What the Label Admits, the Model Asserts: Evidence-Constrained Supervision of Vision-Language Models from Expert Prose

Anonymous authors
Paper under double-blind review

Abstract

Professional inspection practices hold years of photographs paired with expert free-text opinions written for human readers, not for model supervision. We study how to turn such legacy prose into supervision for a vision-language model (VLM) that produces structured, evidence-constrained defect reports, using a private corpus of 4,231 photographs from 586 expert-witness housing-condition reports prepared for disrepair claims in England. Each surveyor opinion follows an evidence–cause–remedy convention; a locally hosted text-only LLM converts only the visible-evidence component into a controlled JSON schema over a 52-type, 20-family defect taxonomy derived from the corpus, so that diagnosed causes never become visual labels, and every record passes structural, taxonomic and semantic validation, with flagged records excluded from training. Three open-weight VLMs of 7–9B parameters (Qwen2.5-VL, Qwen3-VL, GLM-4.1V) are adapted with QLoRA on a single 24 GB GPU under a property-level, near-duplicate-safe split. Adaptation raises type-level micro-F1 from at best 0.186 (4-shot prompting) to 0.432 with disjoint bootstrap intervals, and schema-valid output from 69.6% to 100.0%, while preserving the expert's evidence-only register. Under five-fold property-level cross-validation with a paired property-level bootstrap, swapping among three backbones from two vendors moves micro-F1 by at most 0.035, several times less than adaptation itself. A label-content ablation then separates recognition from faithfulness: converting the whole opinion instead of only its evidence leaves recognition unchanged within bootstrap noise, but the model reproduces the register of its labels almost exactly, asserting an unverifiable cause on 95.6% of images against 0.0% for the evidence-only model. In our experiments recognition is driven by adaptation and data volume (no saturation over the tested range); what the model asserts is largely determined by what the label admits. We release the schema, taxonomy, prompts, validation rules and evaluation code; the imagery remains private.

1 Introduction

Every clinical, industrial, insurance and legal domain has accumulated the same asset: images attached to expert free text that was written to persuade or inform a human reader. Housing-condition reporting in England is a sharp instance. Housing-conditions litigation and the statutory duties on damp and mould introduced after the death of Awaab Ishak (UK Parliament, 2023) have increased both the volume of inspections and the expectation that findings are evidenced, consistent and auditable. For every photograph, an expert-witness surveyor states what is visible, separates that observation from an inferred cause, names the affected building element, and recommends a proportionate remedy. Years of such opinions exist, but none of them was written as a training target.

Vision-language models (VLMs) can in principle produce exactly this kind of output: a structured report with several findings per image, each with evidence, severity, confidence and limitations. The practical question is what supervision they should receive. Purpose-built

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.

annotation is expensive and, in a private legal corpus, cannot be crowdsourced. Converting the existing prose is cheap but risky, because the expert's inference ("this is consistent with a leak") is not something a model can verify from pixels, and a model trained to reproduce it will learn to assert causes it cannot see.

This paper studies that conversion as the object of research. We turn a corpus of 586 expert-witness reports (5,257 image–opinion pairs, of which 4,231 photographs with a usable opinion form the dataset) into supervision for three open-weight VLMs, using a locally hosted text-only LLM that converts only the visible-evidence component of each opinion into a controlled schema, followed by structural, taxonomic and semantic validation. We then ask what determines what the adapted model learns. Our findings are:

- Adaptation, not prompting, delivers the task. Against zero-shot, taxonomy-prompted and few-shot baselines on the same backbone, QLoRA adaptation raises type-level micro-F1 by a factor of 2.3 ($0.186 \rightarrow 0.432$, disjoint bootstrap intervals) and takes schema- and taxonomy-valid output from at most 69.6% to 100.0%, without diagnostic vocabulary leaking into generated evidence (§5.1).
- The backbone is second-order. Under five-fold property-level cross-validation, three backbones from two vendors (Qwen2.5-VL-7B, Qwen3-VL-8B, GLM-4.1V-9B) differ by at most 0.035 micro-F1 under a paired property-level bootstrap, several times less than the adaptation gain; the newest backbone is reliably but modestly better (§5.3).
- Label content sets faithfulness, not recognition. Converting the whole opinion instead of only its evidence component leaves recognition unchanged within noise but changes what the model asserts: cause vocabulary appears in 95.6% of its evidence sentences against 0.0% for the evidence-only model, tracking the label rate almost exactly (§5.4).
- Failure follows the supervision. The adapted models' one regression — reporting defects on defect-free images — appears exactly where the converted corpus is thinnest (2.0% of training targets are empty), and we show it is common to all backbones (§5.6).

The transferable recipe (evidence/diagnosis separation, the validation stack, the leakage-safe property-level split, and multi-granularity evaluation with no-defect and unsupported-claim rates) is released with the schema, taxonomy and prompts; the dataset is private and is not the contribution.

2 Related work

Computer vision for building inspection. Inspection research is organised around classification, detection and segmentation of narrowly defined defect classes, most often cracking and concrete deterioration (Spencer et al., 2019; Kang et al., 2025), and reviews call for evaluation under realistic field conditions and for modelling defects rather than only detecting them (Chen et al., 2024). Operational systems couple detectors with BIM and augmented reality to manage findings (Tan et al., 2024). All of these return predetermined labels or masks rather than an evidence-based report.

VLMs in construction and inspection. Recent work prompts VLMs for safety-hazard identification (Adil et al., 2025), introduces ConstructionSite 10k with captioning, VQA and grounding annotations and finds that pretrained VLMs need further training before field use (Chen & Zou, 2025), and benchmarks 18 multimodal models on façade pathology, finding strong semantic perception but weak localisation (Zhong et al., 2026). These studies rely on purpose-built annotations; we study supervision derived from legacy expert prose, and our target is a multi-issue structured report rather than a caption or a mask.

Supervision from expert free text. The closest structural analogue is radiology, where reports have a findings section (what is seen) and an impression section (what it means), and report-generation and labelling work must decide which to learn from (Johnson et al.,

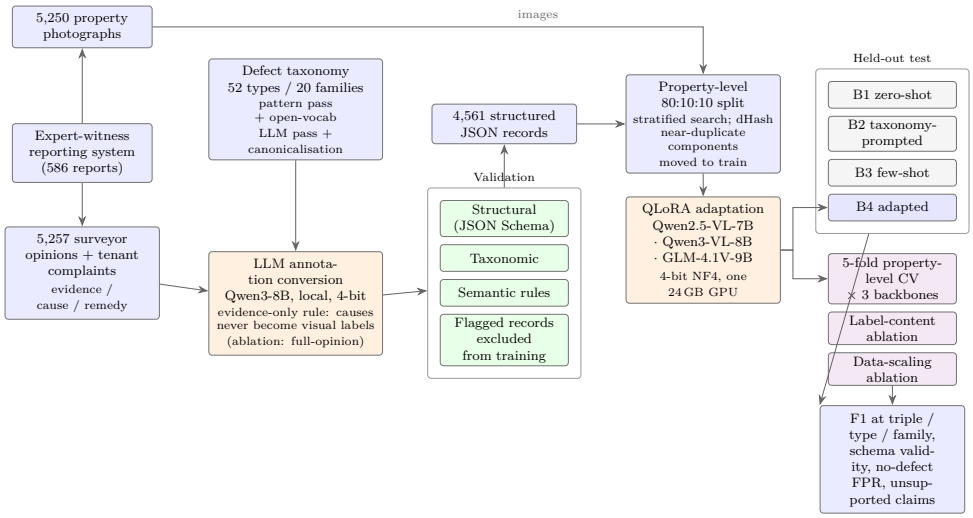


Figure 1: Pipeline. Legacy opinions are converted by a local text-only LLM under the evidence-only rule, validated at four levels, split at property level with connected-component near-duplicate removal, and used to adapt three VLM backbones that are evaluated against prompted baselines and under label-content and data-scaling ablations.

2019a;b; Jing et al., 2018). Rule-based and learned labellers extract structured labels from such reports (Irvin et al., 2019; Smit et al., 2020), and LLMs are now used as annotators more broadly (Wang et al., 2021; Gilardi et al., 2023), in the tradition of weak supervision (Ratner et al., 2017). Our evidence-only rule is the findings/impression split applied at conversion time; unlike prior work we ablate it and measure its effect on the downstream VLM.

Hallucination and unsupported claims. Object hallucination in captioning and VLMs is measured with CHAIR and POPE (Rohrbach et al., 2018; Li et al., 2023). In an inspection report the analogous failure is a claim the image cannot support — an invented defect on a clean wall, or a stated cause. We measure both: a no-defect false-positive rate on defect-free images and a lexical unsupported-claim rate.

Leakage. Group-level splitting and duplicate control are known requirements for credible evaluation (Kaufman et al., 2012; Kapoor & Narayanan, 2023); our corpus has both hazards and §3.5 treats them explicitly.

3 From expert prose to supervision

3.1 Corpus

The corpus was extracted from the relational database of a live expert-witness reporting system operated by a UK surveying practice. Two tables carry the supervision: reports (586, one per property inspection) and report items (5,257, each pairing one photograph with a tenant complaint and a surveyor opinion; 4,592 have a non-empty opinion, 4,055 of them distinct because one opinion is often attached to several photographs of the same condition). All 5,250 referenced photographs were recovered (PNG 3,020, JPEG 1,280; 151–4,032 px per side); a further ∼21,000 photographs exist in the system without any attached opinion. The tenant complaint ("Bathroom — there is damp and mould present") is treated as context only. The opinion follows a stable three-part convention that the whole method rests on: (i) visible evidence ("elevated moisture-meter readings, black mould and spoilt decoration to the plaster walls"), (ii) causal interpretation hedged in professional register ("this is all consistent with moisture damage …likely an intermittent or historic leak"), and (iii) recommended remedy. Only (i) is treated as a source of visual ground truth; (ii)

informs limitation fields and (iii) the recommendation field. The label distribution is heavily skewed towards damp- and mould-related conditions (Figure 3); explicit no-defect items are rare (2.0% of training targets have an empty issue list). Addresses, party names and any occupant-identifying metadata are excluded from all processing.

3.2 Taxonomy

The taxonomy was derived empirically. A deterministic pattern pass over the evidence paragraphs (n-gram profiling and iteratively refined rules) yielded 47 provisional defect types in 20 families covering 88.2% of items. An open-vocabulary LLM pass then classified every item against the provisional taxonomy while being allowed to propose new snake_case types; 145 raw labels were canonicalised by alias and home-family rules, five genuinely new types were promoted (e.g. vermin contamination, gutter blockage), giving 52 canonical types, of which 49 occur in the built dataset. Building elements follow a parallel 19-class list. The families and types are listed in Appendix B. The two extractors were built independently, and their disagreement (family-level full agreement 56.2%) is used as a review trigger rather than resolved automatically.

3.3 Evidence-constrained conversion

A locally hosted text-only LLM (Qwen3-8B, 4-bit NF4 (Yang et al., 2025)) drafts one structured record per report item; running it locally keeps the private text on-premises. The record schema is

{inspection_summary, issues: [{building_element, defect_family, defect_type, location_text, visual_evidence, severity, confidence, recommended_action, limitations}], overall_limitations}.

The conversion prompt (Appendix A) receives the canonicalised class hints for the item and is required to emit one issue object per hinted defect; enumerated fields are enforced by JSON Schema and (family, type) membership in code. The rule that matters for this paper is rule 2: visual_evidence may contain only what the surveyor states is visible, paraphrased; a cause or diagnosis may never appear there and is instead recorded as uncertainty in limitations ("the cause cannot be confirmed from the image alone; the surveyor considered it consistent with a leak"). Severity and confidence must be grounded in the surveyor's wording (decorative wording $\rightarrow$ minor; hazard language $\rightarrow$ severe; hedges $\rightarrow$ lower confidence) and never invented. Duplicate opinion texts are converted once and cached; a structurally invalid response is retried once. Of 4,592 items, 4,561 (99.3%) produced a valid record.

3.4 Validation

Records pass four levels of checking. Structural: valid JSON, required fields, permitted enumerations, no extra fields. Taxonomic: element, family and type permitted and consistent. Semantic: rules that flag a no-defect summary with issues, a defect summary with no issues, high confidence with hedged source wording, a severe rating without risk wording, duplicate issues, and definitive claims of hidden causes (subsidence, rising damp, active leak, asbestos, dry rot, structural failure) in evidence fields; 293 records (6.4%) were flagged, most often for an empty issue list where the source does not read as no-defect. Human: a side-by-side review interface exists in which a surveyor confirms or corrects records; the flags order its queue. Expert review had not been completed at the time of writing, so in every experiment reported here flagged records are excluded from training (183 items) and retained in the evaluation partitions, where they are counted (35; 75 more are placeholder items, excluded below); removing the 14 flagged test images changes type-level micro-F1 by at most 0.005 in every configuration (Appendix G). The effect of expert correction is left to future work (§6). A further 140 items whose "opinion" is a placeholder (Updated, DELETE) are excluded from every partition, because any label for them could only have come from the tenant complaint. Fields that depend on non-visual information are constrained at build time: moisture-meter readings are kept as evidence only when the instrument is in frame, and are otherwise generalised to the visible surface condition.

3.5 Property-level, near-duplicate-safe splitting

Photographs from the same property are often near-duplicates of one condition, so all items of a report stay in one partition. The 80:10:10 assignment is chosen by a seeded randomised greedy search (1,000 trials) that jointly minimises deviation from the item ratio, the family distribution, the no-defect share and the element distribution; the selected split holds family proportions within one to three points across partitions. The same photograph can also appear in two reports (re-inspections, shared instructing parties), which property grouping cannot catch. We compute a 64-bit difference hash for every photograph and call two images near-duplicates at Hamming distance ≤ 6; 13 near-uniform photographs whose hashes carry almost no bits are excluded from matching because they would match everything. Cross-report near-duplicate pairs (327) are grouped into connected components, and every image of a component that touches more than one partition is moved to training (16 components, 152 images), so no evaluation image has a near-duplicate — direct or transitive — in training. Re-scoring an adapter trained before this refinement on the cleaned test partition changed its type-level micro-F1 by less than 0.002, so the removed leakage had not inflated the earlier estimates; it is removed on principle. The final partitions hold 3,284/470/477 image-level samples (a photograph referenced by several items becomes one multi-issue sample). For cross-validation, the training and validation reports are re-partitioned into five property-level folds by the same search, the test partition is left untouched, and the near-duplicate rule is re-applied per fold.[1]

4 Adaptation and evaluation

Training format. Each sample is a chat with a fixed system instruction ("report only visible conditions; never assert hidden causes; return JSON"), the image, a fixed user request and the structured target; only target tokens contribute to the loss. Before every full run we verify the loss mask by decoding exactly the supervised tokens of one sample and require the model to overfit a 16-sample subset.

Backbones and recipe. We adapt Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), its successor Qwen3-VL-8B-Instruct (Bai et al., 2025a), and the architecturally unrelated GLM-4.1V-9B-Thinking (Hong et al., 2025) with one recipe: QLoRA (Hu et al., 2022; Dettmers et al., 2023) with 4-bit NF4 base weights and bf16 compute, LoRA rank 16 ($\alpha = 32$, dropout 0.05) on all language-model projection layers, vision tower and merger frozen, two epochs, learning rate 10^{-4} with cosine decay and 3% warm-up, batch size 1 with 8-step gradient accumulation, gradient checkpointing, and a 512-visual-token budget per image (converted to pixels per backbone's patch size). One run takes about 2.5 h on a single 24 GB RTX 3090; all runs use the same seed. Hyper-parameters are listed in Appendix D.

Baselines. On Qwen2.5-VL-7B we compare B1 zero-shot with the training prompts only; B2 prompt-engineered, adding the full taxonomy, element list and evidence rules to the system prompt; B3 few-shot, adding synthetic worked examples (two, four or eight examples) (no corpus text or image is used as an exemplar); and B4 the QLoRA-adapted model. Base weights, quantisation and image-token budget are identical, so the comparison isolates prompting against adaptation. Frontier API models are not baselines because the imagery must remain on-premises.

Protocols. (i) Main comparison: B1–B4 on the held-out test partition. (ii) Backbone cross-validation: each backbone trained five times (one per fold) and validated on its fold, reported as mean ± standard deviation with paired fold-wise win counts, then confirmed once on the test partition; because every backbone is scored on identical images, backbone differences are tested with a paired property-level bootstrap (2,000 property resamples). (iii) Label-content ablation: the conversion is re-run with rule 2 removed, so that the surveyor's

[1] The folds were built with the per-fold rule before the connected-component refinement; re-checking with the final rule finds 7 validation images in residual transitive components, and excluding them from scoring changes every backbone's cross-validation mean type-level F1 by <0.001 (no fold by more than 0.002), so we report the runs as trained.

Table 1: Main comparison on the held-out test partition (477 images, 526 reference issues, 25 defect-free images), Qwen2.5-VL-7B throughout. ND-FPR: share of defect-free images on which an issue is reported. Unsupp.: unsupported-claim rate.

Configuration	Schema	$F1_{triple}$	$F1_{type}$	$F1_{family}$	Macro-F1	ND-FPR	Unsupp.
B1 zero-shot	12.8%	0.000	0.000	0.002	0.012	32.0%	0.2%
B2 prompted	69.6%	0.058	0.130	0.305	0.167	48.0%	0.0%
B3 few-shot (2 examples)	62.5%	0.069	0.155	0.346	0.190	28.0%	0.0%
B3 few-shot (4 examples)	63.3%	0.068	0.186	0.354	0.204	24.0%	0.0%
B3 few-shot (8 examples)	63.5%	0.067	0.181	0.355	0.215	40.0%	0.0%
B4 QLoRA-adapted	100.0%	0.298	0.432	0.595	0.403	56.0%	0.0%

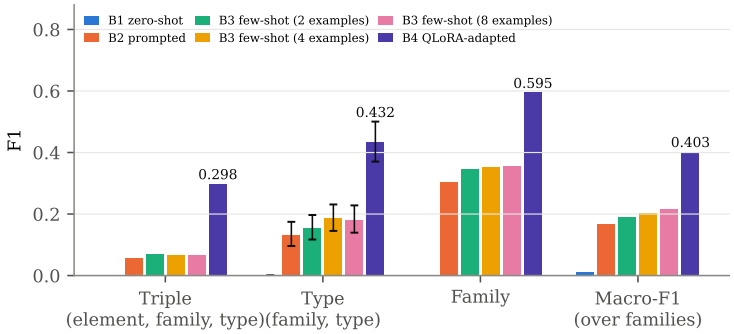


Figure 2: Recognition performance of the configurations in Table 1 at the three matching levels and macro-F1 over families. Whiskers are 95% property-level bootstrap intervals on type-level micro-F1; the adapted model's interval does not overlap any baseline's.

stated cause is written into visual_evidence ("staining to the ceiling consistent with a leak from the bathroom above"); everything else in the prompt, the taxonomy and the training recipe is unchanged, and the model is trained on Qwen2.5-VL-7B and evaluated against the evidence-only test targets. (iv) Data scaling: training on seeded property-level subsets of 25% and 50% of training reports chosen by the same stratified search, with validation and test unchanged.

Metrics. Predicted and reference issues are matched per image by multiset intersection at three granularities: triple (element, family, type), type (family, type) and family; we report micro precision/recall/F1 pooled over issues and macro-F1 over families with non-zero support, so invalid predictions count as false positives rather than being ignored. Structured-output quality is the share of outputs that parse, satisfy the schema and use only taxonomy labels. No-defect behaviour is the false-positive rate (share of defect-free reference images on which at least one issue is reported) and the mean number of invented issues on those images. Severity and confidence agreement with the reference is measured on type-matched issues. The unsupported-claim rate is the share of outputs whose summary or evidence asserts one of the banned definitive diagnoses; for the ablation we additionally report a broader cause-term rate using the lexicon in Appendix C. Uncertainty is quantified with 95% cluster-bootstrap intervals that resample the 63 test properties with all their images (Efron & Tibshirani, 1994).

5 Results

5.1 Adaptation versus prompting

Table 1 and Figure 2 report the six configurations. Adaptation raises type-level micro-F1 by a factor of 2.3 over the best prompted baseline (0.432 vs. 0.186; 95% intervals 0.371–

Table 2: Three backbones under one recipe. Top: five-fold property-level cross-validation. Bottom: held-out test partition. Every one of the 15 cross-validation and 3 test runs is 100% valid JSON and ≥99.3% schema-valid, so the comparison concerns recognition only.

Backbone	$F1_{triple}$	$F1_{type}$	$F1_{family}$	Macro-F1	ND-FPR
Five-fold property-level cross-validation (mean ± std over folds)					
Qwen2.5-VL-7B	0.277 ± 0.027	0.375 ± 0.043	0.528 ± 0.050	0.333 ± 0.053	0.672 ± 0.145
Qwen3-VL-8B	0.313 ± 0.040	0.399 ± 0.052	0.549 ± 0.051	0.358 ± 0.032	0.666 ± 0.161
GLM-4.1V-9B	0.291 ± 0.047	0.383 ± 0.065	0.541 ± 0.062	0.341 ± 0.042	0.690 ± 0.133
Held-out test partition					
Qwen2.5-VL-7B	0.298	0.432	0.595	0.403	56.0%
Qwen3-VL-8B	0.310	0.433	0.585	0.418	44.0%
GLM-4.1V-9B	0.305	0.415	0.566	0.379	56.0%

0.501 vs. 0.145–0.231). More exemplars do not close the gap (4 examples 0.186; 8 examples 0.181). Zero-shot recognition is effectively nil: almost every B1 output parses as JSON, but only 12.8% satisfy the schema, because the model invents its own vocabulary (106 distinct off-taxonomy family labels that can never match a reference). Supplying the taxonomy (B2) or worked examples (B3) recovers structure only partially (at most 69.6% schema-valid), whereas the adapted model is schema- and taxonomy-valid on 100.0% of test images, which downstream software requires. Attribute quality follows recognition: on type-matched issues the adapted model agrees with the reference severity in 77.9% and the reference confidence in 78.4% of cases (Appendix E). The unsupported-claim rate is 0.2% for zero-shot and zero for every other configuration; on its own this is weak evidence, because the schema has no field for a diagnosis and the rate only catches banned vocabulary leaking into evidence text, so §5.4 tests the evidence-only rule directly by changing what the labels contain.

5.2 Where the adapted model fails

Per-family F1 (Appendix F, Figure 4) divides by supervision volume and visual distinctiveness. Mould, the regulatory priority of the corpus, and vegetation are the best families (F1 0.81 and 0.83); glazing, cracking, masonry and damp staining follow at 0.58–0.70. Surface-finish failure, the largest family, has high precision but low recall (0.78/0.36): failed decorations co-occur with, and are visually subsumed by, damp staining and mould in the same photograph, and the model reports the dominant condition. Component damage shows the reverse profile (0.43/0.65) and acts as a catch-all. Families with about a dozen or fewer test items (timber rot, gutter drainage, gaps, deformation, leak evidence) are unlearned, and joint/seal failure and electrical are barely learned. Within recognised defects the largest remaining error is element attribution: micro-F1 falls from 0.432 at type level to 0.298 when the element must also match.

5.3 The backbone is second-order

Tables 2 and 3 show three things. First, the backbone is a small effect. Per-model fold statistics overlap (fold-to-fold standard deviations of ±0.027–0.047), but the paired bootstrap over 491 properties resolves the difference: Qwen3-VL is reliably better than Qwen2.5-VL under cross-validation (+0.034 (95% CI +0.022 to +0.049) triple-level, +0.021 (95% CI +0.008 to +0.036) type-level; it wins all five folds at triple level and four at type level), GLM-4.1V sits between the two, and on the single 63-property test partition every difference is inside its interval (-0.015 (95% CI -0.053 to +0.023) for the same pair). The largest backbone effect is therefore about 0.03 micro-F1, against a gain of 0.25 from adaptation over the best prompted baseline, about seven times larger. Second, the differences that persist concern caution and the long tail: Qwen3-VL has the best macro-F1 in both protocols and the lowest no-defect false-positive rate on the test partition (44.0% against 56.0% for both other backbones) (cross-validation rates, 0.67–0.69, do not separate the backbones). Third, the recipe transfers across vendors: GLM-4.1V, an unrelated architecture that normally opens

Table 3: Paired property-level bootstrap of backbone differences (2,000 resamples). Each cell is the micro-F1 difference of the second backbone minus the first, with its 95% interval; cross-validation pools the five validation folds.

	Cross-validation (491 prop.)		Test (63 prop.)	
Backbone change	$\Delta F1_{triple}$	$\Delta F1_{type}$	$\Delta F1_{triple}$	$\Delta F1_{type}$
Qwen2.5-VL-7B → Qwen3-VL-8B	+0.034 [+0.022, +0.049]	+0.021 [+0.008, +0.036]	-0.015 [-0.053, +0.023]	-0.003 [-0.045, +0.038]
Qwen2.5-VL-7B → GLM-4.1V-9B	+0.014 [-0.001, +0.031]	+0.006 [-0.011, +0.024]	-0.018 [-0.051, +0.015]	-0.019 [-0.052, +0.016]
Qwen3-VL-8B → GLM-4.1V-9B	-0.020 [-0.032, -0.007]	-0.016 [-0.028, -0.001]	-0.002 [-0.036, +0.035]	-0.015 [-0.051, +0.023]

Table 4: Label-content ablation on Qwen2.5-VL-7B. Both models are scored against the evidence-only test targets. Cause terms: share of outputs whose summary or evidence contains a term from the cause lexicon (Appendix C).

Conversion rule	Cause terms in labels	$F1_{type}$	$F1_{family}$	Macro-F1	ND-FPR	Unsupp.	Cause terms in outputs
Evidence-only (ours)	0.3%	0.432	0.595	0.403	56.0%	0.0%	0.0%
Full-opinion	90.9%	0.420	0.578	0.387	52.0%	0.0%	95.6%

with a reasoning block, learns to answer directly with valid JSON and lands in the same band on every metric. What the adapted models learn is therefore set by the labels and the training format, which motivates the next two experiments.

5.4 Label content: evidence-only versus full-opinion conversion

Table 4 isolates label content. The two arms were built from the same items with the same exclusions and the same near-duplicate list; the two training sets share 3,219 images, and on 3,086 of them the (family, type) multiset is identical, so the arms differ almost only in what the evidence sentence says. Cause vocabulary (Appendix C) appears in 0.3% of evidence-only evidence fields and in 90.9% of full-opinion ones (the evidence-only rule moves it to the limitation fields: 9.5% versus 0.2%). Recognition does not separate the arms: type-level micro-F1 is 0.420 against 0.432 (intervals 0.359–0.485 and 0.371–0.501), family F1 0.578 against 0.595, macro-F1 0.387 against 0.403, and the no-defect rate moves within its wide interval (52.0% against 56.0%). What separates them is what the model asserts. The full-opinion model writes a cause into its evidence on 95.6% of test images; the evidence-only model on 0.0%. Each model reproduces the register of its labels almost exactly (90.9% → 95.6%; 0.3% → 0.0%), which is the direct evidence for the paper's premise: what is admitted into the label is what the model will say. The banned-diagnosis rate is zero for both arms, confirming that the narrow lexicon is blind to this behaviour: the full-opinion model hedges exactly like the surveyor ("consistent with a leak from above") while asserting, for almost every image, a cause it cannot verify from pixels. The corpus holds no ground truth for causes beyond the surveyor's inference, so those assertions cannot be scored, which is why they do not belong in the label of a system whose output must be auditable against the photograph.

5.5 Data scaling

Table 5 trains on stratified property-level subsets of 163 and 237 reports (933 and 1,844 samples) against the full 443 (3,284). Type-level micro-F1 rises $0.373 \rightarrow 0.404 \rightarrow 0.432$, family F1 $0.532 \rightarrow 0.556 \rightarrow 0.595$, and macro-F1 $0.329 \rightarrow 0.350 \rightarrow 0.403$: about +0.03 type-level F1 per doubling with no sign of saturation over the tested range, and the last doubling helps the long tail most (macro-F1 +0.05). Two conclusions follow. The adaptation gain is not a large-data artefact: a quarter of the properties already gives 0.373 type-level F1, well above the best prompted baseline (0.186), and schema validity is 97.9% at that size. And performance had not saturated over the tested range. The no-defect rate does not improve with volume (52%, 40%, 56%), consistent with it being a balance problem rather than a size problem (§5.6).

Table 5: Training on stratified property-level subsets (Qwen2.5-VL-7B), validation and test unchanged.

Training properties	Reports	Samples	Schema	$F1_{type}$	$F1_{family}$	Macro-F1	ND-FPR
25%	163	933	97.9%	0.373	0.532	0.329	52.0%
50%	237	1,844	99.6%	0.404	0.556	0.350	40.0%
100%	443	3,284	100.0%	0.432	0.595	0.403	56.0%

5.6 No-defect behaviour

The one regression under adaptation is that the model reports something on genuinely defect-free images: 56.0% of the 25 defect-free test images receive at least one issue (95% interval 37.0–92.9%), worse than few-shot prompting (24.0%), and every backbone in Table 2 shows an elevated rate. Of the 25 defect-free references, 23 carry explicit no-defect wording in the source opinion, so the references are genuine. The invented issues are not random: 8 of the 15 issues reported on these images are component damage, the catch-all family that also shows the lowest precision among frequent families (§5.2), so the model has learned that a domestic scene usually contains something reportable. The failure appears where supervision is thinnest (2.0% of training targets are empty), which is consistent with insufficient negative-example coverage rather than a backbone-specific failure; testing it requires adding negatives, which is future work.

6 Discussion and limitations

What transfers. The results argue for spending effort on the label pipeline, where faithfulness is decided, before the backbone: convert only the evidence component so the model is never supervised to assert what it cannot see, validate and exclude what fails, and evaluate structured reports together with no-defect and cause-term behaviour so that fluent but unsupported output is penalised. None of this is specific to housing; radiology, insurance and industrial inspection have the same findings-versus-impression structure in their expert text.

Limitations. Ground truth originates from prose that may omit visible defects, and one opinion often covers several photographs, so image–label pairs are imperfectly specific. The conversion is machine-drafted: automated validation removed what it could detect, but expert correction of every record is incomplete and, with the raw-versus-corrected training arm it enables, is the most important follow-up. All runs use one seed, so seed variance is unmeasured. Some evidence is non-visual (meter readings) and some conditions are indistinguishable in one image (active versus historic moisture). The class distribution is long-tailed, so rare-family results are exploratory; the corpus reflects one practice's instructions, housing stock and cameras; and the system evaluates single images where professional assessment integrates several views, measurements and history. Finally, the dataset and adapted weights cannot be released (see the Reproducibility Statement).

7 Conclusion

Legacy expert prose can supervise open-weight VLMs that produce structured, evidence-constrained defect reports on one consumer GPU; the two things such a model learns have different drivers. Recognition is driven primarily by adaptation and data volume: a factor of 2.3 in type-level F1 over prompting, 100.0% schema validity, no saturation over the tested range, and at most 0.035 from the backbone. Faithfulness is driven by the label: what is admitted into it largely decides whether the model reproduces the expert's caution or the expert's inference. Over-reporting on defect-free images is consistent with scarce negative examples rather than model choice; negative-example augmentation is the next experiment.

Ethics statement

The corpus consists of photographs of occupied private dwellings, tenant complaints and the professional work product of an expert-witness practice, obtained under a data-sharing agreement with that practice; one author is an officer of the company that owns the data, which will be disclosed on de-anonymisation. Addresses, party names and occupant-identifying metadata were removed before processing; no face or identity analysis is performed, and the model is never used to infer attributes of occupants. Housing-condition reports influence repair obligations, litigation and tenant welfare, so the system is decision support only: outputs are labelled as machine-generated, carry mandatory limitation fields, and require review by a qualified surveyor before any consequential use. Neither the imagery nor the adapted weights will be released.

Reproducibility statement

The dataset is private; everything else is released in the anonymised supplementary code: the extraction and profiling scripts, the taxonomy (Appendix B), the conversion prompts for both variants (Appendix A), the JSON Schema and semantic-validation rules (Appendix C), the stratified splitting and near-duplicate procedure (§3.5), the training script with all hyperparameters (Appendix D), the evaluation prompts for B1–B3, and the scoring implementation including the bootstrap. Aggregate corpus statistics and per-fold results are given in the appendices so that a practice holding a comparable corpus can reproduce the pipeline end to end.

AI use statement

We used a commercial large language model assistant throughout this work, as follows. Writing: drafting and editing the manuscript, including first drafts of the results and discussion prose from the generated tables, which the authors revised. Code: writing and refactoring the data-processing, training, evaluation and analysis code, which the authors reviewed, tested and ran. Methodology: proposing or refining parts of the experimental design (the connected-component near-duplicate removal, the label-content ablation protocol and the paired property-level bootstrap), which the authors adopted after discussion. Data: designing the dataset cleaning rules (leakage removal, placeholder exclusion). Interpretation: assisting the authors in interpreting the results by drafting the prose around the metric tables. Literature: assisting the search for related work; every citation was checked by the authors. A locally hosted open-weight LLM (Qwen3-8B) is part of the studied method: it drafts the structured training targets from expert prose as described in §3.3. Generative AI was not used to generate synthetic evaluation data or to formulate or prove mathematical claims. The authors verified every number, claim and citation against the underlying metric files and sources, and take full responsibility for the final content.

References

Muhammad Adil, Gaang Lee, Vicente A. Gonzalez, and Qipei Mei. Using vision language models for safety hazard identification in construction. arXiv preprint arXiv:2504.09083, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923, 2025b.

Junjie Chen, Isabelle Chan, and Ioannis Brilakis. Shifting research from defect detection to defect modeling in computer vision-based structural health monitoring. Automation in Construction, 164:105481, 2024.

Xuezheng Chen and Zhengbo Zou. Are large pre-trained vision language models effective construction safety inspectors? arXiv preprint arXiv:2508.11011, 2025.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems, volume 36, 2023.

Bradley Efron and Robert J. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall/CRC, 1994.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30): e2305016120, 2023.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pp. 590–597, 2019.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2577–2586, 2018.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6:317, 2019a.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042, 2019b.

Suhyun Kang, Seungho Kim, and Sangyong Kim. Automatic detection and classification process for concrete defects in deteriorating buildings based on image data. Journal of Asian Architecture and Building Engineering, 24(4):2773–2787, 2025.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9):100804, 2023.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4):1–21, 2012.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 292–305, 2023.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. Proceedings of the VLDB Endowment, 11(3):269–282, 2017.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4035–4045, 2018.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 1500–1519, 2020.

Billie F. Spencer, Jr., Vedhus Hoskere, and Yasutaka Narazaki. Advances in computer vision-based civil infrastructure inspection and monitoring. Engineering, 5(2):199–222, 2019.

Yi Tan, Wenyu Xu, Penglu Chen, and Shuyan Zhang. Building defect inspection and data management using computer vision, augmented reality, and BIM technology. Automation in Construction, 160:105318, 2024.

UK Parliament. Social housing (regulation) act 2023, c. 36, s. 42 ("awaab's law"). https://www.legislation.gov.uk/ukpga/2023/36, 2023.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Want to reduce labeling cost? GPT-3 can help. In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4195–4205, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Hui Zhong, Yichun Gao, Luyan Liu, Hai Yang, Wang Wang, Haowei Zhang, and Xinhu Zheng. Can large multimodal models inspect buildings? a hierarchical benchmark for structural pathology reasoning. arXiv preprint arXiv:2603.20148, 2026.

A Schema and conversion prompts

Record schema. Every target is a JSON object with inspection_summary (one neutral sentence mentioning only conditions in the issue list), issues (a list, empty for a defect-free image) and overall_limitations (always including "The assessment is limited to conditions visible in the photograph."). Each issue has building_element (19-class enumeration), defect_family and defect_type (taxonomy, Appendix B), location_text, visual_evidence, severity $\in$ {minor, moderate, severe, unknown}, confidence $\in$ {low, medium, high}, recommended_action and limitations (issue-specific uncertainty only).

Conversion system prompt (evidence-only variant, abridged). You are a building-surveying annotation specialist converting a surveyor's written opinion about one property photograph into a structured inspection record. The surveyor text follows a convention: first VISIBLE EVIDENCE, then CAUSE interpretation ("consistent with…"), then RECOMMENDED REMEDY. Use each part for its matching field. (1) Emit exactly one issue object per distinct visible defect; cover every hinted defect. (2) visual_evidence: only what the surveyor states is visible, paraphrased briefly; never place a cause or diagnosis here; causes belong in limitations as uncertainty. (3) location_text: short location, "unspecified" if absent. (4) Severity grounded in the surveyor's wording only (decorative $\rightarrow$ minor; repair works $\rightarrow$ moderate; safety/hazard language $\rightarrow$ severe; otherwise unknown); never invent severity. (5) Confidence: high when stated plainly, medium for mild hedging, low for strong hedging. (6) recommended_action: the surveyor's remedy for this defect. (7–8) Element and (family, type) from the controlled lists. (9) Summary mentions only listed conditions; for no-defect records return an empty issue list. (10) Standard overall limitation always present; per-issue limitations issue-specific only. (11) The tenant complaint is context only. (12) Respond with only the JSON object. A worked example follows in the prompt.

Full-opinion variant. Identical except that rule 2 reads "visual_evidence: summarise the surveyor's finding for this defect, including the cause the surveyor gives (e.g. 'Staining to the ceiling consistent with a leak from the bathroom above.')" and the worked example's evidence sentence carries a cause ("…rotten, consistent with moisture ingress").

Training and evaluation prompts. System: You are a housing-surveying image inspection assistant. Examine the property photograph and report every visible housing defect. Report only conditions visible in the image; never assert hidden causes. Respond with ONLY a JSON object of the form {…}. Return an empty issues list when no defect is visible. User: Inspect this property image and report all visible housing defects. B2 appends the element list, the full taxonomy, the enumerations and the rule that evidence must describe only what is visible; B3 further appends two synthetic worked examples (a mould finding and a defect-free image).

B Taxonomy

Families (20): mould; damp and staining; condensation; cracking; surface finish failure; masonry; timber rot; corrosion; joint and seal failure; glazing; component damage; deformation; gaps and openings; suspected hazardous materials; electrical; leak evidence; vegetation; flooring; roofing; gutters and drainage. Canonical types (52) are listed with their families in the released taxonomy.json; the five promoted from the open-vocabulary pass are gutter blockage, vermin contamination, uneven floor surface, dust accumulation and exposed pipework. Building elements (19): internal wall, external wall, ceiling, floor, window, door, skirting, brickwork/masonry, roof, gutters and drainage, plumbing, radiator/heating, bathroom fittings, kitchen units, electrical fittings, balcony, staircase, external ground, unknown. Figure 3 gives the family distribution of the built dataset.

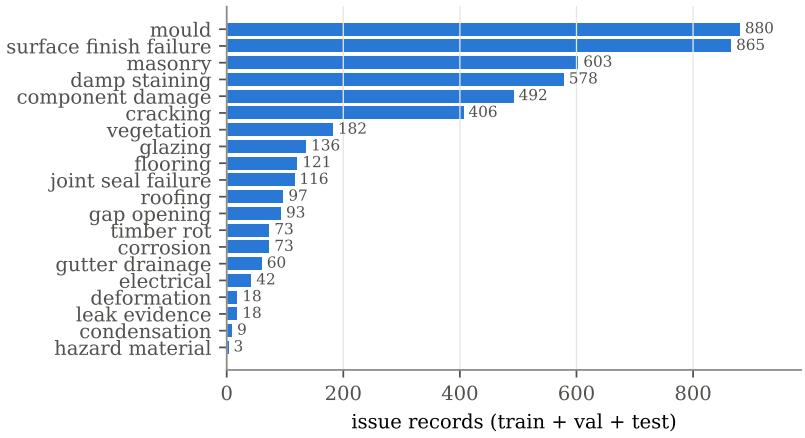


Figure 3: Issue records per defect family in the built dataset (train + val + test; 4,865 issues in total).

C Validation rules and cause lexicon

Semantic flags. A record is flagged when: the summary reads as no-defect but issues are present; the summary reads as a defect but the issue list is empty; the source opinion does not read as no-defect but the issue list is empty; confidence is high while the source wording hedges (possible, possibly, likely, appears, seems, potential, may be); severity is severe without risk wording in the source (safety, hazard, risk, significant, danger, structural); two issues share element, family, type and location; or a banned definitive diagnosis appears in an evidence field. Flag counts by reason are given in the released schema_flags.csv

summary: 119 empty-issue mismatches, 91 severe-without-risk, 41 high-confidence-with-hedge, 7 duplicates, and the remainder banned terms.

Unsupported-claim lexicon (scoring). subsidence, rising damp, active leak, asbestos, dry rot, structural failure — matched case-insensitively as whole words in the summary and evidence fields of an output.

Cause-term lexicon (ablation). consistent with, caused by, due to, as a result of, attributable to, arising from, subsidence, rising damp, penetrating damp, active leak, leak(ing) from, defective {gutter, roof, flashing, pointing, seal, dpc}, structural movement, thermal bridging, inadequate/lack of ventilation.

D Training details

Base weights	4-bit NF4 (double quantisation), bf16 compute
LoRA	rank 16, $\alpha = 32$, dropout 0.05, all language-model projection layers
Frozen	vision tower, multimodal merger
Optimiser	AdamW (paged, 8-bit), lr 10^{-4}, cosine, 3% warm-up, weight decay 0
Schedule	2 epochs, batch 1×8 accumulation, gradient checkpointing
Images	dynamic resolution, at most 512 visual tokens (28^2 px/token for Qwen2.5-VL and GLM-4.1V, 32^2 for Qwen3-VL)
Decoding	greedy, at most 1,024 new tokens
Hardware	one RTX 3090 (24 GB); ≈2.5 h per training run, ≈35 min per 500-image evaluation

Sanity gates run before every full training: a single forward/backward pass with a decoded print-out of exactly the supervised tokens, and a 16-sample overfitting run that must drive the loss near zero.

E Attribute agreement

Configuration	Severity agreement	Confidence agreement	Invented issues / ND image
B1 zero-shot	—	—	0.40
B2 prompted	62.7%	68.7%	0.52
B3 few-shot (2 examples)	59.5%	48.1%	0.36
B3 few-shot (4 examples)	60.4%	52.1%	0.28
B3 few-shot (8 examples)	57.9%	68.4%	0.48
B4 QLoRA-adapted	77.9%	78.4%	0.56

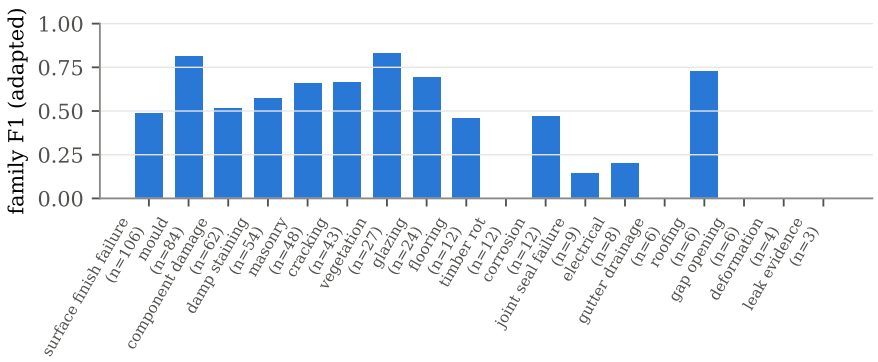


Figure 4: Per-family F1 of the adapted Qwen2.5-VL model on the test partition, ordered by test support. Families with fewer than roughly twenty test examples are unreliable or unlearned.

F Per-family results

Defect family	Test support	Precision	Recall	F1
surface finish failure	106	0.776	0.358	0.490
mould	84	0.795	0.833	0.814
component damage	62	0.430	0.645	0.516
damp staining	54	0.531	0.630	0.576
masonry	48	0.698	0.625	0.659
cracking	43	0.828	0.558	0.667
vegetation	27	0.758	0.926	0.833
glazing	24	0.727	0.667	0.696
flooring	12	0.429	0.500	0.462
timber rot	12	0.000	0.000	0.000
corrosion	12	0.800	0.333	0.471
joint seal failure	9	0.200	0.111	0.143
electrical	8	0.500	0.125	0.200
gutter drainage	6	0.000	0.000	0.000
roofing	6	0.800	0.667	0.727
gap opening	6	0.000	0.000	0.000
deformation	4	0.000	0.000	0.000
leak evidence	3	0.000	0.000	0.000

G Sensitivity to flagged evaluation records

Records that fail a semantic validation rule (§3) are excluded from training but kept in the evaluation partitions, so that evaluation is not restricted to the records the rules find easy. Table 6 rescores every test-partition prediction file with those 14 images removed (463 images remain; no model is re-run). Type-level micro-F1 moves by at most 0.005, family-level F1 by at most 0.005, and every difference between configurations is preserved; the no-defect rate moves by up to five points because the defect-free set is small. Under cross-validation the per-backbone mean type-level F1 changes by at most 0.002. The headline numbers therefore do not depend on the flagged records.

Table 6: Test-partition metrics with and without the 14 semantically flagged reference records.

	All test images			Flagged records removed		
Configuration	$F1_{type}$	$F1_{family}$	ND-FPR	$F1_{type}$	$F1_{family}$	ND-FPR
B1 zero-shot	0.000	0.002	32.0%	0.000	0.002	30.4%
B2 prompted	0.130	0.305	48.0%	0.128	0.300	52.2%
B3 few-shot (2 examples)	0.155	0.346	28.0%	0.153	0.345	30.4%
B3 few-shot (4 examples)	0.186	0.354	24.0%	0.185	0.352	26.1%
B3 few-shot (8 examples)	0.181	0.355	40.0%	0.181	0.355	34.8%
B4 adapted Qwen2.5-VL-7B	0.432	0.595	56.0%	0.438	0.600	52.2%
Adapted Qwen3-VL-8B	0.433	0.585	44.0%	0.435	0.585	39.1%
Adapted GLM-4.1V-9B	0.415	0.566	56.0%	0.417	0.568	52.2%
Full-opinion labels (Qwen2.5-VL)	0.420	0.578	52.0%	0.423	0.583	47.8%
25% training properties	0.373	0.532	52.0%	0.376	0.534	47.8%
50% training properties	0.404	0.556	40.0%	0.409	0.560	34.8%

H Corpus statistics

Reports (properties)	586
Report items (image–opinion pairs)	5,257
Items with a non-empty opinion	4,592
Distinct opinion texts	4,055
Distinct photographs referenced	5,250
Items per report (mean)	9.0
Converted records (valid after $\leq$1 retry)	4,561 (99.3%)
Records with an empty issue list	220 (4.8%)
Records flagged by semantic rules	293 (6.4%)
Items with a placeholder opinion (excluded)	140
Provisional / canonical defect types	47 / 52 (49 observed)
Defect families / building elements	20 / 19
Image-level samples (train / val / test)	3,284 / 470 / 477
Issue records in the built dataset	4,865
Cross-report near-duplicate pairs (dHash ≤ 6)	327
Near-duplicate components spanning a partition / images moved to train	16 / 152
Unattached photographs available for negative mining	$\sim$21,000