# EviInspect: Evidence-Grounded Annotation and Evaluation for Safety-Critical Industrial Inspection

Nikolaos Marios Militsis
nikos.militsis@iti.gr

Achilleas Toumpas
atoumpas@iti.gr

Ilias Koulalis
iliask@iti.gr

Konstantinos Ioannidis
kioannid@iti.gr

Stefanos Vrochidis
stefanos@iti.gr

Centre for Research and Technology Hellas

## Abstract

*Vision-based inspection systems increasingly support safety-critical decision-making in industrial settings. Their reliability depends not only on predictive accuracy, but also on whether visual evidence is interpreted in a manner consistent with domain-specific rules. In many inspection tasks, labels such as hazard level are not intrinsic visual properties, instead, they are defined through external domain references, including standards and empirical studies, that connect observable evidence to downstream operational consequences. However, many existing vision and multimodal reasoning benchmarks implicitly treat such labels as directly observable visual ground truth. As Vision Language Models (VLMs) are increasingly used in inspection pipelines, datasets are needed to evaluate not only prediction accuracy, but also whether decisions are supported by sufficient visual evidence under domain-consistent rules. To address this need, we introduce* ***EviInspect****, an evidence-guided annotation framework in which inspection labels are derived by explicitly linking visual evidence to external domain references. EviInspect combines AI-assisted evidence extraction with human verification to assess whether the available evidence is sufficient to support an assigned label and whether the decision is reproducible given the same evidence. We demonstrate the framework in a Foreign Object Debris (FOD) inspection use case and release* ***FOD-A-H****, an evidence-grounded extension of the public FOD-A dataset, with hazard and size annotations derived from Federal Aviation Administration (FAA) reference material. Using FOD-A-H, we evaluate state-of-the-art VLMs on their ability to derive inspection labels under explicit evidence constraints.*

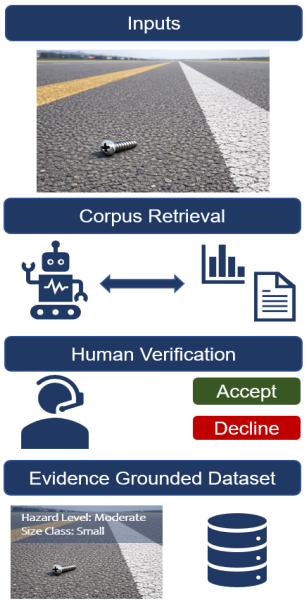



Figure 1. Conceptual overview of EviInspect. Given a runway image and a trusted FOD-A object class, the framework retrieves relevant reference material, records supporting evidence in structured form, and derives hazard and size labels only when evidence sufficiency and decision reproducibility are established under human verification.

## 1. Introduction

Vision-based industrial inspection systems operate within complex workflows in which data quality, domain knowledge, and human decision-making jointly determine overall reliability [16]. In many such settings, inspection labels including defect severity, hazard level, and operational risk cannot be read directly from the image. Instead, they must be derived by applying external domain references, such as standards, technical reports, and empirical studies, to ob-

servable visual evidence. As a result, the correctness of an inspection decision often depends not only on image content, but also on whether the available evidence is interpreted in accordance with domain-specific rules.

This distinction becomes increasingly important as VLMs are introduced into inspection pipelines. Recent multimodal reasoning benchmarks emphasize structured intermediate reasoning and explanation generation [6, 10], but they often assume that the target label is visually well-defined and can therefore be treated as image-grounded ground truth. In safety-critical inspection, however, labels such as hazard level or size class may require applying external domain references to visual cues rather than reading the label directly from the image. Consequently, a model may produce a coherent-looking explanation while still arriving at an insufficiently supported or unsafe inspection decision.

These challenges also affect dataset construction. Industrial workflows often rely on domain experts to define guidelines that are then applied by non-expert annotators. While effective for visual categorization, this strategy becomes less reliable when annotations must encode domain-defined semantics such as severity or risk. In such cases, annotators may overlook relevant evidence, apply rules inconsistently, or rely on subjective judgment, especially when the evidence supporting a label is not recorded explicitly. Recent industrial inspection datasets [1, 2, 4, 17] provide valuable resources for benchmarking and model development, but the external references used in real inspection practice are rarely represented as part of the annotation process itself.

To address this gap, we introduce **EviInspect**, an evidence-guided annotation framework in which inspection labels are derived by explicitly linking visual evidence to external domain references. Rather than assigning labels directly from images, EviInspect structures annotation as a staged procedure: relevant reference material is retrieved and summarized into an explicit evidence record, after which labels are derived only when the available evidence is sufficient and the decision is reproducible under human verification. Figure 1 provides an overview of this process.

We instantiate EviInspect in a safety-critical FOD inspection setting. Specifically, we extend the FOD-A dataset [19] with evidence-grounded annotations for hazard level and size class derived from FAA reference material [9], yielding **FOD-A-H**. In addition to supporting dataset construction, FOD-A-H enables evaluation of VLMs under evidence-constrained inspection settings, where success depends not only on producing a label, but on deriving that label from sufficient, domain-consistent evidence.

The main contributions of this paper are as follows:

- **Evidence-guided annotation framework:** We propose EviInspect, a framework for deriving inspection labels by explicitly linking visual evidence to external domain references.
- **Evidence sufficiency and reproducibility criteria:** We introduce verification criteria that assess whether recorded evidence adequately supports a proposed annotation and whether the decision is reproducible given the same evidence.
- **Evidence-grounded dataset and benchmark:** We release **FOD-A-H**, an evidence-grounded extension of FOD-A with hazard and size annotations derived from FAA reference material, and use it to evaluate VLMs in evidence-constrained inspection settings.

## 2. Related Work

**Benchmarks for Multimodal Reasoning and Safety Evaluation**

Recent benchmarks evaluate multimodal models under structured reasoning, safety, and trust-oriented settings that go beyond single-image perception. For example, MMRB [6] evaluates reasoning across multiple images, while MultiTrust [35] and iSafetyBench [1] assess multimodal trustworthiness and industrial safety understanding. Related efforts such as VLDBench [22] further construct large-scale multimodal datasets for safety-critical applications using semi-automated pipelines with expert review. While these benchmarks advance reasoning evaluation, they place less emphasis on whether labels are grounded in sufficient visual evidence under domain-specific constraints. Our work addresses this gap by centering the annotation process itself and incorporating domain references with human verification.

**Reasoning-Centric LVLM Evaluation and Evidence Grounding**

Recent Vision Language Reasoning Models (VLRMs) show that reinforcement-based post-training can improve multimodal reasoning and encourage explicit intermediate steps. Models such as Vision-R1 [33], MMEureka [15], Ocean-R1 [18], ThinkLite-VL [29], and OpenVLThinker [7] report improved performance on reasoning benchmarks, while complementary approaches incorporate perception-aware signals to strengthen grounding [24, 31]. Other work extends grounding beyond object identity and spatial reference toward richer concepts such as affordances, function, safety, and physical reasoning [23]. However, recent analyses continue to highlight failures in perceptual grounding, especially in semantically rich or multi-image settings. EViSRAG shows that misinterpretation of visual evidence can undermine otherwise coherent reasoning [25]. Our work complements these model-centric advances by focusing on the data and annotation substrate required for evidence-grounded evaluation in safety-critical inspection.

**Vision-Based FOD Inspection**

FOD detection on airport runways is a safety-critical vision task, with prior work focusing mainly on detection and localization performance. FOD-A [19] serves as a widely used public benchmark, and subsequent work has explored detector-based and self-supervised approaches for runway FOD analysis [11, 13, 14, 20, 27]. More recent efforts move toward richer multimodal formulations, including image-text pairing and LLM-assisted FOD analysis [5, 12]. However, existing approaches do not explicitly address hazard severity and operational risk through evidence-grounded annotation, which motivates our extension of FOD-A with FAA-supported hazard and size labels.

# 3. EviInspect: Evidence-Guided Annotation for Industrial Inspection

## 3.1. Problem Formulation

In many industrial inspection tasks, target labels such as severity, hazard level, or operational risk cannot be assigned from image content alone, but must be derived by applying external domain references to observed visual evidence. We therefore formulate annotation as an evidence-guided process in which labels are derived not directly from the image, but from a chain linking image evidence to trusted external references.

Let $I$ denote an inspection image, $c$ an optional trusted semantic anchor such as a ground-truth object class, and $D$ an external reference corpus such as standards, technical reports, or operational manuals. EviInspect constructs a structured evidence record $E$ from observations in the inspection image together with retrieved reference material, and derives inspection labels $y$ only when the available evidence is sufficient:

$$(I, c, D) \rightarrow D_R \rightarrow E \rightarrow y,$$

where $D_R \subset D$ is the retrieved evidence subset. With this formulation, annotation validity depends on both image content and the adequacy of the supporting evidence chain.

## 3.2. Evidence-Guided Annotation Pipeline

EviInspect consists of four stages, as shown in Figure 2.

**Stage 1: Reference Retrieval.** Relevant reference material is retrieved from an external reference corpus consisting of task-relevant standards, technical reports, manuals, or empirical studies. Retrieval uses the trusted semantic anchor when available and otherwise relies on directly observable visual cues from the inspection image. The goal is to assemble candidate supporting evidence rather than assign labels directly. If no sufficiently relevant material is retrieved, the sample is marked as unsupported. Human verification then determines whether this reflects limited corpus coverage or retrieval failure. In the former case, the sample is excluded, while in the latter, it is returned for further retrieval and evidence inspection.

**Stage 2: Evidence Inspection.** Retrieved material is inspected to identify content relevant to the target instance, such as item correspondences, rule definitions, decision criteria, or measured examples. Retrieved documents are treated as visual evidence sources so that layout, tables, and figures remain available during annotation.

**Stage 3: Structured Evidence Recording.** Potentially relevant observations from the inspection image together with the retrieved reference material are summarized into a structured evidence record

$$E = \Phi(I, c, D_R),$$

where $\Phi(\cdot)$ denotes evidence extraction and structuring. In our workflow, the evidence record follows the structured evidence-recording format used in EVisRAG-style reasoning over retrieved visual documents, in which the system first inspects retrieved visual documents, records per-document evidence, and then reasons from the aggregated evidence to derive the final output [25]. In our setting, this record captures image observations together with rule-relevant information extracted from the retrieved references.

**Stage 4: Label Derivation.** Final labels are derived from the structured evidence record:

$$y = G(E),$$

where $G(\cdot)$ denotes the evidence-to-label mapping defined by the referenced domain rules. In our workflow, AI assistance may support earlier stages of retrieval and evidence creation, but final labels are assigned only under human verification when the recorded evidence supports a sufficiently specific and reproducible decision.

## 3.3. Evidence Sufficiency and Human Verification

A central property of EviInspect is that annotation depends not only on the final label, but also on whether the supporting evidence is adequate. We therefore evaluate candidate annotations along two dimensions:

1. **Evidence sufficiency**: whether the recorded evidence provides specific and relevant support for the proposed label;
2. **Decision reproducibility**: whether another annotator given the same retrieved material and evidence record would likely reach the same conclusion.

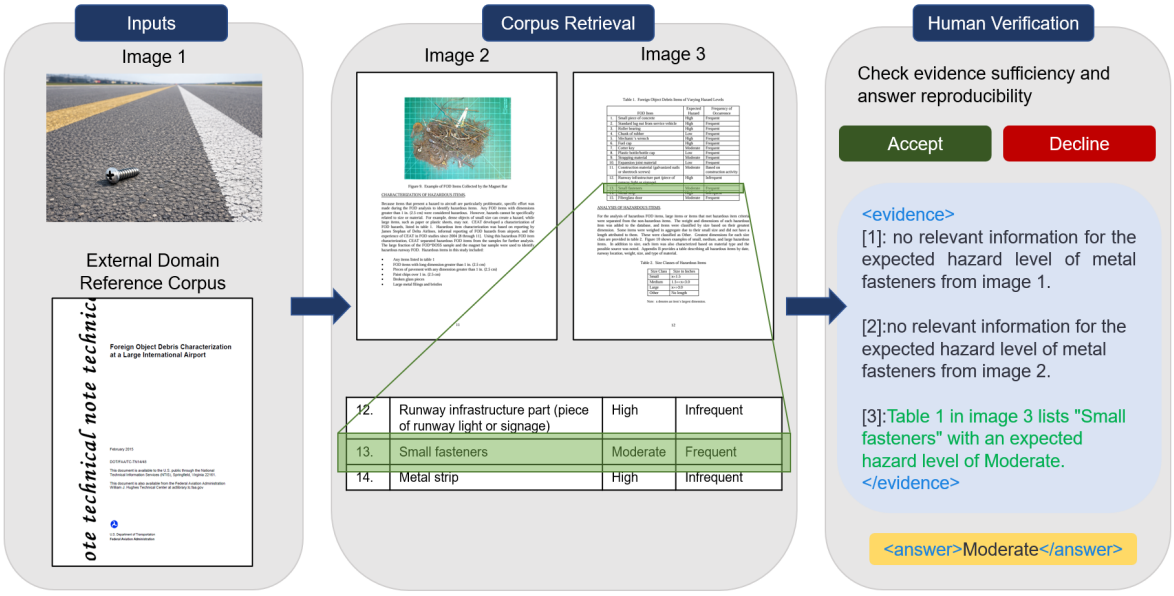



Figure 2. Operational view of EviInspect. Given an inspection image and an optional trusted semantic anchor such as a ground-truth object class, relevant reference material is retrieved and inspected as visual evidence. Candidate evidence is organized into a structured evidence record, from which labels are derived only when evidence sufficiency is established and the annotation is judged reproducible under human verification.

In our workflow, candidate annotations are verified by human annotators using the structured evidence record rather than assigning labels from scratch or searching the reference corpus manually. This supports a more efficient and consistent verification process, particularly for non-expert annotators. A sample is accepted only when the evidence chain is judged sufficiently specific and the resulting decision reproducible under verification. Otherwise, the sample is first marked as unsupported. Under further review, it may be excluded when the lack of support reflects limited corpus coverage, poor image quality, or insufficient evidence for a reproducible decision. The frequency of such cases depends on factors such as reference quality, corpus coverage, image quality, and task difficulty.

### 3.4. Role of AI Assistance

EviInspect supports scalable annotation through AI-assisted evidence inspection and evidence recording. In our application, this workflow is inspired by EVisRAG-style [25] reasoning over retrieved visual documents, where reference pages are treated as visual evidence rather than flattened text. We additionally use Rel-Attn [34] as an auxiliary visualization tool, applying the published method to retrieved reference pages to highlight candidate supporting regions for annotator review; these highlights are used only to speed evidence inspection and are not treated as evidence themselves. Final labels are determined only from the structured evidence record under human verification.

## 4. Experiments and Results

### 4.1. FOD-A-H Benchmark Construction

We instantiate EviInspect in a safety-critical FOD inspection setting by extending FOD-A [19] with evidence-grounded annotations for *hazard level* and *size class*. This setting serves as a concrete instantiation of the framework in a domain where inspection labels are defined through external reference material. While EviInspect is formulated in a domain-agnostic manner, we focus on this single use case to study evidence-grounded annotation and evaluation under controlled conditions. FOD-A provides runway images together with trusted object-class annotations, which we use as the semantic anchor for evidence retrieval and downstream label derivation. As the external reference corpus, we use the FAA technical report *Foreign Object Debris Characterization at a Large International Airport* [9].

FOD-A-H is constructed by conservatively remapping FOD-A classes to FAA-supported concepts and assigning labels only when both hazard level and size class can be justified through an explicit evidence chain under EviInspect and verified by human annotators for evidence sufficiency and decision reproducibility. The starting point is the full FOD-A collection, which contains 33,793 annotated images, split into 25,345 training/validation images and 8,448 test images. We relate the original 31 FOD-A object classes to 15 FAA-aligned target concepts used for hazard-aware characterization (Figure 3). This normalization step is in-

tentionally conservative: only correspondences supported by the FAA report and verified under our framework are retained, with the remaining 18 classes excluded because they are not represented in the reference corpus and therefore cannot support evidence-grounded label derivation. Examples of excluded classes include *Battery*, *Cutter*, *Pliers*, *Pen*, and *SodaCan*. In the final mapping, each retained original FOD-A class is assigned to a single FAA-aligned target concept, so that each image contributes to only one normalized concept label.

In the current construction of FOD-A-H, hazard and size labels are largely determined at the concept level based on FAA-supported mappings, resulting in limited instance-level variation. While this simplifies the annotation setting, the benchmark still evaluates evidence-grounded behavior through retrieval alignment and abstention under insufficient evidence. Extending the framework to settings with stronger instance-level variation is an important direction for future work.

Under this evidence-grounded remapping, FOD-A-H contains 14,355 images in total, with 10,754 images in training/validation and 3,601 images in test. Within this retained subset, 13 original FOD-A classes map to 8 of the 15 FAA-aligned target concepts. The conservative remapping is shown in Figure 3.

Hazard labels follow the FAA categories *High*, *Moderate*, and *Low* [9]. Size labels follow the FAA taxonomy based on greatest dimension: *Small* for $x < 1.5$ in, *Medium* for $1.5 \leq x < 3.0$ in, *Large* for $x \geq 3.0$ in, and *Other* when reliable length-based evidence is unavailable [9]. To assign size labels, we use Appendix B of the FAA report as dimensional support and estimate a representative size for each retained concept from the matched hazardous-item records.

To assess annotation consistency, we performed human verification on a random 10% subset of retained samples. Two annotators were given the image, matched FAA evidence, structured evidence record, and proposed hazard and size labels, and used a structured checklist to judge whether the assignment was sufficiently supported. Inter-annotator agreement was 83.2% (raw agreement) across hazard and size labels. Agreement was higher for hazard labels than for size labels, reflecting the more explicit categorical definitions in the FAA reference material. Disagreements primarily occurred in borderline cases where the available evidence was insufficiently specific, particularly when distinguishing between adjacent size categories or when dimensional evidence was incomplete. In such cases, annotators followed a conservative strategy, marking samples as unsupported when evidence sufficiency could not be established. While the verification was conducted on a 10% subset with two annotators, this process was designed as a consistency check on evidence-grounded label derivation rather than a full independent annotation study. The use

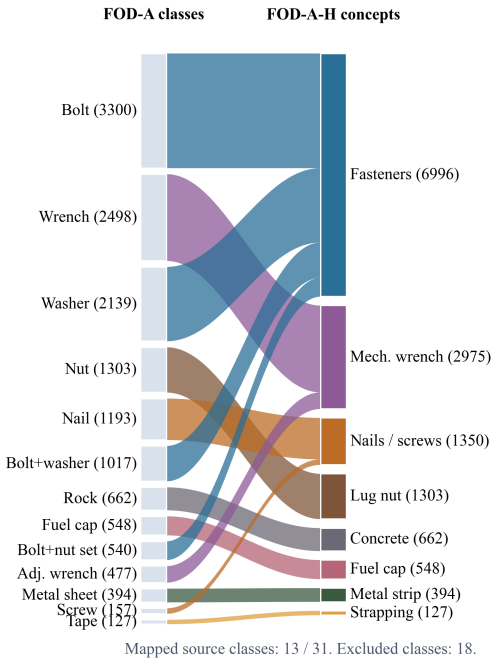



Figure 3. Conservative remapping from original FOD-A classes to retained FOD-A-H concepts. Only source classes with sufficiently supported FAA correspondences are preserved. Unsupported classes are excluded.

of structured evidence records further reduces ambiguity by constraining decisions to explicitly documented evidence.

## 4.2. Dataset Characterization

Figure 4 summarizes the image-level distribution of retained FOD-A-H concepts after conservative evidence-grounded remapping. The resulting benchmark exhibits a strongly long-tailed concept distribution: Small fasteners is the dominant retained category, accounting for 48.7% of FOD-A-H, while several other concepts are substantially rarer. We preserve this imbalance because it reflects the evidence-supported distribution of retained instances in the source data rather than an artificially balanced construction.

FOD-A-H is also imbalanced at the derived label level. Hazard labels are concentrated in the Moderate and High categories, with no retained Low instances under the current evidence-supported mapping. Size labels are dominated by Small, with fewer Medium and Large instances and a substantial number of Other assignments when reliable dimensional evidence is unavailable.

These statistics are important for interpreting benchmark performance. The dataset is imbalanced both at the concept level and at the hazard/size label level, which motivates reporting macro-F1 in addition to Accuracy. The presence of the Other size label further reflects a central property of FOD-A-H: when evidence is insufficient for a reproducible

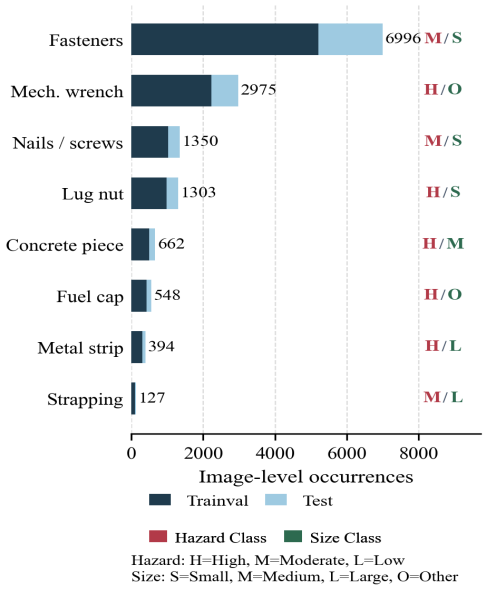



Figure 4. Image-level distribution of retained FOD-A-H concepts after conservative evidence-grounded remapping from FOD-A. Bars show trainval and test counts per retained concept. The right-side tags indicate the associated hazard and size labels for each concept under the FAA-aligned mapping.

dimensional assignment, the benchmark preserves that uncertainty instead of replacing it with a heuristic visual estimate.

### 4.3. Evaluation Settings

We evaluate FOD-A-H under the two settings.

**Class-conditioned grounding:** The model receives the runway image together with the trusted FOD-A object class and must retrieve FAA evidence to derive hazard level and size class. This setting isolates evidence-grounded label derivation from errors in object-class prediction.

**End-to-end grounding:** The model receives only the runway image and must first infer the object class, then retrieve the relevant FAA evidence, and finally predict hazard and size. This setting evaluates the full pipeline, including recognition, retrieval, evidence grounding, and final label derivation.

### 4.4. Retrieval and Sufficiency Protocol

To evaluate VLMs on evidence-grounded label derivation under controlled retrieval conditions, all baselines use the same retrieval backend [32], to ensure that performance differences primarily reflect grounding and reasoning rather than retriever choice. Following recent visual-RAG practice [8, 28, 32], we split the FAA report into page-level images and use these page snapshots as retrieval units [25]. For each test sample, the retriever returns the top-3 FAA page images, which are provided to the model together with the runway image and task prompt. We use VisRAG-Ret [32] as the shared retriever in all experiments.

During dataset construction, we record for each sample the FAA page-image IDs that serve as the required supporting evidence. At evaluation time, we compare these IDs against the retrieved top-$k$ page images. Following prior sufficient/insufficient evidence evaluation [25], we mark a query-context pair as *sufficient* if the retrieved set contains the required supporting page image(s), and *insufficient* otherwise. For insufficient cases, the correct output is *insufficient to answer* rather than a hazard or size label.

Using this protocol, we compare VLMs on more than final answer accuracy. Correct behavior requires producing the correct label when the relevant FAA evidence is retrieved and abstaining when retrieval does not provide the necessary support. Our approach and benchmark therefore evaluate both evidence-grounded prediction and conservative behavior under incomplete retrieval.

### 4.5. Baselines

We compare three groups of baselines spanning general-purpose multimodal models, reasoning-oriented models, and retrieval-augmented models.

**General VLMs:** We include strong general-purpose vision-language models that provide a reference point for multimodal understanding without explicit specialization for retrieval-grounded reasoning: Qwen2.5-VL-7B, Qwen2.5-VL-32B [3], and MiMo-VL-7B-RL [26].

**Vision-language reasoning models (VLRMs):** We include recent reasoning-oriented multimodal models designed for more explicit or extended visual reasoning: Vision-R1 [33], Ocean-R1-7B [18], MM-Eureka [15], ThinkLite-VL-7B [29], and OpenVthinker [7].

**Visual retrieval-augmented generation models (VRAGs):** We include models explicitly designed for retrieval-grounded multimodal generation and reasoning: MMSearch-R1 [30], VRAG-RL [28], and R1-Router [21]. All built upon the Qwen2.5-VL-7B-Instruct architecture [3].

Together, these baselines enable comparison of evidence-grounded label derivation across models with different capabilities, from strong general-purpose VLMs to reasoning-specialized and retrieval-augmented systems. All baselines use the same FAA page-image corpus and the same VisRAG-Ret retriever [32], so differences primarily reflect grounding and reasoning rather than retrieval backend choice. All inferences were conducted on a single NVIDIA RTX PRO 6000 Blackwell GPU, using a unified prompting setup based on the EVisRAG prompt from the original EVisRAG paper [25] unless otherwise specified.

| | Class-Conditioned Grounding | | | | | | End-to-End Grounding | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hazard Level | | Size | | Average | | Hazard Level | | Size | | Average | |
| **Model** | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| **General VLMs** | | | | | | | | | | | | |
| Qwen2.5-VL-7B | 59.75 | 53.21 | 60.73 | 54.22 | 60.24 | 53.72 | 54.19 | 48.19 | 53.81 | 49.70 | 54.00 | 48.95 |
| MiMo-VL-7B-RL | 54.92 | 40.72 | 68.76 | 45.94 | 61.84 | 43.33 | 50.51 | 33.94 | 62.67 | 41.69 | 56.59 | 37.81 |
| Qwen2.5-VL-32B | **69.78** | 61.35 | **78.63** | 66.70 | **74.20** | 64.03 | **65.23** | 54.70 | **71.14** | 60.28 | **68.19** | 57.49 |
| **VLRMs** | | | | | | | | | | | | |
| Vision-R1 | 56.12 | 50.20 | 29.42 | 27.91 | 42.77 | 39.05 | 48.55 | 42.37 | 24.03 | 20.58 | 36.29 | 31.47 |
| Ocean-R1-7B | 47.30 | 48.07 | 53.28 | 52.71 | 50.29 | 50.39 | 40.70 | 43.08 | 46.16 | 46.37 | 43.43 | 44.72 |
| MM-Eureka | 63.52 | 58.27 | 41.12 | 39.91 | 52.32 | 49.09 | 58.89 | 50.92 | 37.16 | 34.07 | 48.02 | 42.50 |
| ThinkLite-VL-7B | 57.18 | 54.28 | 62.19 | 61.54 | 59.69 | 57.91 | 53.78 | 49.68 | 56.65 | 53.88 | 55.22 | 51.78 |
| OpenVthinker | 66.97 | **62.80** | 70.80 | **70.59** | 68.88 | **66.69** | 59.90 | **57.10** | 62.98 | **64.57** | 61.44 | **60.83** |
| **VRAGs** | | | | | | | | | | | | |
| MMSearch-R1 | 63.42 | 59.80 | 58.09 | 57.53 | 60.76 | 58.66 | 57.54 | 55.35 | 54.14 | 53.60 | 55.84 | 54.48 |
| VRAG-RL | 47.18 | 10.28 | 64.82 | 11.55 | 56.00 | 10.91 | 40.39 | 2.90 | 57.20 | 4.34 | 48.80 | 3.62 |
| R1-Router | 61.36 | 16.21 | 60.64 | 15.00 | 61.00 | 15.61 | 54.83 | 11.83 | 53.58 | 7.75 | 54.20 | 9.79 |

Table 1. Performance comparison under class-conditioned grounding and end-to-end grounding. Best results in each column are shown in **bold**. Average columns report the mean across Hazard Level and Size.

### 4.6. Metrics

We report Accuracy and macro-F1 for hazard level, size class, and their average across the two inspection labels. We use macro-F1 because FOD-A-H is imbalanced both at the retained-concept level and at the derived label level, so balanced performance across categories is important for evidence-grounded inspection.

Under the sufficient/insufficient protocol, the gold target is the original label when retrieval is sufficient and *insufficient to answer* otherwise. The reported metrics therefore jointly measure label prediction under sufficient evidence and abstention behavior under insufficient evidence.

### 4.7. Overall Performance

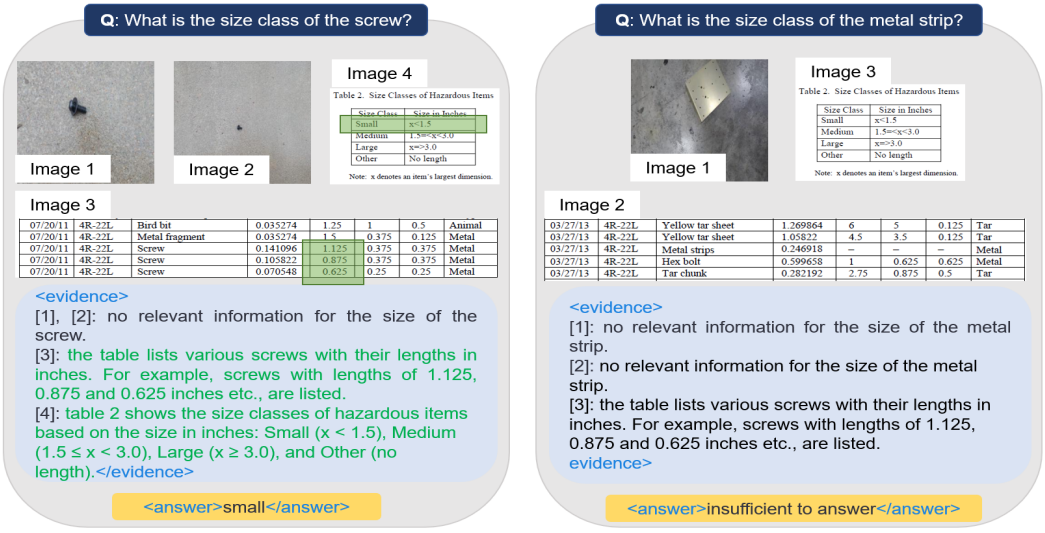
Table 1 evaluates both prediction accuracy and model behavior under evidence constraints. Under the sufficient/insufficient protocol, correct performance requires predicting the correct label when relevant FAA evidence is retrieved and abstaining when it is not, enabling analysis of evidence-aware decision-making. Class-conditioned grounding consistently outperforms end-to-end grounding, indicating that a substantial portion of the difficulty lies in object recognition prior to evidence grounding. However, the remaining gap shows that even with correct class information, aligning predictions with retrieved evidence remains non-trivial. FOD-A-H is challenging across all models. *OpenVthinker* and *Qwen2.5-VL-32B* achieve the strongest performance, but both degrade in the end-to-end setting, highlighting the difficulty of jointly solving recognition, retrieval, and evidence-grounded prediction.

**Model family behavior.** General-purpose VLMs provide the most stable performance, with relatively balanced Accuracy and macro-F1, suggesting reliance on learned visual and semantic priors. Reasoning-oriented models (VLRMs) show mixed results, indicating that improved reasoning does not consistently translate to better alignment with external domain references. Retrieval-augmented models (VRAGs) are the least stable: although retrieval provides access to relevant FAA material, these models often fail to correctly use it and do not reliably adjust predictions when evidence is insufficient.

**Failure modes.** A key pattern is the gap between Accuracy and macro-F1, especially for VRAG-based models. Moderate Accuracy with low macro-F1 indicates inconsistent performance across classes. In insufficient-evidence cases, many models predict labels instead of abstaining, effectively treating the absence of evidence as evidence of presence. Stronger models show better calibration between prediction and abstention, but still fail when retrieved evidence is only partially relevant or lacks sufficient specificity.

**Error sources.** Errors arise from multiple stages: object recognition failures affect retrieval, retrieved evidence may be misinterpreted, and models often fail to recognize when evidence is insufficient. Hazard prediction primarily requires alignment with FAA categories, while size prediction requires interpreting dimensional ranges, making it more sensitive to incomplete or ambiguous evidence.

Overall, success on FOD-A-H requires correct recog-




(a) Sufficient evidence.

(b) Insufficient evidence.

Figure 5. Qualitative examples of evidence-guided size grounding in FOD-A-H. **(a)** two screw images at different visual scales are assigned the same size label (*Small*) because the FAA size taxonomy provides consistent dimensional support. **(b)** for the metal strip, the retrieved material is relevant but does not provide sufficiently specific dimensional evidence, so the correct outcome is `Insufficient`.

nition, retrieval, evidence alignment, and conservative behavior under insufficient evidence, highlighting current limitations of VLMs in safety-critical, evidence-grounded decision-making.

### 4.8. Qualitative Evidence-Grounding Results

To complement the quantitative results, we provide qualitative examples that illustrate the evidence-grounded decision patterns evaluated by FOD-A-H. Figure 5 shows two size-grounding examples from FOD-A-H. In the first case, two screw images at different visual scales receive the same size label because the assignment is supported by matched FAA records and the FAA size taxonomy rather than image scale alone. In the second case, the retrieved material is relevant to the depicted metal strip but does not provide sufficiently specific dimensional support for a reproducible size assignment, so the correct outcome is `insufficient to answer`.

## 5. Conclusion

We introduced **EviInspect**, an evidence-guided annotation framework for industrial inspection settings in which labels must be derived from visual evidence interpreted through external domain references. By structuring annotation around reference retrieval, evidence recording, and human verification of evidence sufficiency, EviInspect supports transparent and reproducible label construction for safety-relevant tasks.

We instantiated the framework in airport FOD inspection and introduced **FOD-A-H**, an evidence-grounded extension of FOD-A with hazard and size annotations derived from FAA reference material. Experiments with state-of-the-art VLMs show that FOD-A-H remains challenging under both class-conditioned and end-to-end grounding, highlighting the combined difficulty of recognition, retrieval, and evidence-grounded decision-making.

More broadly, our results suggest that progress in vision-based industrial inspection should be evaluated not only by prediction accuracy, but also by whether decisions are grounded in sufficient evidence under domain-specific rules, enabling safer and more auditable inspection workflows.

While EviInspect is formulated in a domain-agnostic manner, we evaluate it here in a single FOD inspection setting with one FAA reference corpus. FOD-A-H also uses concept-level hazard and size mappings, resulting in limited instance-level variation, but still evaluates evidence-grounded behavior through retrieval alignment and abstention under insufficient evidence. Extending to broader domains and stronger instance-level variation is an important direction for future work, along with studying how evidence-grounded supervision can improve model development and human-AI inspection pipelines.

## Acknowledgements

This work was supported by SEISMEC project funded by the European Commission under grant agreements No 101135884.

## References


[1] Raiyaan Abdullah, Yogesh Singh Rawat, and Shruti Vyas. isafetybench: A video-language benchmark for safety in industrial environment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1433–1442, 2025. 2

[2] Haoping Bai, Shancong Mou, Tatiana Likhomanenko, Ramazan Gokberk Cinbis, Oncel Tuzel, Ping Huang, Jiulong Shan, Jianjun Shi, and Meng Cao. Vision datasets: A benchmark for vision-based industrial inspection. *arXiv preprint arXiv:2306.07890*, 2023. 2

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 6

[4] Aaditya Baranwal, Abdul Mueez, Jason Voelker, Guneet Bhatia, and Shruti Vyas. Synspill: Improved industrial spill detection with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1414–1423, 2025. 2

[5] Hanglin Cheng, Ruoxi Zhang, Ruiheng Zhang, Yihao Li, Yang Lei, and Weiguang Zhang. Intelligent detection and description of foreign object debris on airport pavements via enhanced yolov7 and gpt-based prompt engineering. *Sensors*, 25(16):5116, 2025. 3

[6] Ziming Cheng, Binrui Xu, Lisheng Gong, Zuhe Song, Tianshuo Zhou, Shiqi Zhong, Siyu Ren, Mingxiang Chen, Xiangchao Meng, Yuxin Zhang, et al. Evaluating mllms with multimodal multi-image reasoning benchmark. *arXiv preprint arXiv:2506.04280*, 2025. 2

[7] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025. 2, 6

[8] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*. 6

[9] Edwin E. Herricks, David Mayer, and Sidney Majumdar. Foreign object debris characterization at a large international airport, 2015. DOT/FAA/TC-TN14/48. 2, 4, 5

[10] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 2

[11] Rahima Khanam and Muhammad Hussain. What is yolov5: A deep look into the internal features of the popular object detector. *arXiv preprint arXiv:2407.20892*, 2024. 3

[12] Marios Krestenitis, Eftichia Badeka, Ilias Koulalis, Konstantinos Ioannidis, and Stefanos Vrochidis. Enhanced defect detection in airport runway infrastructure using image-text pairing. In *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–7, 2024. 3

[13] Necip Sahamettin Kucuk, Hakan Aygun, Omer Osman Dursun, and Suat Toraman. Detection and classification of foreign object debris (fod) with comparative deep learning algorithms in airport runways. *Signal, Image and Video Processing*, 19(4):1–15, 2025. 3

[14] Alka Kumari, Abhishek Dixit, and Pooja Agrawal. Deep learning based foreign object debris (fod) detection on runway. In *2024 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6. IEEE, 2024. 3

[15] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, 2025. 2, 6

[16] Emiel Miedema, Sabine Waschull, and Christos Emmanouilidis. Towards trustworthy artificial intelligence for decision-making: A lifecycle perspective on knowledge-and data-driven artificial intelligence systems. *Computers in Industry*, 174:104409, 2026. 1

[17] Raffaele Mineo, Amelia Sorrenti, Robin Faro, Gabriele Mineo, Francesco Cancelliere, and Alberto Faro. Pcb-said: A low-cost camera-based dataset for few-shot smd assembly inspection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1351–1357, 2025. 2

[18] Lingfeng Ming, Yadong Li, Song Chen, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Oceanr1: An open and generalizable large vision–language model enhanced by reinforcement learning, 2025. Accessed 2026-02-05. 2, 6

[19] Travis Munyer, Peichi Huang, Chenyu Huang, and Xin Zhong. Fod-a: A dataset for foreign object debris in airports. In *International Conference on Machine Learning and Applications (ICMLA)*, 2021. 2, 3, 4

[20] Travis Munyer, Daniel Brinkman, Xin Zhong, Chenyu Huang, and Iason Konstantzos. Foreign object debris detection for airport pavement images based on self-supervised localization and vision transformer. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1388–1394. IEEE, 2022. 3

[21] Chunyi Peng, Zhipeng Xu, Zhenghao Liu, Yishan Li, Yukun Yan, Shuo Wang, Zhiyuan Liu, Yu Gu, Minghe Yu, Ge Yu, et al. Learning to route queries across knowledge bases for step-wise retrieval-augmented reasoning. *arXiv preprint arXiv:2505.22095*, 2025. 6

[22] Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, et al. Vldbench evaluating multimodal disinformation with regulatory alignment. *arXiv preprint arXiv:2502.11361*, 2025. 2

[23] Aadarsh Sahoo and Georgia Gkioxari. Conversational image segmentation: Grounding abstract concepts with scalable supervision, 2026. 2

[24] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao,

Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 2

[25] Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Chi Chen, Zhiyuan Liu, and Maosong Sun. Visrag 2.0: Evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025. 2, 3, 4, 6

[26] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025. 6

[27] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pages 1–6. IEEE, 2024. 3

[28] Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 6

[29] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025. 2, 6

[30] Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025. 6

[31] Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, et al. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models. *arXiv preprint arXiv:2505.24164*, 2025. 2

[32] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. 6

[33] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *CoRR*, 2025. 2, 6

[34] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[35] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Advances in Neural Information Processing Systems*, 37:49279–49383, 2024. 2