VERIFIABLE GROUNDED GENERATION FOR ROAD DAMAGE REPORTS WITH REVISABLE EVIDENCE CONTRACTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Road photographs can support maintenance decisions only when suspected damage is translated into reports that explain what is observed and what remains uncertain. This translation is difficult because a detector may overlook nearby defects, while a written description may claim more than its predictions justify. We investigate these challenges through RoadTrust, a workflow connecting detector evaluation, targeted retraining, and evidence-based report generation. Candidate detectors are compared under a common evaluation protocol, and the selected architecture is trained and tested separately within each image collection. Controlled augmentation studies examine how object size and arrangement affect the visual evidence available for reporting. RoadTrust retains detections alongside available depth readings and optional location metadata, then checks generated descriptions against these records. Its reporting method distinguishes fidelity to an observation record from correctness about the physical road. Executable evidence contracts bind statements to their record identities, required fields, and canonical wording, providing a mechanism for checking which statements remain supported when evidence changes. An intervention audit examines altered wording, record substitutions, withdrawn evidence, and unrelated edits. A complementary writer–verifier workflow and human assessment examine whether natural-language claims are supported and whether their evidence is correct. This design places detector behavior, claim verification, and human judgment within a common inspection workflow. RoadTrust provides a practical framework for studying the full path from image to report, while its usefulness to inspectors and municipal maintenance requires field assessment.

1 INTRODUCTION

Road agencies increasingly receive photographs from phones, vehicles, and residents, and a detector can mark suspected damage in each image within a fraction of a second. A maintenance decision, however, depends on a report that states what was observed, how the observation was obtained, and which questions still need a site visit. A box around a suspected pothole settles only the first of these points, since it fixes neither the depth of the defect, nor its location, nor the urgency of repair. Once a language model turns such boxes into fluent prose, an inspector can no longer see which sentences rest on evidence and which ones fill a gap. The same concern arises wherever structured records are rendered as text for a person who must act on them, so road inspection offers a concrete and checkable instance of a wider problem.

Two sources of error meet in such a report, since the detector may overlook defects, especially small ones lying next to a detected neighbour, while the written description may claim more than the predictions justify. Detection benchmarks score boxes, while attribution metrics score whether generated text agrees with its source (Gao et al., 2023, Min et al., 2023). In our view the missing piece lies between these two lines of work: a report should change in exactly the places where its supporting evidence changes, because inspection records are routinely corrected, enriched, or withdrawn after the first pass. Agreement with a frozen source cannot express this requirement, and neither can a hash over the whole record, which invalidates every sentence after any edit. We therefore separate

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.

Table 1: Each part of RoadTrust addresses a separate question on the path from image to report, and every finding is bounded by the collections, records, and raters described in Section 4.

Component	Question addressed	Finding
Detector selection and per-collection training	How complete is the visual evidence?	Every collection retains many missed objects, and misses gather in scenes with several defects
Object size and arrangement augmentation	Can training reduce clustered misses?	Wider scale variation recovers multi-defect objects, although count and size stay confounded
Evidence contracts and intervention audit	Does a report change exactly where its evidence changes?	Only per-statement dependencies pass every intervention, and the audit detects every non-equivalent injected fault
Writer-verifier route and human assessment	Can free prose and human raters check evidence?	Raters agree on claim support, while agreement on evidence correctness stays near chance

record fidelity, whether a sentence agrees with its record, from *physical correctness*, whether the record agrees with the road.

RoadTrust studies this distinction as a workflow that connects detector evaluation, targeted retraining, and evidence-based report generation (Figure 1). Candidate detectors are compared under one evaluator, and the selected architecture is then trained and tested separately within each of five image collections, while a controlled augmentation study probes how object size and arrangement shape the evidence available to a report. Each inspection record keeps its detections together with any paired depth reading and optional location tag. Statements about a record come from *evidence contracts*, which bind every sentence to the record identity, the fields it requires, and a canonical wording rule, and which are checked again whenever the record changes. An intervention audit confronts the contracts with altered wording, substituted records, withdrawn evidence, and unrelated edits, while a separate writer-verifier route with human assessment covers free-form prose.

Checking field values and hashing whole records each fail one intervention in our audit, and the required behaviour appears only when dependencies are declared for each statement (Table 1). For a restricted reporting language this withdrawal property follows directly from the construction, and we state it formally so that an extended audit, including deliberately faulty implementations, can test it. The human assessment points to a lesson that should hold for grounded generation in general: raters agree closely on whether a claim is supported, yet their agreement on whether its visual evidence is correct is barely separable from chance, largely because incorrect evidence almost never reached them. The detection experiments show that missed defects cluster in crowded scenes, so an empty or partial detection result cannot be read as a clean road. All of our guarantees concern consistency with supplied records, and confirming the road itself remains a task for field inspection.

2 Related work

Smartphone road-damage detection grew from a single-country collection (Maeda et al., 2018) into multinational benchmarks (Arya et al., 2021), culminating in the four-class RDD2022 dataset (Arya et al., 2024a). Challenge entries have raised accuracy through stronger detectors and ensembles (Arya et al., 2024b) and through inference optimizations for constrained hardware (Arya et al., 2024c). Newer collections extend the setting to dense scenes (Alkalbani et al., 2026), top-down imagery (Xiao et al., 2025), and safety-oriented urban pothole monitoring (Benedetti & Bortolotto, 2026). Published scores from these systems cannot be compared with results on our prepared splits, so we evaluate available detector families under recorded recipes and keep those experiments apart from reproductions of named challenge systems. Model selection belongs to the same comparison, and DomainBed (Gulrajani & Lopez-Paz, 2021) and WILDS (Koh et al., 2021) show how validation choices and distribution differences shape the gains a study reports.

Confidence calibration poses a related problem, because a confidence of 0.9 is useful as a probability only if comparable predictions are correct about 90% of the time (Guo et al., 2017). Selective prediction studies the trade-off that arises when uncertain outputs are withheld (Geifman & El-

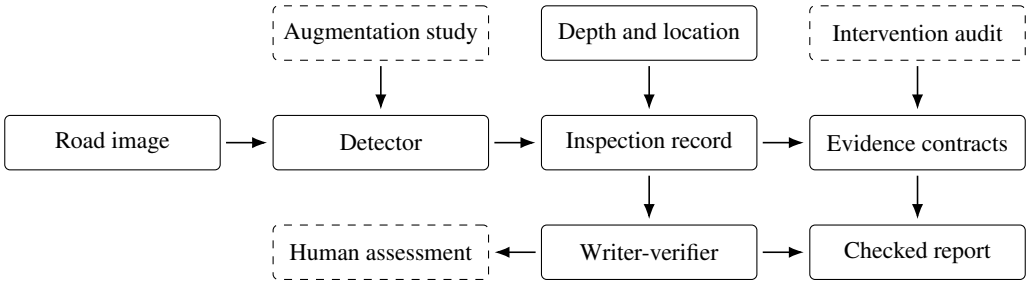


Figure 1: RoadTrust passes each road image through a detector into an inspection record that also holds optional depth and location, and statements reach the report through evidence contracts or through the writer-verifier route, with dashed boxes marking the studies that test each stage.

Yaniv, 2019). Conformal methods provide distribution-free versions of such guarantees for general risk control (Angelopoulos et al., 2024) and for object detection (Andeol et al., 2023). Our empirical threshold carries no conformal guarantee, and its behaviour must be measured on the population where it is used. We therefore report the threshold as an operating choice fitted on a small calibration partition, together with its measured effect on retained and missed candidates.

Data-to-text generation has long faced the same faithfulness problem, as neural summaries of basketball box scores frequently state facts that the source records do not contain (Wiseman et al., 2017), and PARENT therefore scores table-to-text output against the source table as well as the reference (Dhingra et al., 2019). ToTTo instead constrains generation by highlighting the cells a sentence must use (Parikh et al., 2020). For open-ended answers, ALCE separates answer correctness from citation quality (Gao et al., 2023), and FActScore evaluates individual factual propositions (Min et al., 2023). Agreement with a source does not establish visual truth, and a verdict from an LLM judge carries biases of its own (Zheng et al., 2023). All of these measures assess a finished text against a frozen source, whereas inspection records keep changing after the report is written.

Retracting conclusions when their support disappears is the defining behaviour of truth maintenance systems, which store the justification of each belief and withdraw beliefs whose justifications fail (Doyle, 1979), including under several competing assumption sets (de Kleer, 1986). Why-provenance identifies the source tuples responsible for a query answer (Buneman et al., 2001), and provenance semirings generalize that dependence to annotated relations (Green et al., 2007). Counterfactual evidence removal also appears in retrieval-augmented generation, where RAGONITE compares answers with and without retrieved evidence (Roy et al., 2024), while Proof-Carrying Numbers places verification in the renderer for numeric claims (Solatorio, 2025). RoadTrust applies justification tracking to rendered report sentences, binding each sentence jointly to its wording, its record identity, and its necessary fields, so that it survives edits to irrelevant fields. We do not claim to originate truth maintenance, provenance, counterfactual attribution, or proof-carrying output, and the present study does not measure superiority over those systems. Our contribution lies in adapting the binding to inspection reports and in an audit that shows where each simpler binding fails.

3 Evidence contracts for inspection reports

3.1 Records and permitted statements

An inspection record holds detector candidates and optional metadata, and each candidate carries a box, a predicted class, a confidence, and a review decision, while the metadata may add a location tag or a paired depth reading. A missing value is kept distinct from a verified measurement, so an available depth array alone cannot support a pothole depth in centimetres without a physical scale, alignment, and a road reference. The contract language currently has six rule families: suspected candidate, depth availability, missing location, supplied location tag, calibrated metric depth, and unassessed safety. Each rule specifies both the fields it needs and the conditions those fields must satisfy. A location tag records whether it describes the camera or the reported damage and whether it came from manual input or EXIF, and a metric-depth statement requires a finite, nonnegative value in metres together with explicit verification of scale, alignment, and road reference. Those

verification flags are trusted input assertions, since the renderer does not independently validate sensor calibration.

3.2 Generation, checking, and withdrawal

Let E be a record with identity i. Rule r declares a fixed set of field paths D_r, a validity condition g_r, and a deterministic text renderer t_r, and $E|_{D_r}$ denotes the values at those paths. If every required field is present and $g_r(E|_{D_r})$ holds, the compiler emits

$$c_r = \big(i, r, D_r, v_r, t_r(v_r), H(i, r, D_r, v_r)\big), \qquad v_r = E|_{D_r}, \tag{1}$$

where H is a canonical-content hash that in the implementation also covers the record kind. The hash is an audit commitment, not a signature authenticating who collected the evidence. Verification recomputes the rule, required fields, values, identity binding, and displayed sentence, so a certificate supplied by a caller cannot choose a shorter dependency list or attach a different sentence to valid numeric fields. Unsupported or malformed claims are withheld. When an existing record is revised, each previously issued claim is checked against the revision, and only the claims whose contracts still hold remain available.

Consider a record with a suspected pothole and a depth array. It may support two sentences: "The detector marks candidate 1 as a suspected pothole, retained for review" and "A paired depth array is available." Removing the depth-availability field withdraws the second sentence, whereas the first survives because it relies on different fields. Adding an operator note changes neither sentence. A new physical-depth statement becomes possible only after all of its additional prerequisites are supplied, and free-form descriptions of the same record follow the separate writer-verifier route evaluated in Section 5.4.

3.3 Withdrawal guarantee

Write $C(E)$ for the claims emitted by the compiler. Assume fixed rules, complete dependency declarations, a trusted record identity, deterministic rendering, and no hash collision. Let E^{-S} be the same record after deleting field set S, without changing other values. For a previously issued claim,

$$V(E^{-S}, c_r) = 1 \quad \Longleftrightarrow \quad D_r \cap S = \varnothing. \tag{2}$$

Deleting a required field prevents evaluation of the rule, and if no required field is deleted, the inputs, validity condition, identity, and rendered sentence are unchanged. This proves exact withdrawal of affected claims and retention of unaffected claims, and the same argument shows that changing fields outside D_r leaves the claim valid. Appendix A states the scope and the corresponding regeneration result.

These properties follow from the restricted language and do not certify arbitrary LLM prose, discover missing prerequisites automatically, or establish physical truth. Their practical value is that the reporting layer acquires behaviour an inspector or test harness can check directly. Completeness of each declaration with respect to wording holds by construction, because the renderer receives only the declared values and cannot read any other field. Completeness with respect to meaning does not hold automatically, since a declared decision field derived from a confidence score can become stale when the score is revised without recomputing the decision upstream. For unrestricted prose, RoadTrust retains the prompted writer-verifier route, whose judgments are evaluated empirically. The two routes are complementary, since the first offers exact guarantees for a narrow language and the second offers coverage of open wording without them.

4 Evaluation design

4.1 Visual evidence and model selection

Five image collections enter the main study: RDD2022 (Arya et al., 2024a) supplies the multi-national four-class benchmark, PothRGBD (Yurdakul & Taşdemir, 2025) is the only collection here with paired depth arrays, Chitholian (Rahman & Patel, 2020) and the multi-weather MWPD (Parvin et al., 2025) add independently collected pothole images, and HRP4K V1.01 (Chen,

2026) adds high-resolution scenes. Prepared training/validation/test counts are 28,331/3,721/5,663; 730/87/140; 2,234/61/101; 520/76/88; and 4,203/900/900 images, respectively, and the Chitholian training count includes 1,015 derived crops. Identity audits check exact images and derived copies across splits, but they do not establish road-segment or video independence, for which the available records lack identifiers. Appendix H preserves split history and attribution, including the earlier mixed Chitholian test pool.

YOLO26n (Ultralytics, 2026) is the selected architecture for the dataset-specific series after a two-stage RDD2022 comparison. We first train YOLOv5s, YOLO11n, YOLO12n, and YOLO26n for 12 epochs on the full RDD2022 training split and evaluate their validation-selected checkpoints with one common evaluator. AP50-95 determines the ranking, while parameter counts and recorded wall time describe the resource trade-off. We then allocate 12 additional epochs to the two highest-ranked models, as a 12+12 continuation from selected weights with a restarted optimizer rather than one uninterrupted 24-epoch schedule, and no test score enters architecture selection.

Every PothRGBD, Chitholian, MWPD, and HRP4K model starts afresh from the same official COCO-pretrained YOLO26n weights, not from an RDD2022-trained or another dataset's checkpoint. Each receives 24 epochs under the same recorded optimizer, loss, augmentation, batch size, resolution, and seed, and its own validation split selects the checkpoint before evaluation on its own test split. The RDD2022 model likewise starts from COCO weights but follows the 12+12-epoch selection protocol. The identity audit finds no train-test image reuse, and test labels do not select checkpoints. Because the fixed test sets were inspected in earlier work, they remain developmental rather than pristine holdouts. Earlier YOLO11n experiments and the five-person judgments retain their original checkpoint and stimulus identities, and neither is reassigned to the new models.

Detection quality is summarized by two complementary measures, where AP summarizes box matching across confidence thresholds and required overlap, and AP50-95 averages overlap requirements from 0.50 to 0.95. F1 combines precision, the fraction of predictions that match annotations, with recall, the fraction of annotated objects recovered. Fixed-threshold F1 uses confidence 0.25 and one-to-one matching at intersection-over-union 0.50. Scores on different collections are not a paired model comparison, because cameras, labels, scene distributions, and test sizes differ.

The controlled augmentation study uses an earlier YOLO11n PothRGBD detector and changes object scale and arrangement. Twelve runs compare control and joint composition augmentation at 12 and 24 epochs with three seeds, and nine additional runs vary mosaic, scale, or translation individually at 24 epochs. Checkpoint selection uses 43 validation images, and calibration uses a separate 44. All validation images contain one pothole, while 29 test images contain several, which makes density and object size potential sources of selection bias. Full recipes and per-seed results are retained in Appendices F and I.

4.2 Reporting and intervention tests

The archived reporting benchmark contains 40 images and 90 detector candidates. Two box-selection conditions each generate 130 inputs, namely 90 candidate records and 40 metadata records. The contract audit therefore uses 260 records from 40 unique images, not 260 independent scenes. The contract compiler produces 420 atomic statements from these records. Neither this experiment nor its synthetic edits adds new human annotations, so the audit tests implementation behaviour rather than annotator judgment.

We compare the earlier field-only verifier, a whole-record binding baseline, the complete contract, and two ablations removing text or record binding. The whole-record baseline binds the record and canonical text together, and it is a simple strong alternative when any record change should invalidate the entire output. Each method receives four paired conditions: an unchanged record, an unsupported replacement sentence, a changed record identity, and an added unrelated audit note. Correct behaviour means accepting the unchanged and unrelated-edit cases and rejecting the other two. We additionally delete every nonempty combination of the declared dependency fields and check both withdrawal and regeneration.

Each condition applies one templated perturbation per record, and the 420 archived statements contain only 13 distinct sentences from four of the six rule families, so we extend the audit in four directions. Cross-image swaps attach every archived claim to every record of the same kind from a

Table 2: Independent YOLO26n training and evaluation within each collection, where every test score uses a checkpoint selected on that collection's validation split and rows are descriptive rather than paired comparisons.

Dataset	Epochs	Test images	AP50-95	AP50	F1	Missed objects
RDD2022	12+12	5,663	22.41	46.62	49.49	4,818
PothRGBD	24	140	35.23	54.84	60.57	103
Chitholian	24	101	45.19	72.92	68.53	93
MWPD	24	88	28.88	57.56	57.31	105
HRP4K V1.01	24	900	15.82	31.56	35.23	704

different image. Field edits change every leaf value of every archived record one at a time, so that retention is expected exactly when the edited path lies outside the claim's declared dependencies. Text edits apply contradicting rewrites, unsupported additions, and meaning-preserving paraphrases to each claim. Property-based checks generate 10,000 records with boundary and malformed values across all six rule families and compare the compiler with a separately written reading of the rules in Section 3. A mutation test finally injects nine realistic implementation faults and asks whether any of these checks detects each fault.

The natural-language experiment uses separately prompted GPT-5.6 Sol writer and verifier contexts with archived inputs and outputs. Five human participants rate 160 claims from the 40 images under a protocol that distinguishes whether a statement is supported from whether its visual evidence is correct, and they do not rate the contract outputs. We report nominal Krippendorff's alpha, which adjusts agreement for category prevalence, with 5,000 image-cluster bootstrap samples. Separately, 20 human participants review 50 YOLO26n image cases without rating generated claims. The complete five-person protocol and pairwise matrices appear in Appendices L and O, and the scope of the visual review appears in Appendix D.

5 Results

5.1 Detector selection

YOLO26n leads the 12-epoch RDD2022 screening with 19.75% common-validation AP50-95 and the fewest parameters, ahead of YOLO12n, YOLO11n, and YOLOv5s (Appendix B). After the 12-epoch continuation, YOLO26n reaches 21.84% against 20.24% for YOLO12n and also leads AP50 and fixed-threshold F1, so it becomes the architecture of every dataset-specific model. On the 5,663-image RDD2022 developmental test, evaluated after selection and never used to choose between the finalists, the selected checkpoint obtains 22.41% AP50-95, 46.62% AP50, and 49.49% fixed-threshold micro F1. The historical YOLO11n system records a higher 24.52% AP50-95 on its sealed test evaluation after a longer and different training history, which shows that the common-validation winner is not automatically the highest-scoring deployed candidate. This conclusion is conditional on the prepared split, one seed, and the tested budgets, and recorded wall times describe observed cost rather than an architecture speed ranking. We keep both results visible because a report generated from either detector inherits its particular pattern of misses.

5.2 Missed damage and object arrangement

Each row of Table 2 uses its own training and test images, with AP and F1 in percent, four classes for RDD2022, and single-class potholes elsewhere. The differences therefore reflect datasets as well as learned parameters, and they do not show that training on one collection transfers to another. Every collection nonetheless leaves a substantial number of annotated objects undetected at the fixed operating point, from 93 on the Chitholian test to 4,818 on RDD2022. The PothRGBD, Chitholian, and MWPD tests were used during RoadTrust development and the HRP4K official test was inspected in an earlier diagnostic, so we do not call these tests previously unseen holdouts. HRP4K also lacks video identifiers, which prevents its official split from establishing video-level separation. These observations motivate keeping "the detector found a candidate" distinct from "the road has been fully inspected" in every report.

Table 3: Paired record-intervention audit on 260 inputs from 40 images, counting correct behaviour for constructed edits and swaps rather than for observed field errors.

Verifier	Clean	Text edit	Record swap	Irrelevant edit
Earlier field check	260/260	0/260	0/260	0/260
Whole-record binding	260/260	260/260	260/260	0/260
Without text binding	260/260	0/260	260/260	260/260
Without record binding	260/260	260/260	0/260	260/260
Full contract	260/260	260/260	260/260	260/260

Missed objects cluster in crowded scenes: with the earlier PothRGBD detector, 92 of 209 annotated potholes are missed, and the 29 multi-pothole images account for 82 of those misses, so 20.7% of test images contain 89.1% of missed objects. The same scenes contain smaller objects, since the median annotated pothole covers 14.74% of the frame in test against 48.47% in the single-pothole validation split, and object count alone is therefore not identified as the cause. The controlled augmentation study responds to this pattern: at 24 epochs, joint composition augmentation raises mean multi-pothole AP50-95 from 11.17% to 20.57%, and scale-only augmentation raises mean fixed-threshold F1 from 59.43% to 63.84% across three seeds (Appendix I). Neither result shows that a report lists every defect, and the incomplete factorial design does not identify interactions among augmentation settings. Object size and arrangement thus change the visual evidence a report receives, which is precisely what the reporting layer cannot see from the detections alone.

Score calibration does not remove this distinction, because on the 12-epoch reporting checkpoint of the augmentation study temperature scaling worsens the Brier score from 0.163 to 0.174 and expected calibration error from 7.96% to 12.78%. Because this monotone transform cannot improve ranking, the report uses an equivalent raw threshold of 0.3319, which retains 114 annotation matches and 48 non-matches, compared with 116 and 62 at confidence 0.25. The gate thus removes fourteen unmatched candidates at a cost of two matches. Its 90% precision target is empirical, fitted on 44 images, and not achieved as a guaranteed bound on the test population. Appendix F distinguishes this checkpoint from the one behind the earlier PothRGBD counts.

The independently trained PothRGBD YOLO26n model was also run through the reporting endpoint on the 40 fixed benchmark images. It matched 30 annotations, produced 14 unmatched predictions, and missed 39 annotated potholes (Appendix E). All responses preserved unknown location, metric depth, and safety status when evidence was absent. This check is a technical interface test on reused images, not a new human evaluation. Additional 20-epoch architecture comparisons on DRDD and grouped UAV-PDD2023 remain separate exploratory experiments in Appendix N.

5.3 Intervention audit

The earlier verifier accepts clean records but also accepts all 260 unsupported text replacements and all 260 identity swaps (Table 3). Its whole-evidence hash causes the opposite error on unrelated edits, rejecting every unchanged claim after an irrelevant note is added. Checking some field values and hashing the whole input therefore do not by themselves establish the intended relation between a sentence and its evidence. Whole-record binding blocks both attacks but rejects unrelated edits, which isolates the specific benefit of declaring dependencies separately. The full contract handles all four constructed conditions correctly, while removing text binding reproduces the text-edit failure and removing record binding reproduces the identity-swap failure. The complete contract also satisfies all 2,940 claim-level withdrawal checks and 1,820 record-level regeneration checks. Such perfect scores are the expected consequence of a finite rule system rather than evidence of language understanding, and the extended audit in Table 4 therefore asks a harder question, namely whether the checks would notice a contract that is wrong.

Real swaps between images show why identity binding matters beyond a synthetic suffix: 21.6% of candidate claims carry exactly the values of some candidate in another image, and every metadata claim does, so a value-only check would let a certificate issued for one image validate inside another image's report. The rendered sentence is identical in these cases, which makes the failure one of provenance rather than wording and explains why agreement-based checks cannot see it. Every leaf edit behaves as the dependency declaration predicts, and the generated records extend the same

Table 4: Extended audit of the full contract on the archived records and on 10,000 generated records, where the last column reports the weakest ablation or baseline for each test and paraphrase rejection measures the conservative cost of exact wording.

Stress test	Checks	Full contract	Weaker binding
Cross-image swap, candidates	31,256	100% rejected	21.6% pass a value-only check
Cross-image swap, metadata	18,720	100% rejected	100% pass a value-only check
Leaf edit outside dependencies	4,620	100% retained	0% kept by the earlier field check
Leaf edit inside dependencies	780	100% withdrawn	n/a
Contradicting rewrite	1,020	100% rejected	0% caught without text binding
Unsupported addition	1,260	100% rejected	0% caught without text binding
Meaning-preserving paraphrase	1,860	100% rejected	0% rejected without text binding
Generated records, six rules	1,093,509	100% as specified	n/a
Injected faults	9	8 detected	one equivalent mutant

checks to supplied geotags, zero and boundary coordinates, non-finite values, and calibrated metric depths that the archive lacks. The audit detects eight of the nine injected faults, including a guard that ignores the road reference, an off-by-one latitude bound, a missing flag read as false, and a renderer that reads an undeclared field. The surviving mutant lets the location guard accept non-finite coordinates, but the canonical hash independently rejects non-finite values, so that mutant never issues a claim and is equivalent to the correct implementation.

The same table exposes two costs of the design, since the contract rejects all 1,860 meaning-preserving paraphrases, including changes of case or spacing, so any wording outside the canonical renderer must pass through the writer-verifier route or a new rule. All 360 edits to candidate confidence scores also leave the corresponding "retained for review" claims valid, because the rules declare the decision field rather than the score from which the decision was derived. Both behaviours follow from the declared scope rather than from implementation errors, and we report them as design limits. The contract is integrated into the executable report serializer, while the earlier verifier is preserved as a snapshot for replay (Appendix C).

5.4 Human assessment

In the writer-verifier experiment, each box-selection condition produces 130 sentences, and the verifier flags one sentence containing two unsupported causal propositions. A separate 20-case control set contains ten supported statements and ten injected errors, all distinguished by that verifier. These observations establish behaviour on the archived examples, but filtering and rescoring with the same judge cannot establish zero hallucination risk, which leaves human judgment as the independent check on the prose route.

In the five-person assessment, claim support reaches 96.63% agreement and $\alpha = 0.931$, with interval [0.891,0.963]. Evidence correctness has similarly high raw agreement at 96.75%, but only $\alpha = 0.300$, with an interval from just below zero to 0.489. Almost all evidence labels are positive, with 781 of 800 marked yes, 19 unclear, and none no, so agreement says little about the rejection of incorrect visual evidence. With an interval that reaches chance level and no negative evidence judgments, the estimate is too uncertain to establish reliable discrimination of incorrect evidence. We preserve the submitted judgments and do not treat these ratings as an evaluation of the contract renderer. Claim support is thus judged consistently, whereas judging visual evidence correctness reliably needs a design that deliberately presents incorrect evidence to the raters. A separate round in which 20 participants reviewed 50 YOLO26n image cases is documented in Appendix D, but its centrally entered worksheets supply no per-person votes, so we draw no quantitative conclusion from it.

6 Discussion and conclusion

Any system that renders structured records as text faces the choice our audit exposes: checking field values misses substituted sentences and identities, while hashing whole records discards valid statements after harmless edits. Declaring dependencies for each statement resolves both failures, and the

formal statement together with the mutation test shows that this resolution follows from the declaration and can be checked, rather than depending on the particular rules we wrote. The human study adds a second point that transfers beyond roads, namely that an agreement coefficient on evidence correctness is informative only when raters actually encounter incorrect evidence, so assessments of grounded generation should enrich negative cases by design. We expect both observations to apply to clinical, financial, or asset-management reports assembled from structured records. The present study, however, tests them only on road-inspection records, and that extension remains a hypothesis.

The experiments do not directly compare RoadTrust with RAGONITE, Proof-Carrying Numbers, or FActScore, which provide conceptual context rather than measured baselines, and the prompted LLM verifier has not yet been run on the extended interventions, for which the supplement provides a stratified 300-item challenge set with hidden expected labels. The augmentation contrasts rest on three seeds and an incomplete factorial design, so they do not identify individual interactions or general causal effects. The contract rules are hand specified, authentic evidence is assumed, and the benchmark reuses 40 images, so the study does not establish a broadly novel learning algorithm. Exact wording is also conservative, since the contract rejected every meaning-preserving paraphrase we tested, and semantic dependencies such as a decision derived from a confidence score must still be declared by the rule designer. Stronger generalization evidence would require unseen claim families, independent natural-language verifiers, and errors constructed by assessors who did not design the rules. Learning the dependency structure from data is a natural research direction rather than a completed result.

For municipal use, the next validation step is a road- or session-disjoint collection with verified location, calibrated depth, and inspector adjudication. A counterbalanced user study could compare detector overlays with evidence-dependent reports on missed defects, incorrect claims, appropriate deferral, and review time. Open311 or GIS integration could then test whether records link correctly to road assets (Open311, n.d.), while spatial differences in resident reporting would still need consideration (Liu et al., 2024). RoadTrust contributes a tested way to preserve and revise what an inspection system can support, and its benefit to inspectors and maintenance planning is left to that field evaluation.

AI USE STATEMENT

The authors made the final research decisions and take responsibility for the paper and supplement; no language model produced or imputed a human rating. OpenAI Codex with GPT-5.6 Sol and GPT-6 and Claude (Anthropic) assisted with developing the evidence-contract framework, formulating its mathematical claims and writing proofs, designing the audits and synthetic-data generation procedures, implementing methods, and interpreting results, as well as drafting, translation, and literature suggestions. GPT-5.6 Sol also served as the prompted writer and verifier studied in Section 5.4, with prompts and raw outputs archived as study data. Verification included checking reported results against archived records, running `run.py verify`, `replay`, and `audit` for artifact integrity and reporting audits, and checking bibliographic details of newly added references against published records; these checks do not independently establish the proofs or rederive all detector results or human-study intervals from raw data.

ETHICS STATEMENT

The intended application supports inspection with human review. Detection scores do not establish road safety or authorize maintenance. The rubric study collected coded judgments from five human participants; a separate visual review involved 20 human participants. The paper reports no participant identities or review times. A prospective field study must protect identifiable people and vehicles and examine uneven performance across acquisition conditions. Geotags remain local until a user exports the report. Shared records require appropriate location privacy controls. No municipal deployment or inspector field study is reported.

REPRODUCIBILITY STATEMENT

The supplement supports artifact verification, replay from archived reporting records, and inference with supplied checkpoints, not end-to-end re-derivation from raw data. Raw images, depth arrays,

prepared split files, and participant workbooks are not included; `external_data/` paths identify external dependencies, not bundled files. Hashes establish content consistency, not training provenance or independent replication. For the five-rater study, aggregate pairwise matrices suffice to recompute point estimates of nominal alpha, but not image-bootstrap intervals or workbook-to-label extraction. Workbook hashes do not remove this limitation. Image-bootstrap intervals condition on fixed models; the retraining study separately reports three-seed variation. The extended contract audit and the mutation test run on CPU through `python run.py audit`, which regenerates the results reported in Table 4 from archived records and seeded synthetic records. A stratified LLM-verifier challenge set and its scoring script are included in `data/evidence_contracts_v2`. Appendix N documents experimental extensions, and Appendix O documents archived reporting outputs and replay. No independent third-party training replication is reported.

REFERENCES

Asma Musabah Alkalbani, Muhammad Saqib, Ahmed Salim Alrawahi, Abbas Anwar, Chandranath Adak, and Saeed Anwar. RDD4D: 4d-attention-guided road damage detection and classification. *IEEE Access*, 14:40078–40092, 2026. doi: 10.1109/ACCESS.2026.3666237. URL https://doi.org/10.1109/ACCESS.2026.3666237.

Leo Andeol, Thomas Fel, Florence de Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In Harris Papadopoulos, Khuong An Nguyen, Henrik Boström, and Lars Carlsson (eds.), *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 36–55. PMLR, 13–15 Sep 2023. URL https://proceedings.mlr.press/v204/andeol23a.html.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, Alexander Mraz, Takehiro Kashiyama, and Yoshihide Sekimoto. Deep learning-based road damage detection and classification for multiple countries. *Automation in Construction*, 132:103935, 2021. doi: 10.1016/j.autcon.2021.103935. URL https://doi.org/10.1016/j.autcon.2021.103935.

Deeksha Arya, Hiroya Maeda, Sanjay Kumar Ghosh, Durga Toshniwal, and Yoshihide Sekimoto. RDD2022: A multi-national image dataset for automatic road damage detection. *Geoscience Data Journal*, 2024a. doi: 10.1002/gdj3.260. URL https://doi.org/10.1002/gdj3.260.

Deeksha Arya, Hiroya Maeda, and Yoshihide Sekimoto. From global challenges to local solutions: A review of cross-country collaborations and winning strategies in road damage detection. *Advanced Engineering Informatics*, 60:102388, 2024b. doi: 10.1016/j.aei.2024.102388. URL https://doi.org/10.1016/j.aei.2024.102388.

Deeksha Arya, Hiroshi Omata, Hiroya Maeda, and Yoshihide Sekimoto. ORDDC'2024: State of the art solutions for optimized road damage detection. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 8430–8438, 2024c. doi: 10.1109/BigData62323.2024.10825254. URL https://doi.org/10.1109/BigData62323.2024.10825254.

Roberto Benedetti and Valerio Bortolotto. BS Dataset: A tailor-made urban road pothole dataset for real-time detection and safety-oriented monitoring. *Sensors*, 26(16):5267, 2026. doi: 10.3390/s26165267. URL https://doi.org/10.3390/s26165267.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *Database Theory, ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pp. 316–330, 2001.

Hanshen Chen. HRP4K. Zenodo dataset, version 1.01, 2026. URL https://zenodo.org/records/21941931.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4884–4895, 2019.

Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 31–40, 2007. doi: 10.1145/1265530.1265535. URL https://www.cs.ucdavis.edu/~green/papers/pods07.pdf.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf?id=lQdXeXDoWtI.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11. Software, version 11.0.0, 2024. URL https://github.com/ultralytics/ultralytics.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804. URL https://doi.org/10.1016/j.patter.2023.100804.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, 4 edition, 2018. URL https://us2.sagepub.com/en-us/nam/content-analysis/book258450.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html.

Zhi Liu, Uma Bhandaram, and Nikhil Garg. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 4:57–65, 2024. doi: 10.1038/s43588-023-00572-6. URL https://www.nature.com/articles/s43588-023-00572-6.

Hiroya Maeda, Yoshihide Sekimoto, Toshikazu Seto, Takehiro Kashiyama, and Hiroshi Omata. Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images. *Computer-Aided Civil and Infrastructure Engineering*, 33(12):1127–1141, 2018. doi: 10.1111/mice.12387. URL https://doi.org/10.1111/mice.12387.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Open311. GeoReport v2. Official API specification, n.d. URL https://wiki.open311.org/GeoReport_v2/. Accessed September 19, 2026.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1173–1186, 2020.

Shahnaj Parvin, Foysal Munsy, Md Tanzeem Rahat, Aminun Nahar, Kamruddin Nur, and Debasish Ghose. Robust multi-weather pothole detection: An enhanced YOLOv9 trained on the MWPD dataset. *Results in Engineering*, 28:107817, 2025. doi: 10.1016/j.rineng.2025.107817. URL https://doi.org/10.1016/j.rineng.2025.107817.

Atikur Rahman and Sachin Patel. Annotated Potholes Image Dataset. Kaggle, dataset, version 1, 2020. URL https://doi.org/10.34740/kaggle/dsv/973710.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. Evidence contextualization and counterfactual attribution for conversational QA over heterogeneous data with RAG systems. *arXiv preprint arXiv:2412.10571*, 2024. URL https://arxiv.org/abs/2412.10571. Accepted at WSDM 2025.

Aivin V. Solatorio. Proof-carrying numbers (PCN): A protocol for trustworthy numeric answers from LLMs via claim verification. *arXiv preprint arXiv:2509.06902*, 2025. URL https://arxiv.org/abs/2509.06902.

Ultralytics. Ultralytics YOLO26. Official model documentation, 2026. URL https://docs.ultralytics.com/models/yolo26/. Accessed 24 September 2026; experiments use the recorded local software and checkpoint hashes.

Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2263, 2017.

Xi Xiao, Zhengji Li, Wentao Wang, Jiacheng Xie, Houjie Lin, Swalpa Kumar Roy, Tianyang Wang, and Min Xu. TD-RD: A top-down benchmark with real-time framework for road damage detection. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888616. URL https://doi.org/10.1109/ICASSP49660.2025.10888616.

Mustafa Yurdakul and Şakir Taşdemir. PothRGBD: RGB and Depth Images of Potholes. IEEE DataPort, dataset, 2025. URL https://doi.org/10.21227/z8eq-sf60.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A EVIDENCE-CONTRACT SPECIFICATION AND PROOF

The contract implementation uses a finite set of rule identifiers, each mapped to an immutable dependency list. A claim is issued only when every dependency resolves and the rule's guard accepts its values. Verification reconstructs that list from the rule identifier, checks the values against the current record, checks record identity, and recomputes the displayed sentence. This prevents a caller from omitting a prerequisite or attaching arbitrary wording to a valid payload.

Withdrawal property. Fix a supported claim $c_r \in C(E)$ and delete a field set S without changing the record identity or other values. If $D_r \cap S \neq \varnothing$, at least one required lookup fails, so verification rejects. Otherwise all lookups return their original values. The guard, typed payload, and canonical sentence are identical, so verification accepts. This proves Equation 2. The argument assumes all dependencies are declared, fixed rules, deterministic serialization, a trusted identity, and no hash collision; it does not establish that an input measurement is accurate.

Regeneration property. Every guard in this implementation requires its declared fields to be present; absence is never interpreted as a positive fact. Under field deletion alone, a rule that was previously unavailable cannot become available. Therefore

$$C(E^{-S}) = \{c \in C(E) : V(E^{-S}, c) = 1\}.$$

An explicit unknown status can support a sentence describing missing information, but deleting that status removes the sentence rather than manufacturing an unknown value. Field replacement or addition can create new claims, so the displayed equality is restricted to deletion.

Boundary of the guarantee. A caller able to forge the trusted input record can make a false physical statement consistent with that forged record. Hashes provide content commitments, not authenticated acquisition. A missing prerequisite in the rule design also invalidates the intended semantic guarantee. Free-text paraphrases are outside the certified language: changing a canonical sentence requires a new rule or an independently assessed language interface.

Audit protocol. The four archived input files yield 260 record-condition pairs from 40 image identities and 420 atomic claims. Each pair receives the same text replacement, identity replacement, and unrelated-note addition. Deletion tests enumerate all nonempty subsets of declared dependency paths for that record. The earlier verifier is archived before repair. The whole-record baseline binds both input and output, while ablations disable one contract check at a time. Source hashes, certificates, counts, and generated tables are retained in `review_experiments/evidence_contracts_v1`. The experiment runs on CPU without detector retraining or an external model API.

The current rules have at most five field dependencies; checking all claims is linear in the number of dependencies visited, excluding the size of serialized field values. Exhaustive subset enumeration is used only for this audit and grows exponentially with the number of distinct paths. Deployment verification does not enumerate subsets: it rechecks each contract directly.

B DETECTOR SELECTION DETAILS

Table 5: Twelve-epoch RDD2022 screening on 3,721 validation images, where wall time records the whole process including evaluation and interruptions rather than isolated training throughput.

Model	AP50-95 (%)	F1 (%)	Params (M)	Wall time (h)
YOLOv5s	16.71	45.23	7.021	17.74
YOLO11n	17.55	44.13	2.591	25.27
YOLO12n	18.30	44.05	2.569	4.98
YOLO26n	19.75	47.27	2.505	5.58

Table 5 lists the 12-epoch screening. YOLO12n records a shorter process duration than YOLO26n, whereas YOLO26n offers higher localization accuracy and a slightly smaller model. Because pro-

cess duration includes interruptions and concurrent workload, these timings describe observed cost rather than an architecture speed ranking. The accuracy-first rule therefore advances YOLO26n and YOLO12n, without a retrospectively fitted accuracy-time score. The native trainer's validation scores use a different evaluation path and are not substituted for the common-evaluator values in either table.

Table 6: Top-two continuation with 12 additional epochs from each selected 12-epoch checkpoint under a restarted optimizer, scored by the same common validation evaluator.

Model	AP50-95 (%)	AP50 (%)	F1 (%)	Added time (h)
YOLO26n	21.84	47.01	50.34	3.80
YOLO12n	20.24	43.74	48.26	4.68

C Extended audit and mutation test

The extended audit uses the same 260 archived records as Table 3. Cross-image swaps pair each claim with every record of the same kind whose image identifier differs, which excludes the two box-selection conditions of one image because their metadata records are copies of each other. Leaf edits perturb one field at a time, flipping booleans, incrementing integers, scaling floats, and appending a marker to strings, and a claim is expected to survive exactly when the edited path neither equals nor lies under nor above one of its declared paths. Text edits comprise three contradicting or strengthening rewrites for candidate claims and two for each metadata rule, three unsupported additions per claim, and between four and five meaning-preserving variants per claim, including case, spacing, and synonym changes. The generated records use seed 391 and mix valid values with boundary coordinates, non-finite numbers, booleans in numeric fields, unverified calibration flags, and randomly deleted fields, which yields 1,891 location-tag and 1,490 metric-depth claims across 10,000 records.

Table 7: Injected implementation faults and the audit checks that fail under each one, where the only surviving fault is blocked by the canonical hash before any claim is issued.

Injected fault	Patched component	Failing checks
Metric guard ignores road reference	renderer	specification
Latitude bound off by one degree	renderer	specification
Non-finite coordinates accepted by guard	renderer	none (equivalent)
Boolean accepted as metric depth	renderer	specification
Missing depth flag read as false	field lookup	specification, withdrawal, regeneration
Record identity not checked	verifier	record swap
Sentence text not checked	verifier	text edit
Support hash covers whole record	issuer	specification, withdrawal, outside edit
Text reads an undeclared field	issuer	outside edit

The specification used in the property checks was written from the rule descriptions in Section 3 without calling the compiler, so disagreement signals a gap between the stated rules and their implementation. The same author wrote both, and the check therefore cannot rule out a shared misreading of the intended rules. The mutation test patches one component at a time at run time and reruns a compact version of every check on the archived records and on 3,000 generated records. A fault counts as detected when any check falls below 100%. Scripts `run_extended_contract_audit.py` and `run_contract_mutation_test.py` write their outputs to `data/evidence_contracts_v2`.

D Methodology checks

These follow-ups use the historical YOLO11n checkpoints and are retained separately from the new YOLO26n experiments.

Calibration sensitivity. Five hundred image-disjoint resamples of the 87 validation images use 61 images for calibration and 26 for evaluation. This increases the number of split perturbations, not the number of independent images. The checkpoint was selected using overlapping validation data, so this is a conditional sensitivity analysis, not a fully nested generalization estimate.

Table 8: Calibration-rule sensitivity reported as mean percentages of precision, candidate coverage, and annotated-object recall across dependent split perturbations, where Wilson-rule precision is defined only for resamples in which that rule selects a gate.

Rule	Gate selected (%)	Precision (%)	Coverage (%)	Recall (%)
Absolute minimum 10	100.0	90.73	75.42	96.88
Relative minimum 10%	100.0	90.73	75.42	96.88
Wilson lower bound	50.2	93.79	32.15	43.12

The relative rule removes the fixed candidate-count dependence but selects the same gates in these resamples. The Wilson rule raises conditional precision while withholding a gate in nearly half the resamples and sharply reducing annotated-object recall. It is not a distribution-free guarantee after threshold search or under image clustering. The selected YOLO26n service uses a fixed raw operating point of 0.25, with no probability-calibration claim and no inherited YOLO11n threshold.

Density-aware selection pilot. A separate historical PothRGBD-to-UAV pilot trained three 24-epoch YOLO11n runs and selected checkpoints using 43 single-pothole PothRGBD images and 20 positive UAV images, including six multi-pothole scenes. The selected checkpoints reached 35.34% to 38.34% AP50-95 on the developmental PothRGBD test, but none matched an annotated pothole at the fixed operating point on the 374-image UAV test. The mixed-domain selection score therefore did not repair this perspective shift. This negative result is not part of the new dataset-specific YOLO26n comparison and does not replace a prospective road-disjoint holdout.

Additional visual review. Twenty human participants took part in a separate review of 50 YOLO26n image cases. The frozen package samples ten cases from each of five collections, using annotation-count strata rather than model outcomes: two images without an annotated pothole, four with one, and four with at least two where available. Reviewers saw each original image beside its predicted boxes at confidence 0.25. The package contains 44 candidate boxes across the 50 cases. For the HRP4K cases only, the displayed boxes come from an RDD2022-trained checkpoint evaluated zero-shot; the main HRP4K detector table instead uses a separately trained in-domain model. The available review workbooks were entered centrally, and the analysis archive does not preserve each participant's original submission. Participation is therefore reported, but workbook-level label agreement is not treated as 20 independent votes or as a detector-accuracy estimate. No review-time or personal-identity data are analyzed. The five-person quantitative assessment remains tied to its original YOLO11n text stimuli; an independent agreement analysis of the YOLO26n cases would require the individual submissions.

E Independent PothRGBD YOLO26n service regression test

This run uses checkpoint SHA-256 prefix `5c6405d0f7ef`. The historical HTTP results in Appendix F use YOLO11n checkpoint `cce85af87abe`; their counts and latencies belong to that separate execution. The reporting endpoint now defaults to a validation-selected PothRGBD YOLO26n checkpoint initialized directly from the official COCO weights. It did not inherit an RDD2022-trained checkpoint. We exercised the endpoint on the same 40 fixed technical-benchmark images after locking the checkpoint and raw-score threshold of 0.25. The benchmark images were used in earlier RoadTrust development, so this is an interface regression test rather than a new generalization sample.

All 40 requests returned reports bound to the same checkpoint hash. The warm-request median was 0.139 s and the 95th percentile was 0.179 s on the local CPU service while other training jobs were active; these values are not a controlled throughput benchmark. The assertions found no invented location, metric depth, or safety assurance in the 40 responses. The detector still missed 39 annotated

Table 9: Measured HTTP-service results for the independently trained PothRGBD checkpoint, where counts match predicted boxes to image annotations at the fixed operating point and do not represent human judgments of report correctness.

Checkpoint	TP	FP	FN	Recall (%)
PothRGBD in-domain	30	14	39	43.48

Table 10: The prepared data inventory summarizes the subsets used in this study, while the RDD2022 checkpoint comes from an earlier stage and the reported crops remain training entries rather than separate scenes.

Dataset	Original train	Crops	Validation	Test	Evaluated classes
RDD2022	28,331	0	3,721	5,663	Four damage classes
PothRGBD	730	0	87	140	Pothole
Chitholian	1,219	1,015	61	101	Pothole, core subset
MWPD	520	0	76	88	Pothole
HRP4K V1.01	4,203	0	900	900	Pothole

objects, and annotation matching does not independently verify the prose or the physical road. The five-person assessment used older YOLO11n stimuli and is not attributed to this checkpoint.

F Extended detection and reporting study

F.1 Data and evaluation method

F.1.1 Datasets, preparation, and identity checks

The original RoadTrust benchmark contained four prepared collections (Table 10); the new independent-training series also evaluates HRP4K V1.01. RDD2022 covers four damage classes, whereas the others are evaluated for potholes only. The counts refer to the subsets used here, not the complete source archives. This distinction is particularly important for Chitholian: an earlier 226-image test pool combined 101 core images with 125 auxiliary images from a different image family. We restrict the dataset-specific test result to the 101 core images so that its evaluation population is explicit, while retaining core images, auxiliary images, and derived crops in training. Appendix H documents the earlier mixed-pool evaluation.

DRDD, UAV-PDD2023, and RoadSens-4M are used only in the additional external evaluations in Appendix N. They are not pooled into the four-collection main benchmark or used to revise its reported checkpoints.

The identity audit compares bytes, decoded content, original stems, and confirmed near-duplicates. It finds no exact split overlap, confirmed derived-image overlap with validation/test, or test image in training records. Removing 22 confirmed near-duplicates reduces the earlier RDD2022 test from 5,685 to 5,663 images. This checks image reuse, not street independence: geographic metadata is unavailable (Appendix H).

For HRP4K, we retain the author-provided 4,203/900/900 image folders. A separate image-only audit found no exact train-validation duplicate. Two train-validation pairs had 64-bit perceptual-hash distance four; visual inspection of their original 4K images found different streets and traffic structures, and the decisions are archived. An earlier audit found no exact overlap of its test images with the training and validation images. Neither check establishes video- or road-disjoint sampling, because the archive has no corresponding group identifiers. The official test split had already been used in a frozen-checkpoint diagnostic before this independent training run, so it is not described as untouched.

F.1.2 DETECTOR STATUS AND CHECKPOINT PROTOCOL

The selected RDD2022 detector architecture is YOLO26n with 640-pixel inputs. Historical experiments in this appendix use YOLO11n (Jocher & Qiu, 2024) unless explicitly identified as the new YOLO26n series. Each prediction contains a bounding box, a damage category, and a confidence score. RDD2022 evaluation includes four damage classes; the other dataset-specific evaluations concern potholes only. We do not transfer predictions or class scores across datasets for the new comparison.

The completed 12-epoch comparison supersedes the earlier five-epoch screening rationale (Table 5). The conditional fourth run was YOLO26n because YOLO12n exceeded YOLO11n on common-validation AP50-95; YOLOv8n was the unselected alternative branch. Public pretrained initializations, checkpoint hashes, model sizes, and observed process durations are retained in the selection manifest.

YOLO26n and YOLO12n advance by the recorded AP50-95 ranking. Both complete a separate 12-epoch continuation from their selected screening weights, with a restarted optimizer and a lower initial learning rate. YOLO26n wins the final common-validation comparison (Table 6). The budget is 24 executed training epochs per selected architecture; a best checkpoint may come from an earlier epoch within either stage. This distinction prevents interpreting the experiment as an uninterrupted optimizer resume or as a convergence guarantee.

The historical target-trained systems use validation images to select their saved checkpoints, but their training recipes differ. The final PothRGBD, Chitholian, and MWPD stages use AdamW, batch size 8, and seed 391, with respective budgets of 34, 26, and 19 epochs and learning rates of 8×10^{-5}, 10^{-4}, and 6×10^{-5}. The reported RDD2022 system instead follows 40 initial epochs and a 12-epoch localization stage; epoch 11 of that stage provides the highest validation score. A continuation to 15 epochs does not improve that score and is excluded. Consequently, the historical matched evaluation compares complete trained systems, not the effect of a shared fine-tuning recipe. Appendix G provides the remaining settings.

The new YOLO26n series instead initializes each PothRGBD, Chitholian, MWPD, and HRP4K run from the same COCO-pretrained weights. It trains only on that collection's own training images for 24 epochs, selects weights using that collection's own validation images, and evaluates only that collection's own test images. The RDD2022 model uses its own training split and the 12+12-epoch architecture-selection protocol. The shared initial weight hash, optimizer settings, and split manifests are retained with each run. This removes the RDD-to-target initialization from the main result; it does not make test sets that were used historically pristine or establish road-level split independence.

F.1.3 DATASET-SPECIFIC EVALUATION RULES

Average precision (AP) summarizes whether predicted boxes match annotated damage across score thresholds and required box overlaps. We compute it from confidence 0.001, after non-maximum suppression removes duplicate boxes at intersection-over-union (IoU) 0.70, with at most 300 detections per image. AP averages IoU requirements from 0.50 to 0.95. Precision, recall, and F1 instead use one fixed operating point: confidence 0.25 and one-to-one matching at IoU 0.50. RDD2022 pools counts over four classes. The other datasets evaluate potholes only. We do not mix AP-optimized thresholds with these fixed-threshold counts.

Each test model is evaluated on its own collection, with a common image size, preprocessing, duplicate-box removal, and matching implementation. The evaluation records AP, fixed-threshold counts, the selected checkpoint hash, and the split-file hashes. Since no two rows share a test population, subtracting their AP scores would not estimate an architecture or data-composition effect. Neighboring captures may remain related despite distinct file identities. Only three PothRGBD timestamp-hour groups are available, too few for reliable time-group analysis.

F.1.4 EQUAL-BUDGET RETRAINING AND SCORE CALIBRATION

Images containing several potholes account for most missed objects, so we test a training change aimed at a wider range of object sizes and arrangements. Both arms start from the same RDD2022 checkpoint and train on the same 730 images at budgets of 12 and 24 epochs, each repeated at seeds

Table 11: Historical YOLO11n checkpoint results retained with their original provenance, with all scores in percent and precision and recall taken at confidence 0.25, so these rows are not YOLO26n measurements.

Dataset	Images	mAP50	mAP50-95	P	R	F1
RDD2022	5663	50.28	24.52	60.00	47.19	52.83
PothRGBD	140	61.33	39.83	77.48	55.98	65.00
Chitholian	101	82.96	49.66	72.05	76.25	74.09
MWPD	88	73.19	36.66	64.81	69.65	67.15

391, 392, and 393. The two budgets start independently from the same source weights. The longer budget also uses its own learning-rate schedule. Mosaic augmentation combines several training images into one. Scaling changes object size, and translation changes position. The control uses mosaic probability 0.2, scale amplitude 0.2, and translation amplitude 0.05. Composition augmentation uses 0.8, 0.5, and 0.1. This is a joint augmentation intervention, so its components are not individually identified. The two arms share the detector, optimizer, loss, and epoch budget within each comparison.

The original 87 validation images are partitioned deterministically into 43 checkpoint-selection and 44 calibration images, with no identity-component overlap. Each run chooses its checkpoint on selection mAP50-95. The highest selection score across six runs within each budget determines its chosen checkpoint before test evaluation. Every validation image contains exactly one annotated pothole, whereas 29 test images contain several. Checkpoint selection can therefore favor the simpler single-pothole distribution; the observed multi-pothole deficit is confounded with this split mismatch rather than being attributable to object count alone. Historical test diagnostics motivated the recipe, and the 140-image test is a developmental benchmark, not a newly acquired population.

Candidate correctness for reporting uses confidence-descending one-to-one annotation matching at IoU ≥ 0.5, separately from the standard AP evaluator. A score floor of 0.05 defines the candidate population. We evaluate temperature scaling (Guo et al., 2017) as a diagnostic, not as a mandatory reporting step. It fits $T \in [0.25, 4]$ by binary negative log-likelihood on calibration candidates:

$$q(s, T) = \sigma\left(\frac{\log(s/(1-s))}{T}\right). \tag{3}$$

Here, s is the original confidence, T controls its adjustment, and q is the adjusted score. σ maps the result to the interval from zero to one. This operation does not change the boxes or their confidence ranking.

We then choose a raw-score report threshold that keeps as many matched calibration candidates as possible while reaching at least 90% precision and retaining at least ten candidates. If no threshold meets both conditions, every candidate is sent for further review. The 90% target applies only to the calibration images and is not promised on test images. We also apply the same selection rule to the adjusted scores. Because temperature scaling preserves rank, an equivalent raw threshold keeps exactly the same candidates. A non-identity calibration transform would be deployed only after a prespecified held-out check shows improved reliability; the present study does not meet that condition.

F.2 Results and interpretation

F.2.1 Historical detection diagnostics

The historical YOLO11n results show why aggregate detection scores need to be read alongside missed-object counts (Table 11). PothRGBD has the highest precision at confidence 0.25, yet its model misses 92 of 209 annotated potholes; the corresponding counts are 57 of 240 for Chitholian and 61 of 201 for MWPD. Class-level results reveal a related distinction: RDD2022 has a micro F1 of 52.83%, but pothole recall is only 35.22%, compared with 57.95% for alligator cracks. Thus, neither high precision nor a pooled score justifies interpreting an empty detection result as the absence of damage.

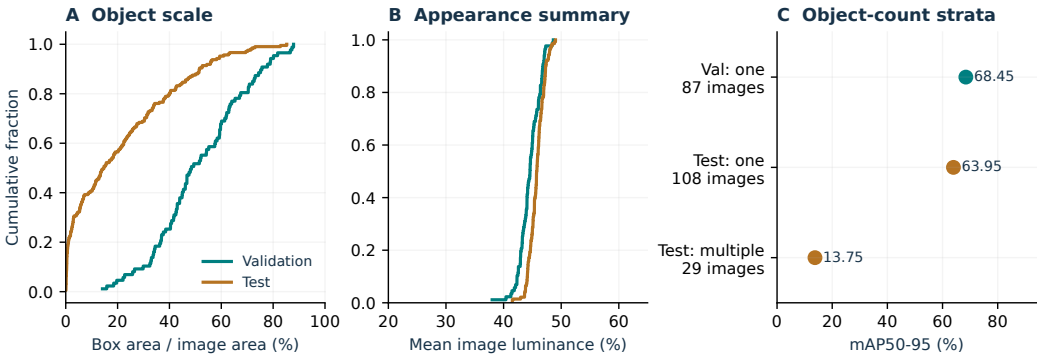


Figure 2: PothRGBD split diagnostics with box area weighted by annotation and luminance weighted by image, where the right panel compares single- and multi-pothole scenes without the three background images and does not isolate the cause of the performance gap.

The historical YOLO11n scores above are retained only to explain earlier observations about missed objects and their reporting consequences. Those systems have a different initialization and training history from the new independent YOLO26n series. Their performance must not be relabeled as the result of training each dataset from the same COCO weights.

F.2.2 Concentration of missed damage in PothRGBD

The historical YOLO11n PothRGBD checkpoint has an archived native Ultralytics validation score of 68.82% mAP50-95. Re-evaluating that checkpoint with the common diagnostic evaluator gives 68.45% on validation and 39.83% on test. The validation-test comparison uses only these latter scores, obtained with matching diagnostic settings; 68.82% belongs to the trainer's separate evaluation path. We use the common-evaluator gap to examine differences between the image groups.

The two splits differ in both object count and object size. All 87 validation images contain exactly one annotated pothole; the test set contains 108 single-pothole images, 29 multi-pothole images, and three backgrounds. The median annotated pothole also occupies less of the frame in test than in validation: 14.74% versus 48.47%. These differences mean that validation-based checkpoint selection favors a population with larger, isolated defects rather than the full range of test scenes.

Brightness provides less evidence of a split difference. Its medians are 44.58% and 45.75% of the 8-bit range, and the distributions overlap (Figure 2). Object scale and scene composition therefore motivate the subsequent diagnostics, but these descriptive comparisons do not isolate their causal effects. Weather, lighting angle, and partial occlusion remain unmeasured.

The contrast becomes clear when the groups are scored separately. Single-pothole images reach 63.95% mAP50-95, while multi-pothole images reach only 13.75%. Those 29 images contain 82 of the 92 missed annotations. Put differently, 20.7% of the test images contain 89.1% of the misses. This matters in practice: showing one detected pothole can hide the fact that a second or third nearby defect was overlooked. The groups were defined after examining the data, and object count is mixed with object size and capture order. We use the pattern to motivate the training experiment, not to claim that object count alone caused the errors.

F.2.3 Measured response to composition augmentation

Joint composition augmentation increases mean mAP50-95 by 2.36 percentage points at 12 epochs and 3.80 points at 24 epochs (Table 12). To examine which changes contribute to this result, nine additional 24-epoch runs vary mosaic, scale, or translation individually. Their paired F1 changes are -0.32, $+4.42$, and $+1.35$ percentage points, respectively. Scale alone gives the largest observed standalone F1 gain, reaches 63.84%, and improves every seed.

The preferred configuration nevertheless depends on the outcome being measured. Joint augmentation has a slightly lower overall F1 of 63.37%, but a higher mean multi-pothole AP of 20.57%,

Table 12: The PothRGBD retraining results report mean ± sample standard deviation across three seeds, while all metrics are percentages and Multi denotes the 29 multi-pothole images.

Epochs	Augmentation	mAP50-95	Multi mAP50-95	F1 at 0.25
12	Control	33.94 ± 0.78	11.32 ± 0.78	58.57 ± 1.25
12	Composition	36.30 ± 0.93	18.81 ± 1.25	60.89 ± 1.76
24	Control	34.69 ± 0.38	11.17 ± 0.48	59.43 ± 1.76
24	Mosaic only	34.48 ± 1.13	12.55 ± 1.55	59.10 ± 1.47
24	Scale only	39.75 ± 0.28	19.26 ± 0.95	63.84 ± 0.10
24	Translation only	34.43 ± 0.72	11.91 ± 1.33	60.78 ± 0.79
24	Composition	38.49 ± 1.48	20.57 ± 0.53	63.37 ± 0.44

compared with 19.26% for scale alone. The single-factor results describe changes from the control settings and do not identify all interactions among augmentation parameters. Appendix M reports the paired variation and the departure from additive effects.

The frozen application benchmark retains the 12-epoch checkpoint so that its detector, gate, and measured requests remain one auditable historical system. This is not a claim that it is the strongest detector in the study. The 24-epoch scale-only configuration has higher mean test F1, but that finding emerged from the developmental test analysis; retrospectively replacing the service model would turn the test into a model-selection set. Promotion therefore requires a fresh holdout, a new raw-score gate, and a regenerated reporting benchmark.

The calibration experiment addresses a different question: whether adjusted scores provide a more reliable basis for reporting. Here, temperature scaling worsens both the Brier score, from 0.163 to 0.174, and the ten-bin expected calibration error, from 7.96% to 12.78% (Figure 4). The fitted adjustment therefore does not generalize from the single-pothole calibration images to this test population. Since temperature scaling also preserves ranking, Algorithm 1 uses the equivalent raw-score threshold instead.

For the historical 12-epoch YOLO11n retraining checkpoint `control_seed391` (SHA-256 prefix `cce85af87abe`), threshold 0.3319 retains 162 candidates: 114 annotation matches and 48 non-matches. At confidence 0.25, this same checkpoint retains 116 matches and 62 non-matches. Its standard detector evaluation also records TP/FP/FN of 116/62/93. By contrast, the earlier target-trained YOLO11n checkpoint (`dcd6c38d6065`) supplies the 117/34/92 counts in Tables 11 and 17. The images and operating threshold are shared, but the weights and training histories differ; the FP difference is therefore not evidence of a disagreement between matching implementations. Standard AP and fixed-threshold confusion counts are the primary detector metrics. The reporting proxy separately assigns candidates in descending confidence order to unmatched annotations at IoU ≥ 0.5, in clipped original-image coordinates, so a changing gate preserves higher-score assignments. Its counts describe the stated checkpoint and gate only. Within `control_seed391`, the gate removes fourteen unmatched boxes at the cost of two matched boxes.

F.2.4 PAIRED DEPTH EVIDENCE

The original PothRGBD archive contains 1,000 RGB images, 1,000 depth arrays, and 1,000 segmentation-label files (Yurdakul & Taşdemir, 2025). Timestamp matching recovers a depth array for all 957 prepared RGB images. Decoded RGB content is identical to the corresponding archive image in 955 cases. Two require additional correspondence checks. Median nonzero depth coverage is 92.80%, 91.43%, and 92.60% in train, validation, and test.

Although raw depth can support review, the archive lacks the sensor scale, camera geometry, alignment evidence, and road reference plane needed to express a depression in centimeters, so the interface displays paired readings and valid-pixel coverage while leaving physical depth and severity unresolved (Figure 8).

F.3 EXECUTABLE EVIDENCE-AWARE INSPECTION

The updated prototype loads a hash-checked YOLO26n deployment policy, retains only its pothole predictions for the inspection interface, and attaches image and checkpoint identities to every record.

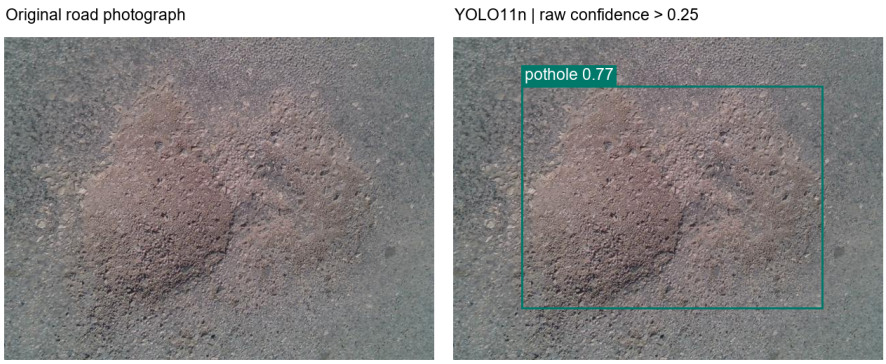


Figure 3: One inspection example pairs the original image from Yurdakul & Taşdemir (2025) with suspected damage from the archived YOLO11n checkpoint, which differs from the service model described in Appendix K.

Algorithm 1 Evidence-aware inspection reporting preserves supported observations, while it defers claims whose required location, depth, or maintenance evidence is absent.

Input: road image x. Selected detector f_θ. Model hash h_θ. Raw-score report threshold τ_s. Optional depth pair (d, h_d). EXIF GPS g_e or manual WGS84 tag g_m.
Output: retained records $\mathcal{O}$. Review queue $\mathcal{Q}$. Annotated image. Location ℓ. Evidence flags. Accepted text and text-review queues.

1. Verify the image identifier, h_θ, and, when supplied, the depth-pair hash h_d.
2. Detect at 640 pixels and remove duplicate boxes:
$$\mathcal{D} = \mathrm{NMS}_{0.70}(f_\theta(\mathrm{resize}(x, 640))), \quad \mathcal{D} = \{(b_i, c_i, s_i) \mid i = 1, \ldots, N\}.$$
3. Apply the frozen raw-score gate without rescaling confidence:
$$A_i = \mathbf{1}[s_i \geq \tau_s].$$
4. Create $r_i = (\mathrm{id}(x), h_\theta, b_i, c_i, s_i, \tau_s)$. Attach d only when the RGB-depth identity check passes. Otherwise mark depth as unresolved.
5. Set $\mathcal{O} = \{r_i : A_i = 1\}$ and $\mathcal{Q} = \{r_i : A_i = 0\}$. Set $\ell = g_m$ if a valid manual tag is supplied, otherwise $\ell = g_e$ for valid EXIF GPS, or $\ell = \varnothing$ when absent. Valid latitude/longitude satisfy $|\phi| \leq 90$, $|\lambda| \leq 180$ and are finite. Invalid manual input is rejected.
6. Mark retained and deferred candidates separately. Export $(\mathcal{O}, \mathcal{Q}, \ell)$ with location source and role. Physical depth and repair priority remain unresolved.
7. Optionally draft narratives from these records. Verify each proposition against its record and hold any sentence with an unsupported proposition for review (Section F.3.1).

A fixed raw threshold of 0.25 replaces the historical checkpoint-specific gate; no YOLO11n calibration is transferred to the new model. When an RGB-depth pair passes the identity checks, the record also includes raw readings and valid-pixel coverage. This structure makes missing evidence explicit: without geographic evidence, sensor calibration, or a maintenance policy, the corresponding location, metric depth, or repair priority remains unresolved. An empty detection result likewise does not establish road safety.

Location metadata follows the same provenance rule. The interface accepts optional EXIF GPS or a manual WGS84 geotag and records whether it refers to the camera or to reported damage. A manual entry takes precedence without erasing the EXIF value. Neither source is independently verified, and a camera position must not be interpreted as the position of every defect. To support review of the visual claims as well, the interface keeps the original photograph beside its predictions (Figure 3).

The archived 40-image HTTP benchmark in this paragraph uses the historical 12-epoch YOLO11n `control_seed391` checkpoint (SHA-256 prefix `cce85af87abe`), not the current YOLO26n

Table 13: The automatic grounding results use the same 40 images in each condition, while U_{text} measures the flagged clause rate and C_{text} measures accepted sentence coverage under LLM judgments.

Box rule	Sentences	Clauses	Flagged sentences	U_{text}	C_{text}
Raw 0.25	130	588	1	0.34	99.23
Report gate	130	582	1	0.34	99.23

service. With no supplied geotags, raw confidence 0.25 produces TP/FP/FN of 29/24/40. Its historical raw-score gate of 0.3319 preserves all 29 matches while reducing non-matches to 18. Warm CPU requests take 0.11 seconds at the median and 0.14 seconds at the 95th percentile; the first takes 2.47 seconds. These archived timings exclude the later LLM workflow. Appendix E instead reports a separate HTTP run on the same images using independently trained PothRGBD YOLO26n (`5c6405d0f7ef`) at 0.25: TP/FP/FN of 30/14/39 and warm median/p95 of 0.139/0.179 seconds. These are distinct model executions under local workloads, not repeated measurements of one checkpoint or a controlled speed comparison. Both runs leave unsupported geography, metric depth, and safety unresolved.

A reporting rule must balance how much it retains against how often retained statements lack support. Let A_i be 1 when candidate i is retained and U_i be 1 when its statement is unsupported. With N candidates, the two measures are

$$\widehat{R} = \frac{\sum_i A_i U_i}{\sum_i A_i}, \qquad \widehat{C} = \frac{1}{N} \sum_i A_i. \tag{4}$$

The first measure is the fraction of retained statements that lack support. The second is the fraction of all candidates retained. Risk is undefined when none are retained. Our technical measurements substitute annotation mismatch for U_i. A human study must instead adjudicate the complete assertion and its evidence. Matched coverage and review effort distinguish useful selectivity from indiscriminate deferral (Geifman & El-Yaniv, 2019).

F.3.1 Automatic grounding check for generated narratives

A municipal inspection record should not gain an invented street, depth, or repair decision when converted to prose. A prompted LLM writer describes every retained and deferred candidate and the unresolved metadata. A separately prompted LLM verifier splits each sentence into propositions and checks them against the supplied JSON. Unsupported sentences enter a review queue while their structured evidence remains available. This extends evidence attribution (Gao et al., 2023, Min et al., 2023) to the reporting layer, without verifying detector predictions or visually inspecting the image.

Let U_j^L indicate that the verifier flags clause j, and G_s indicate that every clause in sentence s passes. With M clauses and S sentences,

$$\widehat{U}_{\text{text}} = M^{-1} \sum_{j=1}^{M} U_j^L, \qquad \widehat{C}_{\text{text}} = S^{-1} \sum_{s=1}^{S} G_s. \tag{5}$$

We apply this check to narratives produced from the same 40 image records under each box rule. Each condition yields 130 sentences, of which one contains two flagged causal propositions; retaining the remaining sentences gives 99.23% sentence coverage (Table 13). In a separate control check, the verifier distinguishes ten injected errors from ten supported synthetic statements. These small controls demonstrate behavior on the constructed cases, not general hallucination-detection accuracy. Moreover, a zero post-filter flag rate is automatic when the same judge both filters and evaluates the text; it is not independent confirmation of correctness.

The study uses separate GPT-5.6 Sol writer/verifier contexts in the OpenAI Codex agent runtime. Archived prompts, inputs, and outputs support audit of the workflow. Five human raters pass both 12-item qualification dimensions and independently complete support and evidence judgments for 160 claims from all 40 images. Claim-support agreement is 96.63%, with nominal Krippendorff's $\alpha = 0.931$ and an image-cluster bootstrap interval of [0.891,0.963]. Evidence correctness has

96.75% agreement but only $\alpha = 0.300$ ([-0.010,0.489]). This discrepancy follows extreme prevalence: 781 of 800 evidence labels are *yes*, 19 are *unclear*, and none are *no*. High raw agreement therefore does not validate the visual-evidence criterion. We do not adjudicate or relabel ratings after observing agreement. Appendix O reports all ten pairwise confusion matrices per criterion and the preserved workbook hashes.

G Additional experimental details

Table 14: Historical YOLO11n fixed-threshold detection counts retained to reproduce the earlier analysis of missed objects, which are separate from the independently trained YOLO26n results in Table 2.

Dataset	Class	TP	FP	FN	F1 (%)
RDD2022	longitudinal crack	1737	1218	2068	51.39
RDD2022	transverse crack	855	625	937	52.26
RDD2022	alligator crack	889	449	645	61.91
RDD2022	pothole	342	257	629	43.57
PothRGBD	pothole	117	34	92	65.00
Chitholian	pothole	183	71	57	74.09
MWPD	pothole	140	76	61	67.15

Training and objective. The historical YOLO11n detector retains its standard localization, classification, and distribution focal losses (Li et al., 2020). Final-stage weights $(\lambda_{\text{box}}, \lambda_{\text{cls}}, \lambda_{\text{dfl}})$ are $(8.5, 0.5, 1.8)$ for PothRGBD, $(10, 0.5, 2)$ for Chitholian, and $(9, 0.5, 2)$ for MWPD. Stage epoch counts exclude earlier training and smoke stages. Training a target single-class head does not establish retention of four-class RDD2022 performance. Diagnostics use Ultralytics 8.4.129, PyTorch 2.6.0+cu124, 640-pixel inputs, batch size 8, square letterboxing, FP32 inference, and zero dataloader workers.

Metric calculation. AP averages ten IoU thresholds and all evaluated classes. Fixed-threshold precision is $\text{TP}/(\text{TP} + \text{FP})$, recall is $\text{TP}/(\text{TP} + \text{FN})$, and F1 is $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$. The implementation uses strict confidence and IoU comparisons for the confusion counts and its standard threshold vector for AP. Each dataset-specific model is evaluated on its own test images with the same input size, score floor, NMS setting, and maximum-detection budget. The single-class models do not require an RDD2022 pothole-class mapping.

H Evaluation history and attribution

Developmental holdouts. The prepared split records do not provide verified road-segment identifiers or geographic grouping suitable for a street-disjoint audit. Image-identity checks establish reuse boundaries, not acquisition-site independence. The prototype's optional EXIF or manually entered coordinates are separate application inputs and are not verified geographic labels for these evaluation splits. The test images had been inspected in earlier development or cross-domain diagnostics. The new in-domain runs select checkpoints on their own validation splits and record selection locks before their respective test evaluations. Earlier validation components were reassigned to training during protocol preparation. These records support a reproducible developmental comparison, not a newly acquired prospective test. Reporting selection and evaluation history separately from image-overlap checks follows the methodological concerns discussed by Kapoor & Narayanan (2023).

Dataset and model attribution. RDD2022 and MWPD are attributed to their published studies (Arya et al., 2024a, Parvin et al., 2025). PothRGBD is cited through its IEEE DataPort dataset DOI (Yurdakul & Taşdemir, 2025). The registered record lists CC BY 4.0, whereas its Kaggle mirror lists MIT. Those records are distinguished in the access audit. A mirror label is not treated as a blanket redistribution authorization. The local archive is pinned by SHA-256 and contains 3,000 members. Chitholian is cited through version 1 of its registered Kaggle dataset DOI (Rahman & Patel, 2020). The DOI names Atikur Rahman and Sachin Patel, while the hosting account is Atikur

Rahman Chitholian. Its Kaggle record labels the license as *Database: Open Database, Contents: Database Contents*. Auxiliary image families are kept distinct from the core evaluation and require their own attribution. YOLO11 follows the official software citation (Jocher & Qiu, 2024), whose documented license is AGPL-3.0. Software and dataset records are not represented as peer-reviewed papers.

Historical evaluation context. A provisional Chitholian evaluation used a 226-image pool formed by 101 core and 125 auxiliary images. The present dataset-specific result uses only the core 101-image test so that its evaluation family is explicit. Earlier source-only and transferred-checkpoint diagnostics remain archived with their original split memberships; they are not rows of the independent-training comparison.

I RETRAINING AND CALIBRATION DETAILS

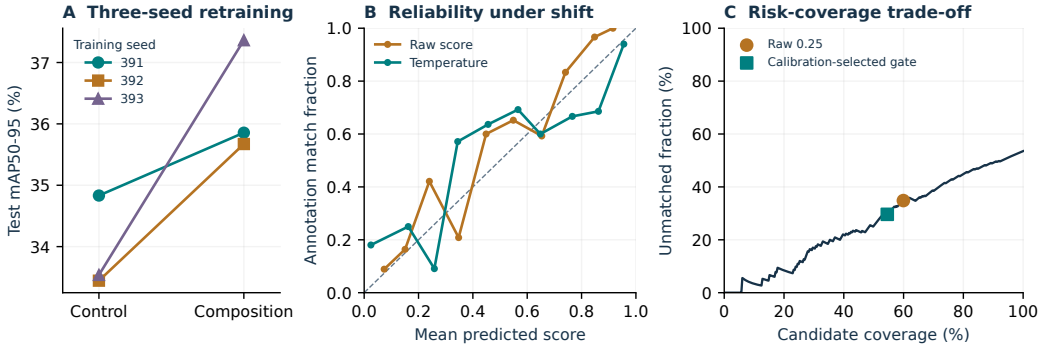


Figure 4: Retraining and reporting diagnostics in which panel A connects same-seed runs across the two arms, while panels B and C use the AP-selected checkpoint with temperature and raw-score gate fitted on the calibration partition and annotation matching as a technical proxy for claim adjudication.

Table 15: Individual runs at both training budgets with AP and F1 in percent, where selection AP determines the chosen checkpoint within each budget and multi-object FN counts misses among 101 annotations.

Epochs	Arm	Seed	Selection AP	Test AP	Test F1	Multi FN
12	Control*	391	64.75	34.83	59.95	81
12	Composition	391	59.22	35.86	61.96	62
12	Control	392	58.92	33.44	58.25	84
12	Composition	392	58.31	35.67	58.85	63
12	Control	393	63.91	33.54	57.50	82
12	Composition	393	59.44	37.36	61.85	61
24	Control*	391	65.07	35.07	61.46	78
24	Composition	391	64.94	40.18	63.70	69
24	Control	392	61.97	34.31	58.33	84
24	Composition	392	60.25	37.83	63.55	69
24	Control	393	61.84	34.69	58.49	84
24	Composition	393	60.46	37.46	62.87	69

* Chosen by selection AP. The application benchmark uses the 12-epoch checkpoint.

Table 16: Replayed calibration fits for the validation-selected YOLO11n checkpoints, where T and τ_q describe the diagnostic temperature transform and the executable report applies raw scores with threshold τ_s.

Epochs	T	τ_q	τ_s	Candidates	Matched/retained
12	0.47003748	0.18413263	0.33187971	77	40/44
24	0.41529334	0.14154522	0.32113245	65	41/45

Table 16 is generated by refitting temperature and both gates from the archived calibration predictions and checking them against the checkpoint lock. Both fits use 44 images, a 0.05 candidate floor, one-to-one IoU matching at 0.5, a 90% empirical precision target, and the historical minimum of ten retained candidates. The 12-epoch checkpoint supplies the frozen application benchmark. The diagnostic T and transformed threshold τ_q are reported for replication; Algorithm 1 uses the corresponding raw threshold τ_s. Both arms use AdamW with initial learning rate 3×10^{-4}, cosine decay to a factor 0.1, weight decay 5×10^{-4}, one warmup epoch, batch size 8, and 640-pixel inputs. Loss weights are $(8.5, 0.5, 1.8)$. Shared augmentation includes HSV amplitudes $(0.01, 0.3, 0.25)$, rotation amplitude 3°, horizontal flip probability 0.5, and no mixup. Mosaic is disabled for the last two epochs. The control runs at seed 391 (12 epochs) and seed 393 (24 epochs) resumed from saved checkpoints after interruptions. All twelve runs completed their assigned 12- or 24-epoch budgets. Within-budget comparisons share the schedule and settings. The budget contrast also changes the schedule duration. Selection, calibration, and test identities were checked against training using byte hashes, decoded-image hashes, source stems, and identity components.

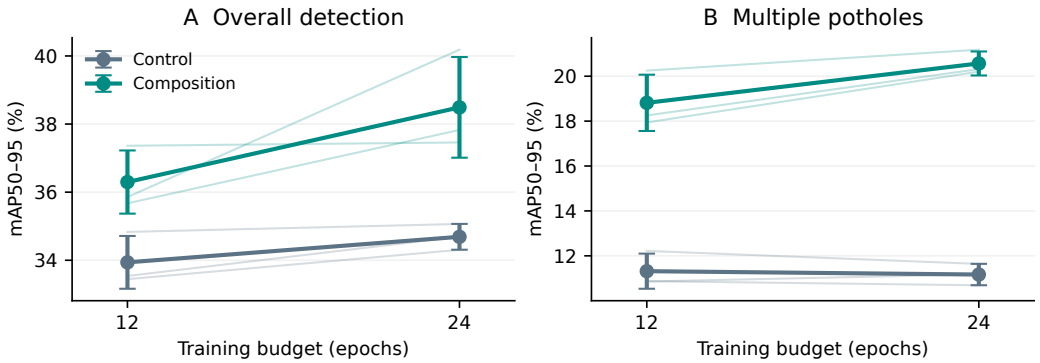


Figure 5: Training-budget comparison in which faint lines connect runs sharing a seed, thick lines show means, and bars give sample standard deviations across three seeds rather than confidence intervals.

The 24-epoch selection rule again chooses control seed 391 before its test evaluation. Its Brier score changes from 0.1722 to 0.1889 after temperature scaling, and ECE changes from 6.77% to 15.43%. Longer training therefore does not resolve the calibration-population mismatch. This checkpoint belongs to the archived YOLO11n calibration study; the current YOLO26n application is benchmarked separately.

Candidate reliability is conditional on predictions surviving the 0.05 score floor and does not measure undetected objects. Ten equal-width bins define expected calibration error. Confidence-ordered greedy matching operates in clipped original-image coordinates at IoU ≥ 0.5. The official AP and fixed-threshold confusion metrics retain their original evaluator. Diagnostic temperature fitting and operating-threshold fitting share the 44-image calibration subset. The precision constraint is empirical, with no held-out calibration guarantee. The same monotone score transformation preserves every accepted set under a corresponding raw-score threshold; because test Brier score and ECE worsen at both budgets, the algorithm records and applies that raw threshold rather than the transformed score.

The 40 technical benchmark images are the first 40 test entries sorted by image-content hash, without filtering on correctness. The service uses CPU inference with two computation threads. Warm latency summarizes requests 2 to 40. The first request includes model loading. Service measurements include current-machine workload and are not isolated hardware throughput measurements. This frozen benchmark uses the 12-epoch checkpoint and should be read as a reproducibility snapshot, not the best observed detector. The 24-epoch scale-only result was identified on the developmental test and requires prospective promotion on new data. The human forms use an earlier fixed PothRGBD checkpoint, whose identifier remains attached to each stimulus, and are separate from the technical benchmark.

J INSPECTION INTERFACE AND ASSESSMENT MATERIALS

The assessment materials distinguish aggregate results from individual inspection records. The dashboard summarizes dataset scope, measured errors, and missing metadata (Figure 6); the executable interface provides the image-level records used in the HTTP benchmark. Deterministic aggregate examples can be reproduced from their recorded inputs. The human follow-up adds 160 claims from all 40 images, presented in five separately ordered workbooks. Some claims deliberately exceed their structured evidence to test assessor judgment. The submitted timing column is excluded from conclusions about review speed.

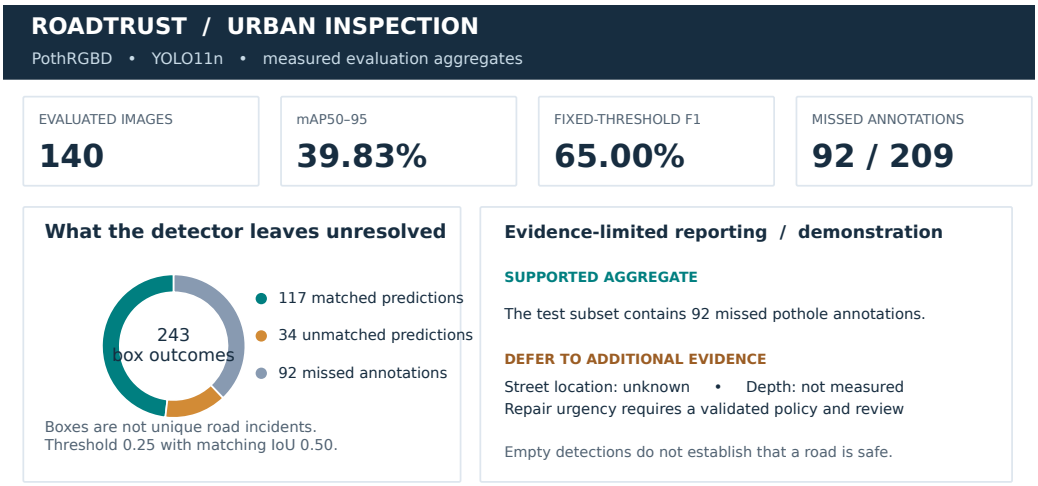


Figure 6: The inspection dashboard presents measured PothRGBD results, while box outcomes remain distinct from unique incidents or repair tickets and missing evidence leaves geographic and physical claims unresolved.

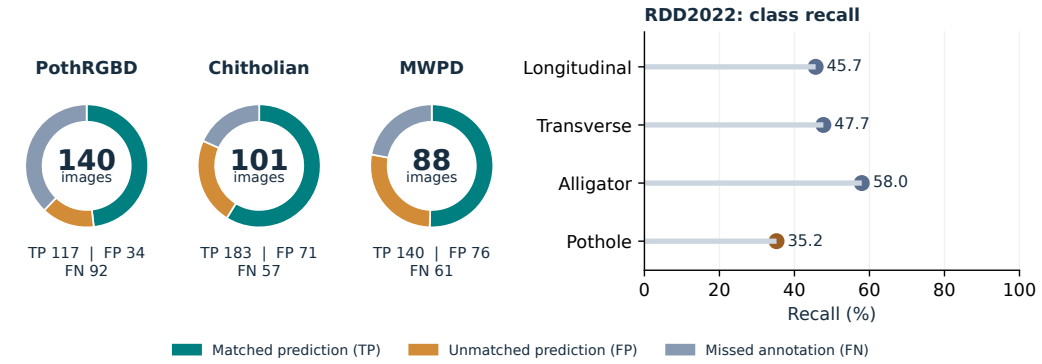


Figure 7: Detection error patterns compare fixed-threshold outcomes with RDD2022 class recall, while target rings normalize TP+FP+FN and class recall uses annotated instances as its denominator.

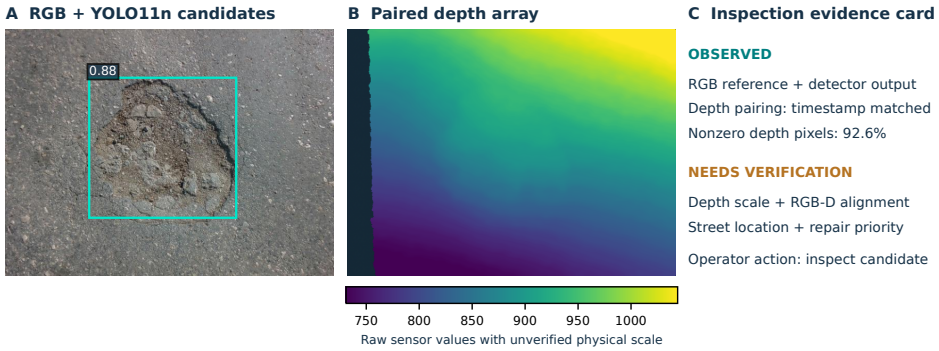


Figure 8: Paired RGB and raw depth evidence from Yurdakul & Taşdemir (2025) for an example with median valid-depth coverage, with zero-valued depth pixels masked and YOLO11n boxes marking candidates whose displayed depth values cannot be read as physical pothole depth without RGB-D calibration.

Table 17: The PothRGBD test groups use confidence 0.25, while the breakdown preserves all 140 images and retains the background cases.

Annotations per image	Images	Annotations	TP	FP	FN
Zero	3	0	0	3	0
One	108	108	98	14	10
Multiple	29	101	19	17	82

K REAL-IMAGE INSPECTION EXAMPLES

The first three pilot cases in the fixed image-hash order illustrate what the reviewer sees before and after detection. Their inclusion does not depend on whether the detector is correct. The complete local gallery retains all ten cases. Original photographs appear on the left and archived YOLO11n predictions on the right. A predicted box is a candidate observation, not an independently verified annotation.

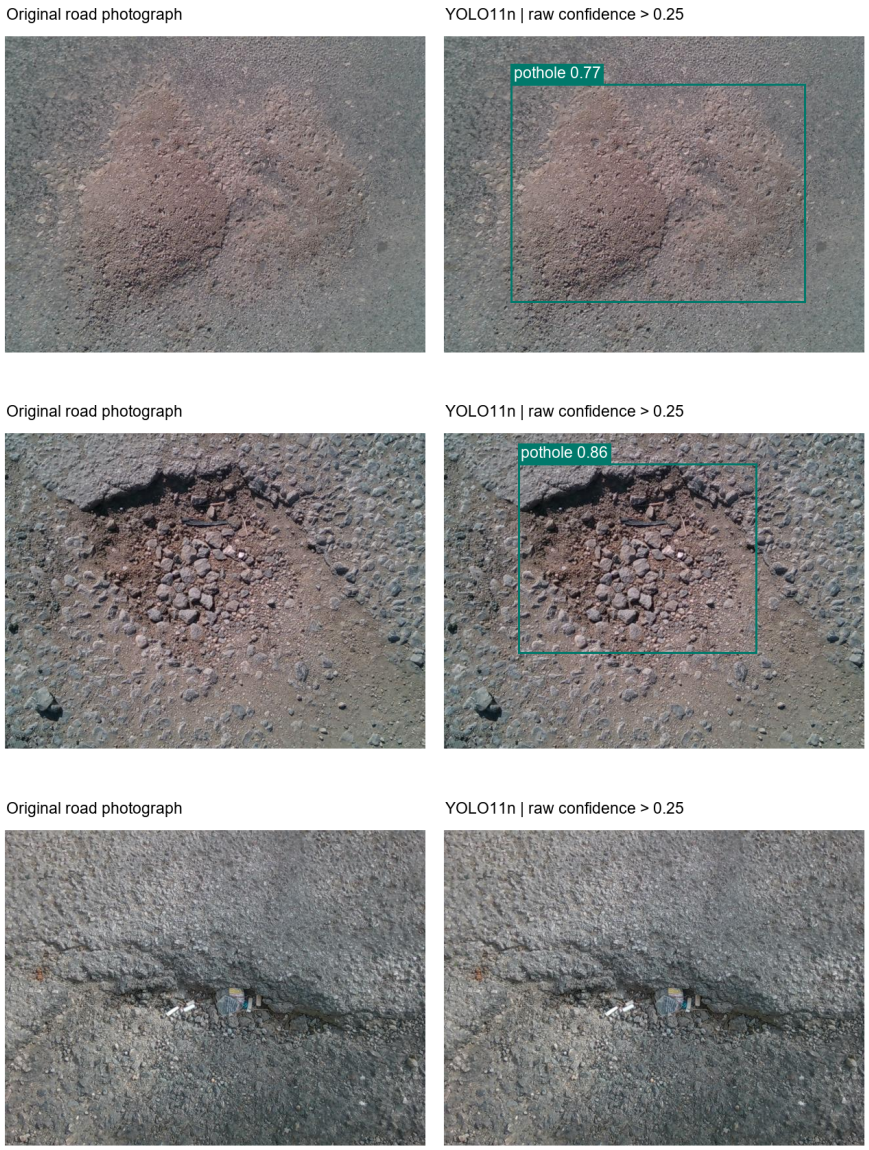


Figure 9: Original images from Yurdakul & Taşdemir (2025) beside archived YOLO11n predictions above confidence 0.25, redrawn with larger labels from the same boxes and rounded scores, while cases 1 to 3 keep the fixed order and illustrate assessor evidence rather than model accuracy.

L INDEPENDENT RATING PROTOCOL

Five assessors independently judge all 160 statements from the same 40 images. Each first completes a 12-item qualification for support and evidence correctness and must score at least 10/12 on each dimension; all five score 12/12. Each receives a separately ordered form and identical evidence. Support is completed before visual evidence correctness. Original judgments are retained without post-hoc relabeling, and any later adjudication must be stored separately.

Support is coded yes when the full statement is supported, no when an important assertion contradicts or exceeds the evidence, and unclear when the material cannot be assessed. Evidence correctness separately asks whether the cited material refers to the right image or object and is described accurately. A correct image can accompany an unsupported assertion. Missing responses remain missing. Active review time begins after the evidence loads and ends when both judgments and the reason are recorded. Interruptions are excluded and documented.

Nominal Krippendorff's alpha is the primary agreement measure for each judgment dimension. Pairwise unweighted Cohen's kappa, exact agreement, and the three-category confusion matrices are also reported. Unclear is a separate category, not a midpoint on an ordinal scale. An ordinal alternative would require a prospectively specified scale of no, partial, and full support, with inability to assess recorded separately. The nominal agreement measures follow Krippendorff (2018) and Cohen (1960).

Uncertainty is estimated with 5,000 bootstrap samples of the 40 images, retaining all four statements and all five raters within each sampled image. Repeated images contribute repeated units. Percentile intervals use the 2.5th and 97.5th percentiles of defined estimates. Undefined repetitions and missing judgments are reported explicitly. Only claims with at least two observed ratings contribute to alpha. A zero expected disagreement yields an undefined coefficient. The intervals are conditional on these participating assessors.

The assessment includes deliberately unsupported statements and has no randomized interface control. Its label frequencies therefore do not estimate deployment error rates, and its timing cannot establish time savings. Agreement measures consistency rather than correctness. The completed ratings are reported with their original labels and without post-hoc adjudication.

Connection to a smart-city workflow. An image-level record could enter a municipal inspection queue after an operator supplies verified geographic coordinates, capture time, and the relevant road identifier. A GIS or citizen-reporting service could then associate repeated observations with a road asset. This is a proposed integration route: the current image collections do not establish those geographic links, and the prototype has not been evaluated as a citywide dispatch system. Repair decisions would still require inspection and a municipal prioritization policy.

M SINGLE-FACTOR AUGMENTATION ABLATION

We increase mosaic probability from 0.2 to 0.8, scale amplitude from 0.2 to 0.5, or translation amplitude from 0.05 to 0.10, changing only one setting per configuration. Each configuration uses seeds 391, 392, and 393, the common source weights, and the same 24-epoch recipe as the existing control and composition runs. All nine new runs complete the fixed budget. The best checkpoint is selected by selection-set AP50-95, while F1 is measured at confidence 0.25 and matching IoU 0.50. Mosaic remains disabled for the last two epochs. Translation-only seed 393 resumed from its saved optimizer checkpoint after epoch 10 following an interruption. It completed the same 24-epoch budget. The archived record identifies this continuation.

The contrasts answer a local question: what changes when one parameter is increased from the control settings? They do not estimate an average effect over all combinations. The image collection has already informed development. This follow-up is not an evaluation on a newly acquired holdout.

Table 18: The single-factor ablation reports mean ± sample standard deviation across three seeds, while paired ΔF1 uses unrounded seed-level scores and the control and composition rows reuse the matched-budget runs.

Augmentation	mAP50-95	Multi mAP50-95	F1 at 0.25	ΔF1
Control	34.69 ± 0.38	11.17 ± 0.48	59.43 ± 1.76	n/a
Mosaic only	34.48 ± 1.13	12.55 ± 1.55	59.10 ± 1.47	-0.32 ± 1.57
Scale only	39.75 ± 0.28	19.26 ± 0.95	63.84 ± 0.10	$+4.42 \pm 1.74$
Translation only	34.43 ± 0.72	11.91 ± 1.33	60.78 ± 0.79	$+1.35 \pm 1.16$
Composition	38.49 ± 1.48	20.57 ± 0.53	63.37 ± 0.44	$+3.95 \pm 1.53$

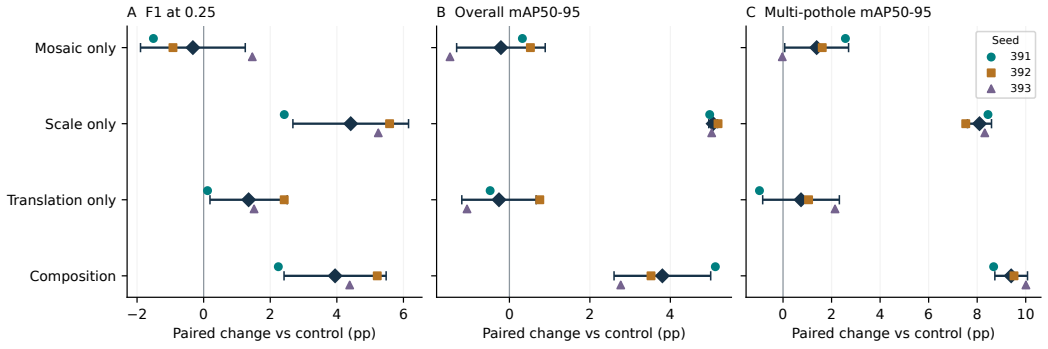


Figure 10: Single-factor augmentation comparison in which symbols show individual seeds, diamonds and bars show the paired mean and sample standard deviation, and joint composition augmentation is included as a reference.

With $Y_{abc,s}$ denoting a score for seed s and low/high settings $a, b, c \in \{0, 1\}$, the individual contrasts and the joint residual are

$$d_{M,s} = Y_{100,s} - Y_{000,s}, \quad d_{S,s} = Y_{010,s} - Y_{000,s}, \quad d_{T,s} = Y_{001,s} - Y_{000,s},$$

$$r_s = Y_{111,s} - Y_{000,s} - d_{M,s} - d_{S,s} - d_{T,s}.$$

The F1 residual is -1.50 ± 2.54 percentage points (mean and sample standard deviation across seeds). Only five of the eight configurations in a full 2^3 design are observed: control, each single-factor change, and the joint change. The mosaic-scale, mosaic-translation, and scale-translation configurations are absent. The residual therefore combines two-way and three-way interactions; cancellation can produce a small residual even when individual interactions exist. It is a descriptive contrast, not an identified interaction mechanism or evidence of a general causal effect. Three seeds give only a limited estimate of training variability.

N EXTERNAL DETECTOR AND TELEMETRY STRESS TESTS

The architecture extension uses seed 391 and a fixed 20-epoch recipe within each of two datasets. DRDD retains its official 1,003/288/197 train/validation/test partition. UAV-PDD2023 is split by source-frame group into 1,700/366/374 images, with zero group overlap, so derived perspective views of a source frame remain in one partition. All models begin from their documented public pretrained weights, validation selects the checkpoint, and the test partition is evaluated afterward. The three common-RDD rows below are separate five-epoch screening runs; they are included for provenance and must not be compared numerically with the 20-epoch DRDD/UAV rows or used to rank datasets.

Table 19: Additional detector measurements in percent, which are meaningful within each dataset block and recipe but not across datasets.

Model / evaluation split	Images	AP50	AP50-95	F1@.25
YOLOv8n / common RDD val	3721	32.24	14.19	38.23
YOLOv12n / common RDD val	3721	29.74	12.72	35.21
YOLOv10n / common RDD val	3721	24.64	10.74	24.41
YOLO11n / DRDD test	197	47.74	24.71	56.51
YOLOv10n / DRDD test	197	55.61	30.81	49.21
RT-DETR-L / DRDD test	197	1.59	0.44	7.85
YOLO11n / UAV test	374	14.43	5.40	22.49
YOLOv10n / UAV test	374	12.56	4.82	12.33
RT-DETR-L / UAV test	374	2.27	0.79	10.61

On DRDD, YOLOv10n has the highest AP50-95 (30.81%), while YOLO11n has the highest fixed-threshold F1 (56.51%). On UAV-PDD2023, YOLO11n leads both AP50-95 (5.40%) and F1 (22.49%). RT-DETR-L performs poorly under this fixed recipe. These are separate in-domain architecture comparisons within each collection. They do not show that one family is universally better, because the experiment uses one seed and does not tune a family-specific schedule.

RoadSens-4M supplies a non-visual telemetry endpoint. We construct 11,911 accepted windows from 101 sessions, keep sessions disjoint across train, validation, and test, and predict Normal, Pothole, or Bump from inertial features. Coordinates, free-text annotations, press duration, and elapsed time are excluded from predictive columns. Logistic regression is selected on validation data and reaches 39.34% test macro-F1, with a session-bootstrap 95% interval of [34.43,42.92]%.

Table 20: Session-disjoint RoadSens-4M telemetry result, with macro-F1 reported in percent for the validation and test sessions.

Telemetry endpoint	Validation macro-F1	Test macro-F1
logistic_regression	38.11	39.34

This baseline establishes an auditable telemetry evaluation rather than a deployable fusion model: it is not joined to the image records, does not estimate geographic coverage, and does not validate maintenance decisions.

O NARRATIVE GROUNDING AND HUMAN ASSESSMENT

Frozen inputs and roles. The 40 original HTTP records contain 90 candidate boxes. Raw confidence 0.25 retains 53 and defers 37. The order-equivalent raw-score report gate retains 47 and defers 43. Both conditions narrate all 90 candidates and 40 image-level metadata records, so the total is 260 short narratives. Two GPT-5.6 Sol writer contexts each handle 20 images. Two fresh GPT-5.6 Sol verifier contexts assess corresponding evidence-text pairs without writer reasoning or prompt. The contexts are separate, but independence of underlying model weights is not established. This is a preliminary descriptive study with archived outputs, not a reproducible model-ranking bench-

mark. Four ten-image batches preserve paired conditions. No generations are revised after receiving verifier judgments.

Prompts and output contract. The writer receives five synthetic correct/incorrect examples covering retained candidates, deferred candidates, missing location/depth/priority, EXIF camera position, and manual zero coordinates. Each output links an unchanged input identifier to one or two sentences. The verifier independently extracts propositions, supplies a binary supported/unsupported label and a JSON Pointer for the supporting field, then aggregates clause labels to sentence labels. Identifiers, complete sentence coverage, pointer validity, and label consistency are checked before export. A malformed response fails closed. Evidence records and rejected sentences are retained. A supported description of a deferred box remains in the review queue. Full prompts and raw outputs accompany the submission materials.

Denominators and interpretation. The raw-threshold condition has 588 clauses and the report-gate condition has 582, reflecting agent segmentation and wording rather than different scene samples. Each has two flagged clauses in one sentence: automatic unsupported rates are 0.34% after rounding. Sentence flag rates are 1/130 or 0.77%. These are paired descriptive results, not independent replicates or evidence that the box threshold improves text grounding. A supported statement that a model predicts a suspected pothole can coexist with a false-positive box. The original annotation-proxy box risks therefore remain separate from text-grounding rates.

Writing $s(j)$ for the sentence containing clause j, a post-gate risk using an independent adjudicated label U_j^H would be

$$\widehat{R}^H_{\text{text}} = \frac{\sum_j G_{s(j)} U_j^H}{\sum_j G_{s(j)}}.$$

It is undefined if no clauses remain and currently unmeasured because independent labels are absent. Replacing U_j^H by the same verifier's U_j^L forces the numerator to zero. We therefore do not present that mechanical result as a measured reduction in hallucination.

Error example and control cases. The flagged metadata wording links unverified depth scale/alignment to unresolved location and repair priority. The verifier treats these causal links as unsupported: the record lists separate missing evidence but supplies no causal relationship between them. This interpretation itself requires human validation. Twenty author-constructed controls, held separate from natural reports, contain ten supported sentences and ten errors (invented location, metric depth, urgency, safety, changed decisions, overclaimed verification, or evidence missingness). Expected labels are not supplied to the verifier. It flags 10/10 errors and 0/10 supported controls. These intentionally small, clear cases cannot estimate general detection sensitivity.

Preliminary rubric ratings. The separate pilot has 40 claims from ten images. It is not the 40-image reporting benchmark. The text-only agent assigns support labels yes/no/unclear to 15/15/10 claims and evidence-correctness labels to 0/15/25. All 40 are routed to human review. These are agent judgments, not human results. The agent lacks original images, overlays, and an independent identity check. Human assessors are supplied visual material. Hash strings alone do not establish that a claim and image are correctly linked. Positive informal author-reported feedback is not converted into structured ratings.

Completed five-rater human assessment. Five human participants completed EXP-coded workbooks, each containing support and evidence-correctness labels for the same 160 claim identifiers across 40 images. The ratings are separate from the preliminary text-only agent pass. Original workbook hashes are preserved; A to E filename aliases are exact copies and are not counted as additional raters. The author identifies the participants as human, but the records do not establish a representative inspector sample, demographic coverage, compensation, or independence beyond the collection instructions. Timing fields are excluded because no external trace establishes active review duration.

Nominal Krippendorff's α treats yes/no/unclear as distinct categories. Image-cluster bootstrap intervals use 5,000 resamples with seed 391, retaining four claims and all five raters for every sampled image. The support result is strong: $\alpha = 0.931$ with interval [0.891,0.963]. Evidence correctness

Table 21: Five-rater follow-up on 160 claims from 40 images, with intervals from 5,000 image-cluster bootstrap samples that condition on these raters.

Criterion	Claims	Alpha	95% CI	Pairwise agreement
Claim support	160	0.931	[0.891, 0.963]	96.63%
Evidence correctness	160	0.300	[-0.010, 0.489]	96.75%

is different. Although 1,548 of 1,600 pairwise comparisons agree (96.75%), 781 of 800 labels are *yes*, 19 are *unclear*, and none are *no*. The resulting $\alpha = 0.300$ with interval [-0.010,0.489] indicates poor chance-corrected reliability and an unstable lower bound. Raw agreement alone would therefore materially overstate validation.

Table 22: Pairwise three-category confusion matrices for the five-rater follow-up, where each entry lists yy, yn, yu then ny, nn, nu then uy, un, uu with rows for the first rater and columns for the second in yes, no, unclear order.

Criterion	Pair	Confusion matrix
Support	EXP_01-EXP_02	99,0,1;0,57,3;0,0,0
Support	EXP_01-EXP_03	99,0,1;0,59,1;0,0,0
Support	EXP_01-EXP_04	98,0,2;0,58,2;0,0,0
Support	EXP_01-EXP_05	97,0,3;0,59,1;0,0,0
Support	EXP_02-EXP_03	98,0,1;0,56,1;1,3,0
Support	EXP_02-EXP_04	98,0,1;0,55,2;0,3,1
Support	EXP_02-EXP_05	96,0,3;0,56,1;1,3,0
Support	EXP_03-EXP_04	97,0,2;0,57,2;1,1,0
Support	EXP_03-EXP_05	96,0,3;0,58,1;1,1,0
Support	EXP_04-EXP_05	95,0,3;0,57,1;2,2,0
Evidence	EXP_01-EXP_02	154,0,6;0,0,0;0,0,0
Evidence	EXP_01-EXP_03	155,0,5;0,0,0;0,0,0
Evidence	EXP_01-EXP_04	156,0,4;0,0,0;0,0,0
Evidence	EXP_01-EXP_05	156,0,4;0,0,0;0,0,0
Evidence	EXP_02-EXP_03	152,0,2;0,0,0;3,0,3
Evidence	EXP_02-EXP_04	152,0,2;0,0,0;4,0,2
Evidence	EXP_02-EXP_05	152,0,2;0,0,0;4,0,2
Evidence	EXP_03-EXP_04	153,0,2;0,0,0;3,0,2
Evidence	EXP_03-EXP_05	153,0,2;0,0,0;3,0,2
Evidence	EXP_04-EXP_05	153,0,3;0,0,0;3,0,1

The full confusion matrices expose every pair rather than collapsing disagreement into one percentage. The lack of negative evidence judgments also limits sensitivity assessment: the study cannot show that raters reliably reject visually incorrect evidence. A future assessment should prospectively enrich incorrect and ambiguous image-claim pairs while preserving their natural prevalence in a separately reported stratum. Representative-inspector agreement and field correctness remain unmeasured.