarms.

Reproducibility Statement

The code and data required for the public-benchmark implementation are included in our anonymous repository at https://anonymous.4open.science/r/anomalies_as_type/. The repository also provides fixed inspection descriptions, prompt templates, and detailed instructions for data preparation, inference, and evaluation. The supplementary material includes per-image predictions, scoring code, protocol locks, and public dataset splits. These records allow the reported semiconductor statistics to be recomputed without the images. Due to security restrictions, the semiconductor industrial dataset cannot be disclosed during peer review. If the paper is accepted, we will explore releasing a subset subject to security review.

AI Use Statement

AI services such as ChatGPT, Gemini, and Claude, assisted with language editing, code generation, and improving presentation quality. Analysis code was drafted with LLM assistance and executed under author supervision. VLM-generated inspection notes are part of the experimental construction of the inspection-knowledge condition described in Section 4, rather than manuscript-writing assistance.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1319–1329. IEEE, 2025.

Yingjun Du, Wenfang Sun, and Cees G Snoek. Ipo: Interpretable prompt optimization for vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126725–126766, 2024.

Artyom Gadetsky, Andrei Atanov, Yulun Jiang, Zhitong Gao, Ghazal Hosseini Mighan, Amir Zamir, and Maria Brbic. Large (vision) language models are unsupervised in-context learners. In *International Conference on Learning Representations*, volume 2025, pp. 25046–25070, 2025.

Yifan Hou, Buse Giledereli, Yilei Tu, and Mrinmaya Sachan. Do Vision-Language Models Really Understand Visual Language? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 23910–23959, 2025. URL https://proceedings.mlr.press/v267/hou25c.html.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379, 2024.

Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19973–20003. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24o.html.

Ziming Huang, Xurui Li, Haotian Liu, Feng Xue, Yuzhe Wang, and Yu Zhou. Anomalyncd: Towards novel anomaly class discovery in industrial scenarios. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4755–4765. IEEE, 2025.

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 *International conference on learning representations*, volume 2025, pp. 87273–87295, 2025.

Hyunju Kang, Woohyun Lee, Jaewon Kim, and Hogun Park. Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa. *arXiv preprint arXiv:2605.20284*, 2026.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 87874–87907, 2024.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean De Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems*, volume 37, pp. 17044–17068, 2024.

Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. In *International Conference on Learning Representations*, volume 2025, pp. 63657–63677, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. In *International Conference on Machine Learning*, pp. 48027–48040. PMLR, 2025.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14298–14308. IEEE, 2022.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 22883–22892. IEEE, 2024.

Xu Yang, Yingzhe Peng, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, and Hanwang Zhang. Lever lm: configuring in-context sequence to lever large vision language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 100341–100368, 2024.

Guangyao Zhai, Yue Zhou, Xinyan Deng, Lars Heckler-Kram, Nassir Navab, and Benjamin Busam. Foundation visual encoders are secretly few-shot anomaly detectors. In *International Conference on Learning Representations*, volume 2026, pp. 97353–97376, 2026.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. In *International Conference on Learning Representations*, volume 2024, pp. 14942–14980, 2024.

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*, volume 2024, pp. 49705–49737, 2024.

Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2001–2025, 2024.

A Implementation Details

Models and inference. The frozen diagnostic models are Gemma4 and Qwen3.6. Public runs query a hosted API; semiconductor runs serve the same checkpoints locally with vLLM because partner images cannot leave the site. The anomaly encoder is `facebook/dinov2-base`. Unless otherwise specified, decoding uses temperature zero and a short answer-only JSON schema. There is no task-specific weight training.

Cohorts. Public evaluation uses five MVTec AD products (cable, grid, metal_nut, screw, transistor; 349 defects and 202 normals) and five Real-IAD products (audiojack, pcb, switch, transistor1, usb; 300 defects and 60 normals). The semiconductor cohort is summarized in Table 2: 11 defect types plus NORMAL, 31 held-out defect images reserved for knowledge construction and separate controls, and 102 evaluation images (90 defects, 12 normals). The 12 normals rotate across three folds as reference, calibration, and evaluation. When used, the detector gate compares the query score with the maximum of the four calibration-normal scores in that fold.

Table 2: **Real-world semiconductor data: label inventory and split.** Four files with conflicting labels are quarantined. Support is globally excluded from evaluation; good is the normal class.

Class	Raw	Retained	Support	Evaluation
ballbridging	12	11	3	8
ballcount	13	12	3	9
ballpitch	8	8	3	5
ballsize	3	3	2	1
chipping	3	3	2	1
good	12	12	0	12
missingball	17	15	3	12
nodevice	12	12	3	9
packagesize	12	12	3	9
sawoffset	12	12	3	9
scratch	5	5	3	2
waterdrop	28	28	3	25
Total	137	133	31	102

Purpose and scope of knowledge construction. The authoring procedures simulate an inspection setting in which a human can supply product-specific criteria. They are an experimental proxy for that knowledge, not a proposed requirement for deployment. The diagnostic comparison treats each description set as fixed, and both diagnostic models receive the same set. Held-out authoring images are excluded from evaluation. Independently authored expert criteria are not evaluated here.

Public knowledge conditions. *Names* supplies the allowed type identifiers. *Definitions* uses texts authored from label strings and generic engineering knowledge, without opening dataset images or image-derived files. These texts are stored in `knowledge_only`; the authoring records are collected by `write_knowledge_priors.py`. They do not use the local semiconductor inspection-programme authoring procedure. *Region observations*, stored in `region_derived`, are written by `google/gemma-4-31B-it` from two declared held-out images per type. The author sees the full query with a DINO proposal, the enlarged query region, the corresponding normal region, and the full normal reference. The saved provenance records no ground-truth-mask or human-localization use for these region notes. *Combined*, stored in `combo_derived`, concatenates the generic definition and region note for each type, in that order, without an additional synthesis call. The local author revision below applies to the semiconductor run, not to all public text sources.

Semiconductor knowledge conditions. The primary semiconductor descriptions are written by `google/gemma-4-31B-it` at revision 842da3794eaa0b77d5f08bae87a17459d91ff475. *Names* supplies type identifiers only. *Generic definitions* and *Text contrasts* are authored from inspection-programme text without images. The former describes the expected appearance and inspection

view; the latter describes distinctions from confusable types. *Individual descriptions* merges per-image reviews with programme text. *Region observations* merges observations from DINO-marked authoring views. *Joint descriptions* uses a separate synthesis call to combine whole-image and region observations with pairwise contrasts from the held-out images. *Joint plus region* concatenates that joint description and its region note. Thus, semiconductor Joint is synthesized from observations and contrasts, whereas public Combined concatenates separately authored definitions and region notes. Each set is locked before its corresponding diagnostic evaluation; separate author and view variants are evaluated as sensitivity controls.

Direct-image controls. Public image controls provide the two held-out images per type used for region authoring. Semiconductor text authoring uses 31 held-out images (two or three per type), while direct-image controls select two per type from that pool, yielding 22 images per fold. The semiconductor text also incorporates programme information and pairwise contrasts. These controls compare the recorded configurations; they do not match the full information available during text authoring or isolate the causal effect of the transmission modality.

Statistical reporting. All reported paired tests use a two-sided exact McNemar test and a $p < 0.01$ threshold. The twelve public knowledge-versus-names contrasts form one Holm-correction family across datasets, models, and knowledge conditions. For those contrasts, threshold statements use adjusted p-values. Other reported tests are unadjusted exploratory comparisons. A result with $p \geq 0.01$ does not establish equivalence or the absence of an effect.

Prediction and validation records. Semiconductor inference uses local offline weights at fixed revisions, checked against the served checkpoint before predictions are written. Scoring code checks support exclusions and source hashes. Prediction records contain image identifier, ground-truth label, fold, condition, model, answer, detector score, and threshold. The supplementary records include the seven knowledge texts for 11 types and three folds, prompt templates, DINO prototype and 1-NN predictions, protocol and lock files, audit records, and exact public product, support, and normal splits.

B Supplementary Results

B.1 Public knowledge contrasts

Table 3: **Knowledge and gate comparisons with fixed four-view inputs.** Type accuracy and normal FPR (%). Top: retrospective references; bold marks the best accuracy. Bottom: combined descriptions before and after the $q = .90$ gate on the same disjoint-bank run.

Knowledge comparison: retrospective bank, ungated

Dataset	VLM	Names	Definitions	Region observations	Combined	FPR names	FPR combined
MVTec AD	Gemma4	46.70	53.01	**67.34**	66.48	80.2	74.8
MVTec AD	Qwen3.6	38.68	44.13	55.01	**55.59**	7.9	25.2
Real-IAD	Gemma4	55.00	52.33	**60.67**	58.67	88.3	71.7
Real-IAD	Qwen3.6	38.33	40.67	**50.33**	50.00	28.3	20.0

Gate comparison: disjoint bank, combined descriptions

Dataset	VLM	Ungated accuracy	Gated accuracy	Ungated FPR	Gated FPR
MVTec AD	Gemma4	67.62	62.46	75.2	11.4
MVTec AD	Qwen3.6	60.17	56.16	28.2	6.4
Real-IAD	Gemma4	58.67	52.67	76.7	3.3
Real-IAD	Qwen3.6	50.33	45.33	41.7	5.0

Tables 3 and 4 retain the complete public comparisons summarized in Figures 2 and 3c, including absolute scores and knowledge variants. Table 5 lists every public knowledge-versus-names difference behind Table 3. Combined and region descriptions have Holm-adjusted $p < 0.01$ in three of

Table 4: **Public content and modality controls.** Type accuracy (%) with four views fixed. Deranged permutes combined descriptions across types within each product; Images appends two labelled held-out defect images per type. Bold marks the best accuracy.

Dataset	Model	Names	Deranged	Combined	Images	Images + combined
MVTec AD	Gemma4	46.70	38.97	**66.48**	49.57	59.60
MVTec AD	Qwen3.6	38.68	44.70	55.59	46.13	**58.17**
Real-IAD	Gemma4	55.00	46.67	**58.67**	52.67	51.33
Real-IAD	Qwen3.6	38.33	40.33	**50.00**	35.00	43.33

Table 5: **All twelve public knowledge-versus-names contrasts.** Differences in type accuracy (percentage points) behind Table 3. Bold marks Holm-adjusted $p < 0.01$ across the twelve contrasts; all other contrasts have adjusted $p \geq 0.01$.

Dataset	Model	Definitions	Region observations	Combined
MVTec AD	Gemma4	+6.30	**+20.63**	**+19.77**
MVTec AD	Qwen3.6	+5.44	**+16.33**	**+16.91**
Real-IAD	Gemma4	−2.67	+5.67	+3.67
Real-IAD	Qwen3.6	+2.33	**+12.00**	**+11.67**

the four dataset/model settings. All generic-definition contrasts and all Real-IAD/Gemma4 contrasts have adjusted $p \geq 0.01$.

B.2 Detector and single-image marking factorial

Table 6: **Anomaly marking with text and gate fixed.** Type accuracy (%) under disjoint normal data and the 90th-quantile gate; differences are in percentage points and all four contrasts have $p < 0.01$.

Dataset	VLM	Original + criteria	DINO boxes + criteria	Δ
MVTec AD	Gemma4	18.05	32.09	+14.04
MVTec AD	Qwen3.6	16.62	26.36	+9.74
Real-IAD	Gemma4	10.33	20.67	+10.33
Real-IAD	Qwen3.6	8.00	15.67	+7.67

Table 6 retains the original and marked accuracies underlying Figure 3b. Table 7 reports the frozen DINOv2 detector alone under the disjoint-normal protocol at the 90th and 95th calibration quantiles (the 89.40% and 83.33% defect recall figures are its $q = .90$ rows) and expands the gated marking comparison in Table 6. Within each dataset and model, the regenerated normal bank, image size, schema, and gate are held fixed. With generic criteria fixed, adding DINO-localized regions raises type accuracy in all four combinations; observed normal false-positive counts do not increase.

B.3 Semiconductor types and controls

Table 8 retains all knowledge conditions, defect and normal metrics, and feature baselines accompanying Figure 3a. Table 9 gives the per-type counts behind the knowledge comparison in Table 8.

Table 10 collects the semiconductor controls reported in Section 5.2: descriptions deranged across types, the held-out defect images shown in place of or in addition to the descriptions, and the two shortcut checks that remove five evaluation images unusually similar to the held-out images or restrict to the 52 defects whose strip-column lookup is incorrect. A knowledge gain remains in both models under both restrictions.

Table 7: **Frozen DINOv2 detector alone and the complete single-image factorial at the 90th-quantile gate.** All rates are percentages under the disjoint normal bank of the single-image study. The detector-only rows use calibration normals for the threshold and report defect TPR and normal FPR at two calibration quantiles. In the factorial rows every arm permits NORMAL, macro-F1 includes normal-to-defect errors and the full candidate inventory, and detector and VLM misses remain in the recall and accuracy denominators.

Image / text	FPR	TPR	Type accuracy	Macro-recall	Macro-F1
Detector only, no VLM					
MVTec AD, $q = .90$	16.83	89.40	–	–	–
MVTec AD, $q = .95$	9.90	81.38	–	–	–
Real-IAD, $q = .90$	5.00	83.33	–	–	–
Real-IAD, $q = .95$	5.00	79.00	–	–	–
MVTec AD / Gemma4					
Original / names	6.44	38.68	15.19	17.86	16.79
DINO boxes / names	8.42	83.09	29.80	30.72	26.14
Original / criteria	2.97	35.82	18.05	21.49	23.50
DINO boxes / criteria	2.97	65.90	32.09	33.78	32.85
MVTec AD / Qwen3.6					
Original / names	0.50	20.34	13.75	17.92	20.07
DINO boxes / names	1.49	41.55	24.07	26.36	24.72
Original / criteria	0.50	24.36	16.62	19.57	22.86
DINO boxes / criteria	0.50	38.97	26.36	28.84	32.71
Real-IAD / Gemma4					
Original / names	0.00	13.33	8.67	8.29	7.97
DINO boxes / names	0.00	43.67	23.00	22.68	21.14
Original / criteria	0.00	17.67	10.33	10.13	8.97
DINO boxes / criteria	0.00	33.00	20.67	20.01	18.76
Real-IAD / Qwen3.6					
Original / names	0.00	9.00	5.67	5.63	6.15
DINO boxes / names	1.67	22.67	10.00	10.56	11.19
Original / criteria	0.00	11.33	8.00	7.79	8.49
DINO boxes / criteria	0.00	26.00	15.67	15.48	14.32

Table 8: **Semiconductor diagnosis with fixed views and descriptions.** Type accuracy, macro-F1, and normal FPR (%). Non-VLM baselines always assign a defect type.

Knowledge	Gemma4 accuracy	F1	FPR	Qwen3.6 accuracy	F1	FPR
Names	15.6	11.6	33.3	4.4	5.8	0.0
Generic definitions	23.3	13.0	33.3	8.9	13.5	0.0
Text contrasts	25.6	12.4	33.3	10.0	6.8	0.0
Individual descriptions	40.0	24.9	83.3	33.3	22.2	50.0
Joint descriptions	55.6	34.1	58.3	45.6	32.5	0.0
Region observations	50.0	34.5	58.3	43.3	26.0	33.3
Joint + region	56.7	37.1	50.0	47.8	31.6	25.0
Majority class (waterdrop)	27.8	–	100.0	27.8	–	100.0
DINO prototype, same held-out images	38.9	31.6	100.0	38.9	31.6	100.0
DINO 1-NN, same held-out images	45.6	34.4	100.0	45.6	34.4	100.0

Table 9: **Correct diagnoses by defect type.** Columns give names, joint descriptions (J), and joint plus region (J+R). Counts retain all evaluation defects, including misses.

Type	n	Gemma4 names	Gemma4 J	Gemma4 J+R	Qwen3.6 names	Qwen3.6 J	Qwen3.6 J+R
ballbridging	8	1	4	5	1	7	6
ballcount	9	0	2	1	0	0	2
ballpitch	5	0	0	1	0	0	0
ballsize	1	1	0	0	0	0	0
chipping	1	0	0	0	0	0	0
missingball	12	8	11	10	0	7	6
nodevice	9	0	8	9	0	6	7
packagesize	9	0	3	2	0	0	0
sawoffset	9	0	0	1	0	0	0
scratch	2	1	0	0	1	1	1
waterdrop	25	3	22	22	2	20	21

Table 10: **Controls on the semiconductor cohort.** Correct types over the defects in each row and false positives over 12 normals; DINO four-view inputs fixed. Deranged rows permute the frozen descriptions across the 11 types; image rows append the 22 held-out defect images per fold (two per type from the 31-image authoring pool), labelled, after the four views. The last four rows restrict the evaluation set to test two shortcuts: similarity to the held-out images and a leave-one-out strip-column lookup that reads no pixels. Deranged and image-only contrasts with their fixed-description counterparts have $p < 0.01$; adding images to joint descriptions has $p \geq 0.01$ for both models.

Knowledge input	Gemma4 correct	FP	Qwen3.6 correct	FP
Names	14/90	4	4/90	0
Joint descriptions (frozen)	**50/90**	7	41/90	0
deranged across types	32/90	2	26/90	0
Joint + region (frozen)	**51/90**	6	**43/90**	3
deranged across types	39/90	6	25/90	1
Names + 22 held-out images	12/90	2	21/90	0
Joint descriptions + 22 held-out images	39/90	2	**43/90**	1
Names, near-held-out candidates excluded	14/85	–	4/85	–
Joint descriptions, near-held-out candidates excluded	47/85	–	39/85	–
Names, column lookup incorrect	6/52	–	2/52	–
Joint descriptions, column lookup incorrect	20/52	–	13/52	–

Answers that follow the deranged text to the label now carrying the true type's description: 17/90 and 14/90 (Gemma4), 3/90 and 4/90 (Qwen3.6).

Table 11: **Sensitivity to authoring views.** Correct types over 90 semiconductor defects. Both description sets are authored by Gemma4; diagnostic views and joint descriptions are fixed. Full-image authoring uses the full query and full normal reference instead of the DINO-marked authoring views. All paired differences have $p \geq 0.01$ (exact McNemar test).

Diagnosing model	Condition	DINO views	Full images	p
Gemma4	Region observations	38/90	39/90	≥ 0.01
Gemma4	Joint + region	42/90	43/90	≥ 0.01
Qwen3.6	Region observations	39/90	30/90	≥ 0.01
Qwen3.6	Joint + region	41/90	37/90	≥ 0.01

Table 12: **Authoring-model sensitivity.** Correct diagnoses over 90 defects; gains in percentage points. Primary author supplies the descriptions in Table 8.

Diagnosing model	Condition	Primary author (Gemma-4-31B)	Authored by Gemma4	Authored by Qwen3.6
Gemma4	Names	14	16	16
Gemma4	Joint descriptions	**50**	39	29
Gemma4	Region observations	45	38	48
Gemma4	Joint + region	**51**	42	34
Gemma4	Joint − names (pp)	+40.0	+25.6	+14.4
Qwen3.6	Names	4	3	3
Qwen3.6	Joint descriptions	**41**	37	14
Qwen3.6	Region observations	39	39	37
Qwen3.6	Joint + region	**43**	41	20
Qwen3.6	Joint − names (pp)	+41.1	+37.8	+12.2

Authoring-model comparisons. The Names condition is rerun without descriptions in each authoring experiment, using identical prompts and image hashes; its differing counts reflect separate diagnostic runs. Relative to the primary-author descriptions in Table 12, Gemma4's accuracy decreases with its own joint descriptions. With Qwen3.6-authored descriptions, both joint and joint-plus-region accuracy decrease for both diagnostic models. These decreases were all statistically significant ($p < 0.01$).

Gemma4's joint-plus-region decrease with its own descriptions and the remaining comparisons in Table 12 have $p \geq 0.01$.

Authoring-view sensitivity. Table 11 compares region observations written from DINO-marked views with observations written from full query and normal-reference images. Gemma4 authors both sets, joint descriptions remain fixed, and both diagnostic models retain identical DINO four-view inputs.

Table 13: **Regenerating the public four-view study with the disjoint normal bank.** Type accuracy (%) with the retrospective bank of Table 3 and with the reference and calibration normals of the single-image study, which are disjoint from evaluation. All paired differences have $p \geq 0.01$.

Dataset	Model	Knowledge	Retrospective	Disjoint	Δ
MVTec AD	Gemma4	Names	46.7	48.1	+1.4
		Definitions	53.0	53.3	+0.3
		Region	67.3	66.2	−1.1
		Combined	66.5	67.6	+1.1
MVTec AD	Qwen3.6	Names	38.7	37.5	−1.1
		Definitions	44.1	45.8	+1.7
		Region	55.0	59.0	+4.0
		Combined	55.6	60.2	+4.6
Real-IAD	Gemma4	Names	55.0	54.7	−0.3
		Definitions	52.3	54.7	+2.3
		Region	60.7	61.0	+0.3
		Combined	58.7	58.7	+0.0
Real-IAD	Qwen3.6	Names	38.3	40.3	+2.0
		Definitions	40.7	43.0	+2.3
		Region	50.3	50.3	+0.0
		Combined	50.0	50.3	+0.3

Differences are computed before rounding. Gate comparisons in Table 3 use the disjoint runs. Repeat variation is measured only on the MVTec combined-condition subset in Table 16.

Table 14: **Frozen-feature baselines on the public benchmarks.** Type accuracy (%). DINO 1-NN and prototype classifiers use the same held-out defect images (two per type) that the descriptions were written from, CLS or mean-patch features, and must assign a type to every image. Qwen3.6 with combined descriptions leads the best variant on both datasets ($p \geq 0.01$). Gemma4 with combined descriptions exceeds every DINO variant, while Qwen3.6 with names alone falls below the MVTec prototype. These latter comparisons were all statistically significant ($p < 0.01$).

Dataset	1-NN CLS	Prototype CLS	1-NN mean patch	Prototype mean patch	Gemma4 names / combined	Qwen3.6 names / combined
MVTec AD	48.42	51.86	45.85	47.56	46.70 / **66.48**	38.68 / 55.59
Real-IAD	44.33	41.67	44.67	46.33	55.00 / **58.67**	38.33 / 50.00

Table 15: **Post-hoc sweep of the semiconductor detector threshold.** The pre-specified rule is the maximum of the four calibration-normal scores per fold; alternatives use a linearly interpolated quantile q of the same four scores. Fold-wise detection AUROC (30 defects versus 4 evaluation normals) is 1.000, 0.967, and 1.000. Correct types are over 90 defects under joint-description knowledge, matching Section 5.3; the last two columns are normal false positives after gating.

Threshold rule	Thr. (fold 2)	Defects passed	Normals passed	Gemma4 correct	Qwen3.6 correct	Gemma4 FP	Qwen3.6 FP
Maximum (pre-specified)	0.536	66/90	3/12	38/90	31/90	1	0
$q = .80$	0.390	87/90	3/12	49/90	41/90	1	0
$q = .90$	0.463	82/90	3/12	46/90	38/90	1	0
$q = .95$	0.499	73/90	3/12	42/90	35/90	1	0
No gate	–	90/90	12/12	50/90	41/90	7	0

Fold 0 and fold 1 thresholds are almost rule-independent (0.271–0.276 and 0.306–0.308); only fold 2, whose calibration scores are 0.207, 0.232, 0.293, and 0.536, separates the rules, and the maximum rule passes 8 of its 30 defects. Under joint-plus-region knowledge the maximum rule gives 39/90 and 32/90.

Table 16: **Run-to-run variation of the hosted public runs.** The MVTec AD combined condition repeated three times on the first 100 cases (50 defects, 50 normals) at temperature zero, against the original run. Type accuracy in %.

Model	Original	Rep. 1	Rep. 2	Rep. 3	sd (pp)	Defects with identical answers
Gemma4	62.0	62.0	62.0	62.0	0.00	49/50
Qwen3.6	42.0	48.0	46.0	48.0	2.45	41/50

SD is computed over the four accuracies (original plus three repeats), with divisor four (ddof=0), on the same 50 defects. Identical answers means agreement across all four runs.

B.4 BANK, BASELINE, GATE, AND REPEAT CONTROLS

Table 13 regenerates the public four-view study with the disjoint normal bank of the single-image study. The disjoint-bank panel of Table 3 reports combined-description accuracy and normal FPR before and after applying the calibration gate offline to the same predictions. Table 14 builds the semiconductor feature baselines on the public benchmarks. Table 15 sweeps the semiconductor detector threshold, and Table 16 repeats the hosted public runs. All VLM diagnosis rows use the two frozen diagnostic models of the main text; semiconductor rows are served locally and public rows through the hosted API. The repeat statistics describe only the MVTec combined-condition subset; they do not estimate variation on other cohorts or conditions. Gemma4's zero accuracy standard deviation does not imply identical predictions: one defect receives different answers across runs.

B.5 DETAILED COMPARISONS UNDERLYING THE RESULTS NARRATIVE

Error transitions. In the public MVTec four-view names-to-combined comparison, Gemma4 corrects 75 wrong-type predictions and recovers 6 defects predicted as NORMAL; Qwen3.6 corrects 9 wrong types and recovers 54 missed defects. In the single-image unmarked-to-DINO-marked comparison, with generic criteria and the gate fixed, missed-defect recoveries are 49, 46, 34, and 26, versus 5, 1, 2, and 1 wrong-type corrections (MVTec/Gemma4, MVTec/Qwen3.6, Real-IAD/Gemma4, Real-IAD/Qwen3.6). These are corrected-case counts before subtracting reverse transitions, not net accuracy gains. Combined descriptions improve over names in 18 of 20 public product/model comparisons.

Modality controls and semiconductor comparisons. In Table 4, adding images to descriptions decreases accuracy in three settings ($p < 0.01$); the MVTec/Qwen3.6 increase has $p \geq 0.01$. On the semiconductor cohort, Qwen3.6 joint descriptions and DINO 1-NN each correctly classify 41/90 defects, with 21 discordant cases in each direction. Qwen3.6 predicts NORMAL for 39/90 defects, giving 56.7% recall versus Gemma4's 93.3%. The retrospective strip-position lookup correctly classifies 38/90 defects; Table 10 reports diagnosis on the remaining cases and after excluding near-held-out candidates.