

Failure prediction and presence cues for budgeted review of Saudi road images

Samer Abaddi

Research Department, IDentity Research, Amman, Jordan

ORCID: [0000-0003-4612-7384](https://orcid.org/0000-0003-4612-7384)

Correspondence: eng.sameralabbadi@gmail.com

Abstract

Limited review capacity makes report selection consequential for automated pavement inspection. This study tested whether regional failure prediction improves the recovery of complete pothole-detector misses beyond image-presence controls. A compact Faster R-CNN detector was evaluated on 4,179 Saudi reports, including 2,802 source-positive reports. Three optimization seeds crossed with three protected folds produced nine correlated configurations. A regional predictor combined spatial features, detector summaries and no-box-first ordering. Residual risk was the proportion of source-positive reports with no matched target after review, assuming perfect correction against supplied annotations. At 10% review (417 reports), candidate risk averaged 24.63%, compared with 25.04% and 25.01% for center- and letterbox-presence controls. Candidate-minus-control differences were -0.412 and -0.385 percentage points, with adjusted 97.5% paired protected-group bootstrap intervals of $[-0.849, +0.107]$ and $[-0.864, +0.087]$, respectively. Superiority over both controls was not established. Fine-tuned presence had the lowest descriptive deployable-policy risk, 24.33%. At this budget, all candidate reviews contained no emitted box, leaving a mean 346.44 emitted-but-incorrect complete misses unreviewed. Unchanged transfer across seven external acquisition settings was heterogeneous: Japan favored the candidate, while external detector F1 ranged from 0.0000 to 0.0919. The contribution is a reproducible, capacity-matched benchmark that separates detection quality from review allocation and identifies how fixed queue priority can exclude a failure type from correction. The findings motivate joint validation of queue allocation and emission thresholds with independent pavement assessment; source-reference oracle recovery does not establish human inspection or safety benefits.

Keywords: Pothole detection; Pavement inspection; Failure prediction; Review allocation; Computer vision; Domain shift; Reproducibility.

Introduction

Reliable road inspection requires more than recognizing visible damage; it requires deciding which automated assessments should not be left unexamined. A detector that returns no pothole boxes leaves an unresolved distinction between an undamaged scene and a missed defect. When inspection images arrive as reports containing several photographs, that ambiguity becomes a resource-allocation problem: which reports should receive additional review when only a fraction can be inspected? Detection and review therefore answer different questions. The former concerns locating annotated objects; the latter concerns recovering consequential omissions within a limited review budget. Improving one does not, by itself, establish improvement in the other.

This distinction is pertinent to the expanding digital inspection infrastructure of the Middle East and North Africa (MENA). Saudi Arabia had more than 15.8 million registered roadworthy vehicles at the end of 2024, 6.9% above 2023. Its 2025 statistical release recorded 17,231 serious crashes in 2024, with 60.3% occurring within cities ([General Authority for Statistics, 2025](#)). In February 2026, the Roads General Authority introduced real-time coordination of survey activities and digital imaging within its national inspection campaign ([Saudi Press Agency, 2026](#)). As image acquisition is expanded, the allocation of scrutiny warrants evaluation alongside automated recognition.

Saudi municipal inspection data include photographs of potholes, excavation barriers, and damaged sidewalks, with annotations developed through an active-learning process ([AlElaiwi et al., 2023](#)). More recent work combines ensemble detection and image captioning to support Saudi visual-pollution reporting ([Almalki & Algethami, 2025](#)). Such advances connect perception with administrative workflows, but reporting capability does not determine whether an additional failure monitor provides useful review ordering. Nor should source annotations be equated automatically with independently verified physical condition. This distinction is

particularly consequential when the same references define both detection errors and the assumed recovery obtained through review.

Geographical transfer further complicates the problem. RDD2022 contains 47,420 images from six countries and more than 55,000 annotations spanning four damage categories (Arya et al., 2022). DA-RDD addresses cross-country pavement detection through image-level and instance-level domain alignment (Lin et al., 2023), whereas Multi-Adversarial Metric-Constrained Domain Generalization learns representations from multiple source domains without target-domain training (Chen et al., 2026). These approaches establish relevant adaptation and generalization precedents. They do not, however, resolve whether a monitor trained with one Saudi inspection source retains useful review priorities when inference is transferred unchanged across acquisition settings.

Failure prediction also has direct methodological antecedents; for example, knowledge-Guided Failure Prediction compares detector representations with visual foundation-model embeddings to identify images on which safety-critical objects may be missed (Zimmermann et al., 2026). Its contribution makes clear that adding a learned monitor is not, by itself, a new research direction. For pothole inspection, the sharper question concerns incremental information: does a regional failure representation improve review allocation beyond a classifier estimating target presence and the observable fact that a detector emitted no boxes? An apparently sophisticated monitor should be evaluated against those inexpensive, task-relevant explanations before its operational value is inferred.

Traub et al. (2024) show that selective-classification metrics can change confidence-method rankings, emphasizing alignment between the metric and the intended decision. Montreuil et al. (2025) extend learning-to-defer to joint classification and regression, including object detection and agent-specific costs. Here, a narrower estimand is adopted: the proportion of all source-positive reports that remain completely missed after budgeted review under an assumed perfect source-reference oracle. The denominator includes reviewed positives; this is not conventional error conditional on accepting automation. Without measured reviewer

accuracy or effort, oracle recovery cannot establish human-review effectiveness, maintenance savings, or crash prevention.

Related MENA transportation research provides a useful systems perspective. [Abaddi \(2026a\)](#) proposes quantum-computing and GPT-4o components for camel–vehicle collision detection, connecting perception with hazard-response decisions. Furthermore, [Abaddi \(2026b\)](#) examines Amman's Abdali interchange using field-supported traffic simulation and operational analytics to compare interventions. The studies illustrate the importance of connecting computational outputs to a defined transport decision.

To summarize the study problem: a complete miss occurs when no annotated pothole is matched in any photograph belonging to a source-positive report. Two mechanisms are distinguished: silent misses with no emitted boxes, and complete misses despite emitted, incorrect detections. Partial object misses are separate. Figure 1 summarizes how observable prediction cues are converted into review priorities, while source references remain outside the locked inference and ranking path. Review capacity is expressed as the number of reports examined. A report containing several photographs is retained as one decision unit. Consequently, the benchmark isolates queue ordering while identifying the additional evidence required to translate computational recovery into field practice.

The objective is to assess whether regional failure prediction adds value beyond two prespecified presence controls, using center crops and letterboxing, at a 10% review budget, while examining failure mechanisms and transfer limitations. Nine matched configurations, comprising three optimization seeds crossed with three held folds, support the comparison. Saudi reports form the primary evaluation cohort; seven external acquisition settings are assessed separately without target adaptation. Regional, presence, and emission-status ablations are examined under fixed comparison rules, with paired group uncertainty and operating-point sensitivities. The study's contributions are:

- A report-level benchmark that distinguishes complete from partial misses and separates silent from emitted-but-incorrect failures under an explicit source-reference recovery assumption.

- A controlled assessment of regional information against strong presence and emission-status baselines, supported by matched configurations, paired uncertainty, and prespecified sensitivities.
- An auditable account of source-specific transfer and computational provenance, preserving frozen predictions, unsuccessful attempts, timing records, and genuine inference demonstrations without claiming deployment safety.

Section 2 presents the methods, and Section 3 reports the results. Section 4 addresses interpretation and limitations. Section 5 concludes the study.

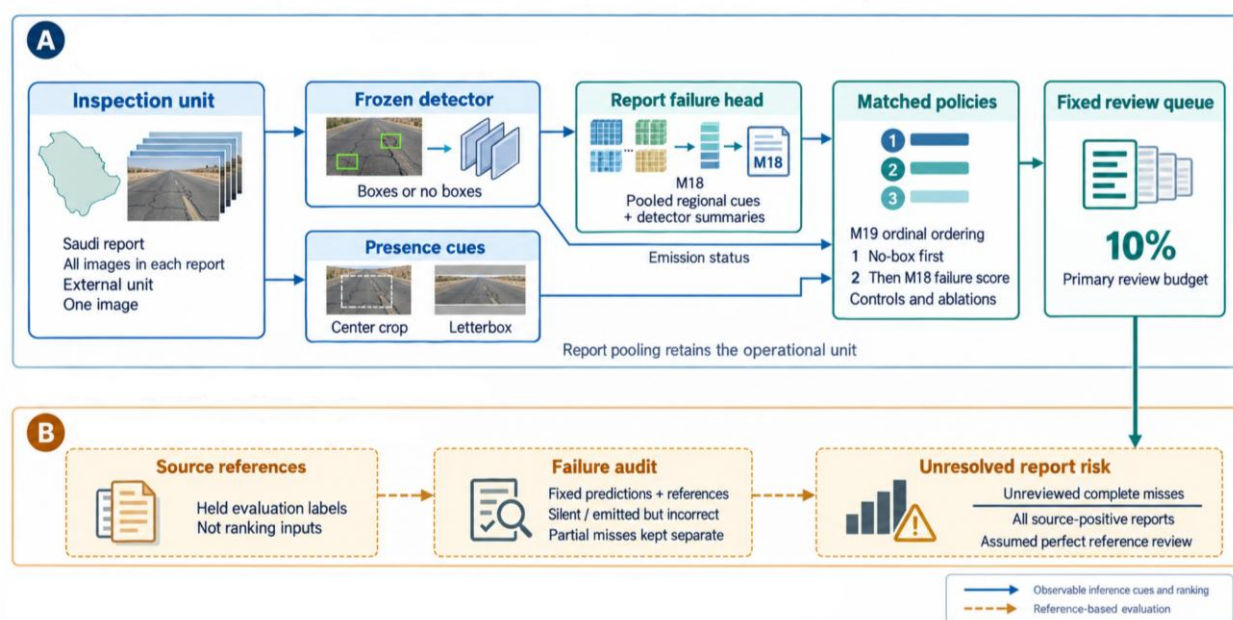


Figure 1. Semantic framework for budgeted review of complete pothole detection failures: (A) From observed images to a fixed review queue, and (B) Reference-based evaluation after ranks are fixed

Note: 3 seeds $\times$ 3 held holds. 7 external acquisition settings. M19 is ordinal

Methods

Study design and reference data

A retrospective, secondary-image benchmark was conducted to determine whether a report-level failure monitor improves the allocation of limited review capacity beyond target-presence cues. The primary setting comprised Saudi road-inspection reports; transfer was

assessed separately across external acquisition settings. Review was represented by a perfect source-reference oracle: selecting a report was assumed to resolve its annotated complete miss. Accordingly, the endpoint measures allocation relative to supplied annotations, rather than observed expert performance, physical pothole severity, or maintenance benefit.

The original partitions were frozen on 19 October 2026. Following adaptive development, the main protocol and comparison family were prospectively frozen on 21 October 2026, before release of the main test predictions. Earlier engineering checks remained development evidence. The study was not externally preregistered, and a local SHA256 record establishes file integrity. Explicit separation of fitting, selection, and evaluation information is consistent with concerns about leakage in machine-learning-based science ([Kapoor & Narayanan, 2023](#)).

The executed Saudi source was the public-road visual-pollution data set, version 4, released on 22 May 2023 ([AlElaiwi, Ahmad, et al., 2023](#)). Its 31,795 labelled images retained their original report identifiers. External data were taken from the [versioned RDD2022 archive](#), released on 29 October 2022; the final data article documents its acquisition diversity ([Arya et al., 2024](#)). The latter supplied seven acquisition settings from six countries, including separate Chinese drone and motorbike collections. Figure 2 locates the source countries.

Across both archives, 79,215 image records were inventoried. The 9,035 official RDD test images without released labels were excluded from scoring and were never assigned negative targets. Deterministic conversion retained all 70,180 labelled images and 32,519 target boxes: Saudi YOLO class 2 and RDD category D40. Labelled images without those targets remained source-negative examples. Cross-source equivalence was operational: annotation completeness and physical pothole boundaries were not independently adjudicated. Saudi normalized centres were converted to continuous pixel-edge coordinates; original numeric VOC edges were retained without an undocumented one-pixel shift. Small boundary overshoots were clamped for 643 Saudi target boxes within the recorded normalized tolerance of 0.000001. One invalid non-target RDD box was excluded; no target box was discarded.

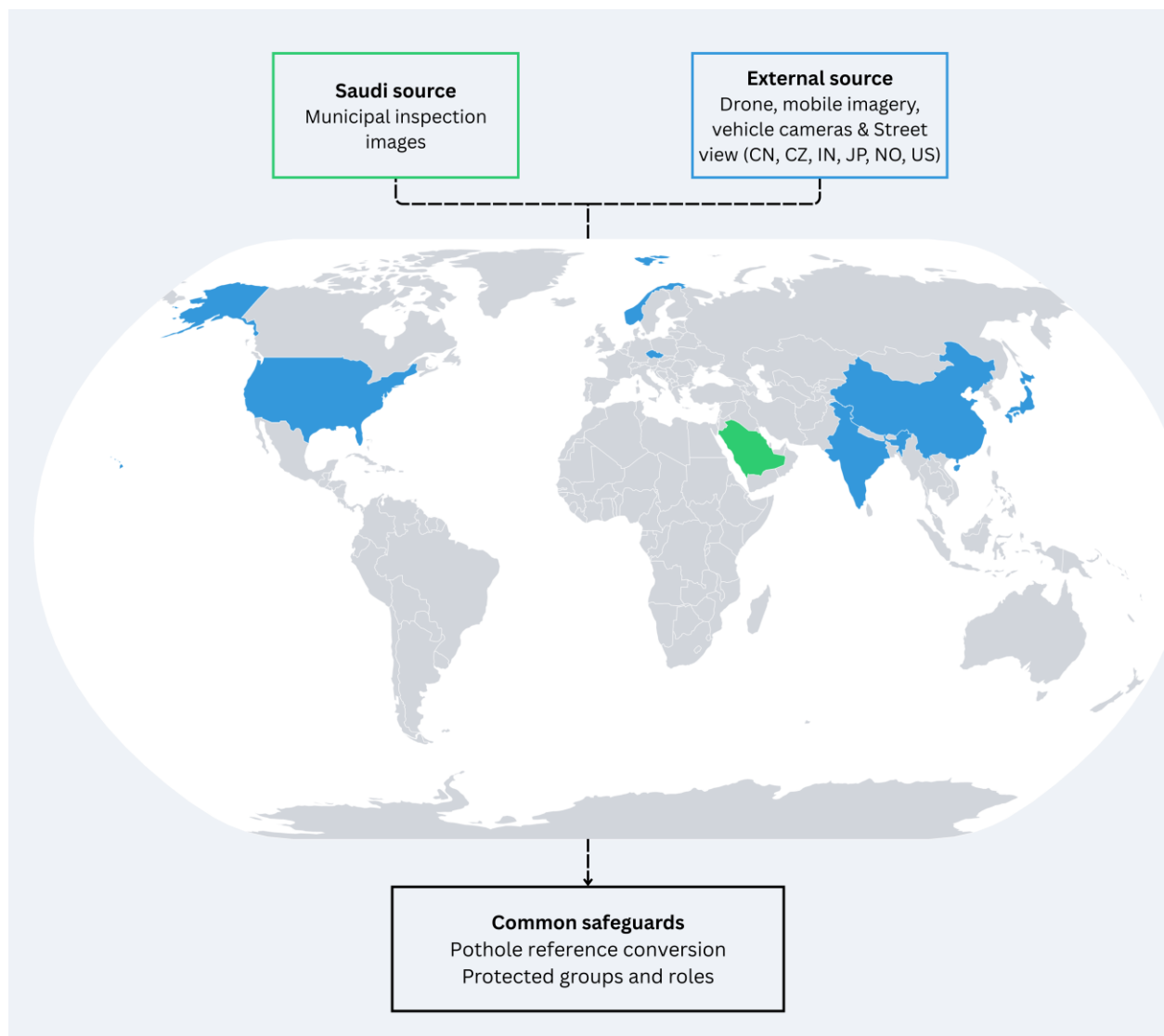


Figure 2. Dataset provenance and the distinction between source records and analysis units. Saudi municipal reports form the primary setting. External inference covers China, Czechia, India, Japan, Norway, and the United States; Chinese acquisition modes remain separate. Camera and platform descriptions follow the source data documentation (AlElaiwi, Ahmad, et al., 2023; Arya et al., 2024).

Conversion verified source-label hashes and read held-out label metadata for schema validation, without using model outcomes for selection. Raster orientation was preserved so that pixel coordinates remained consistent with the supplied boxes. No exclusions were introduced because a trained model performed poorly. Table 1 distinguishes archive rows, byte-unique images, analysis units, and protected groups; these quantities are not interchangeable. For external primary

counting, the minimum-row-ID representative of each identical byte hash was used within its acquisition source. One United States duplicate row was thereby removed from counting, while its original record was preserved.

Table 1. Locked evaluation cohorts and their counting units.

Setting and unit	Acquisition	Source rows	Unique images	Analysis units	Positive units	Protected groups
Saudi Arabia report	Municipal reports	4,778	4,593	4,179	2,802	4,171
China drone image	Drone	1,918	1,918	1,918	43	1,817
China motorbike image	Motorbike	1,597	1,597	1,597	137	1,172
Czechia image	Vehicle smartphone	2,244	2,244	2,244	119	2,095
India image	Vehicle smartphone	6,153	6,153	6,153	1,212	5,910
Japan image	Vehicle smartphone	8,314	8,314	8,314	1,090	8,207
Norway image	Vehicle camera	6,595	6,595	6,595	206	5,625
United States image	Google Street View	3,852	3,851	3,851	95	3,817

** Saudi risk is defined over original reports; external risk is defined over byte-unique images.*

Protected groups and fitting roles

Before model fitting, connected components were formed from Saudi report identifiers, exact SHA256 image matches and qualifying perceptual-neighbour edges. Perceptual screening used a 64-bit discrete-cosine perceptual hash and a 64-bit difference hash, followed by a grayscale root-mean-square difference check. Candidate pairs were joined when their respective Hamming distances were at most 6 and 12 and the grayscale difference was at most 30. The screen yielded 5,099 qualifying edges. Transitive components were assigned to roles. Perceptual neighbours remained distinct observations; identical views were deduplicated within reports. Figure 3 summarizes the fitting rotation and report representation.

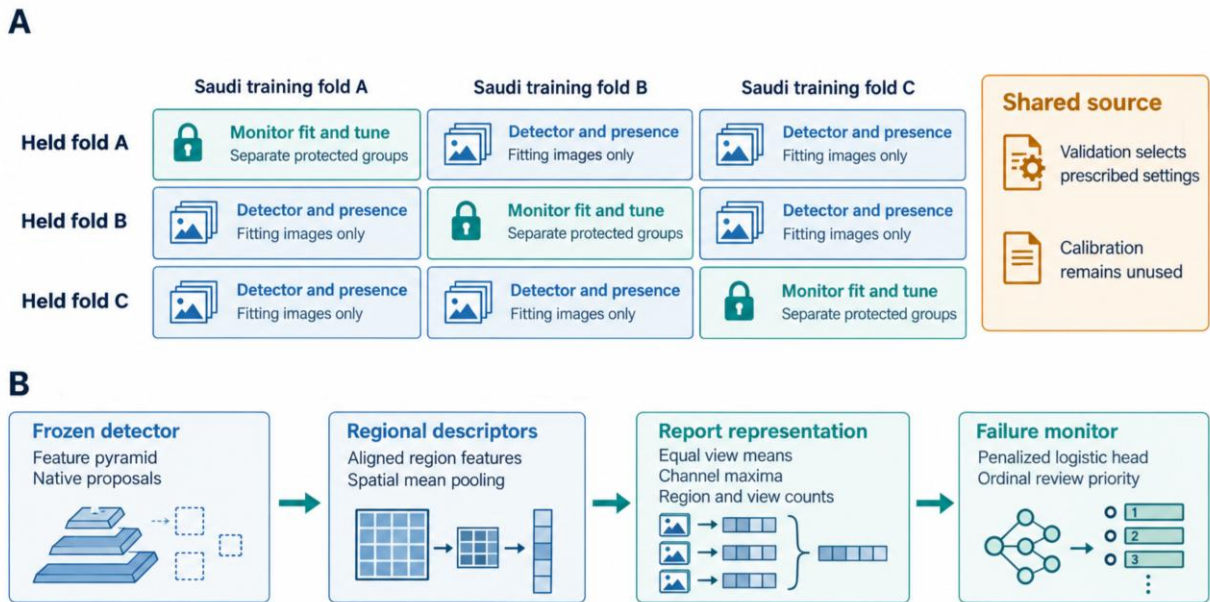


Figure 3. Protected fitting roles and the construction of report-level failure evidence. Panel A shows the rotation repeated for each optimization seed; diagram folds A, B, and C correspond to stored folds 0, 1, and 2. Panel B separates detector-derived regional pooling from subsequent report supervision.

Deterministic, label-free hashing of protected-group IDs assigned Saudi groups to training, validation, calibration, and locked test with nominal fractions 0.60, 0.15, 0.10, and 0.15. External groups were assigned to reserve or locked evaluation with nominal fractions 0.20 and 0.80. Exact assignment salts and thresholds are retained with the code. No report ID, exact image hash, or qualifying similarity edge crossed labelled outer partitions. Saudi calibration and external reserve were unused in the main protocol; neither probability calibration nor target-domain adaptation was subsequently introduced. The resulting Saudi role counts are provided in Table S1.

The Saudi training partition was divided by protected group into three folds (Figure 3). For each held fold, detector and presence models were fitted on the other two folds; groups in the held fold were allocated approximately 80:20 to monitor fitting and monitor tuning. Thus, the failure monitor learned from predictions generated on images excluded from its paired detector's fitting data. Seeds 20261020, 20261029, and 20261030 crossed with held folds 0, 1, and 2 produced nine

matched configurations (Figure 3). These are three optimization seeds applied to overlapping training rotations. The same fold-specific detector and monitor were retained for evaluation.

Models and report failure representation

The detector was TorchVision Faster R-CNN with a MobileNetV3-Large backbone and feature pyramid, drawing on region proposal networks, multiscale feature fusion, and efficient convolutional design (Howard et al., 2019; Lin et al., 2017; Ren et al., 2017). Official COCO weights were used (Lin et al., 2014); the final predictor was replaced by a newly initialized pothole/background head. No road-trained pilot checkpoint was reused. Images were resized with a minimum side of 320 pixels and a maximum side of 640 pixels. The raw detection score floor was 0.01 and at most 100 final detections were retained per image. The last three backbone layers were trainable, while batch-normalization running statistics were frozen. Native RPN and RoI classification and box-regression losses were summed. Table 2 records the executed optimization recipes.

Table 2. Executed model fitting and selection recipes.

Component	Fitting recipe	Selection and eligibility
Pothole detector	3 epochs; unique-image batch 2; SGD momentum 0.9; weight decay 0.0005; body learning rate 0.0002; FPN, RPN, and head rate 0.002; gradient norm limit 10.	Final epoch only. Both rates multiplied by 0.1 after floor($U/2$) and floor($4U/5$) updates.
Frozen presence controls	ImageNet MobileNetV3-Small features; 576 dimensions; separate centre-crop and letterbox views; penalized logistic heads.	Four regularization values; source-validation image log loss. Six unique heads shared across seeds within folds.
Fine-tuned presence	8 epochs; batch 16; AdamW feature rate 0.00001, head rate 0.0001; weight decay 0.0001; gradient limit 1; horizontal flip probability 0.5.	Final epoch only. Initialize from ImageNet and the paired fold's letterbox head and normalization.
Report logistic heads	Equal report weight; fitting-only standardization; L-BFGS up to 300 iterations; maximum finite gradient at most 0.00001.	Regularization in $\{0.001, 0.01, 0.1, 1\}$; monitor-tuning report log loss; larger value breaks ties.

Regional-max comparator	256→512→256→64→32→16→1 ReLU network; AdamW rate 0.001; 20 epochs; batch 128; gradient limit 5.	Weight decay in {0.001, 0.01, 0.1, 1}; maximum-region report log loss; fixed monitor seed 20261027.
----------------------------	--	---

* U denotes the prescribed detector-update budget. The final singleton image batch was retained. AdamW follows the decoupled weight-decay method (Loshchilov & Hutter, 2019). Presence models and report monitors have different supervision units and capacity; they are not asserted to be matched in complexity.

The detector completed 18,321, 18,393, or 18,126 updates according to held fold. The fine-tuned presence model completed 6,112, 6,136, or 6,048 updates, respectively. Image permutations, augmentation draws, learning rates, losses, gradient norms, and recovery states were logged. The centre view used a bilinear short-side resize to 256 pixels followed by a 224×224 centre crop; the alternative was an aspect-preserving 224×224 letterbox. The letterbox used bilinear interpolation and RGB padding (124, 116, 104). Both views used the recorded ImageNet normalization (Table S3), and neural presence fitting also froze batch-normalization running statistics. Standardization was estimated only from fitting images; the fine-tuned model retained its fitted head normalization.

For each final detector, the operating confidence was chosen on the 4,535 byte-unique Saudi validation images, maximizing global object F1 at intersection-over-union (IoU) 0.50 over the grid {0.05, 0.10, ..., 0.95}. Exact integer fractions were compared, with the higher confidence selected on an exact tie:

$$\tau_{sk} = \arg \max_{t \in T} \frac{2TP(t)}{2TP(t) + FP(t) + FN(t)} \quad (1)$$

Here s and k index seed and held fold; TP, FP, and FN denote true positives, false positives, and false negatives. Detections were matched greedily, one-to-one, in descending confidence; coordinate order broke score ties, and the lowest reference index broke equal-IoU ties. Inclusion used scores and IoUs greater than or equal to their thresholds. The selected confidence was propagated to every emitted-box feature, regional target, and report outcome. Eight configurations selected 0.40; seed 20261029 with held fold 2 selected 0.35. Selection did not use report recovery, test budgets, or external labels.

Let Y_i indicate at least one reference pothole in the images owned by unit i , H_i indicate at least one emitted detection overlapping an own-image reference at the specified IoU, and Z_i indicate that the unit emits no boxes. Complete, silent, and emitted complete misses were then defined as:

$$M_i = Y_i(1 - H_i), \quad M_i^{\text{silent}} = M_i Z_i, \quad M_i^{\text{emitted}} = M_i(1 - Z_i) \quad (2)$$

Matching never crossed image boundaries. A match to any source object in any own view prevents a complete report miss, even when other objects remain undetected. Conversely, a negative report emitting a box is a source-reference false alarm. Figure 4 makes this distinction explicit. Regional missed-object targets were defined separately: a proposal was positive when it overlapped a source object at IoU at least 0.50 and no emitted detection overlapped that object at the same threshold. Such a target can occur within a report that is not a complete miss.

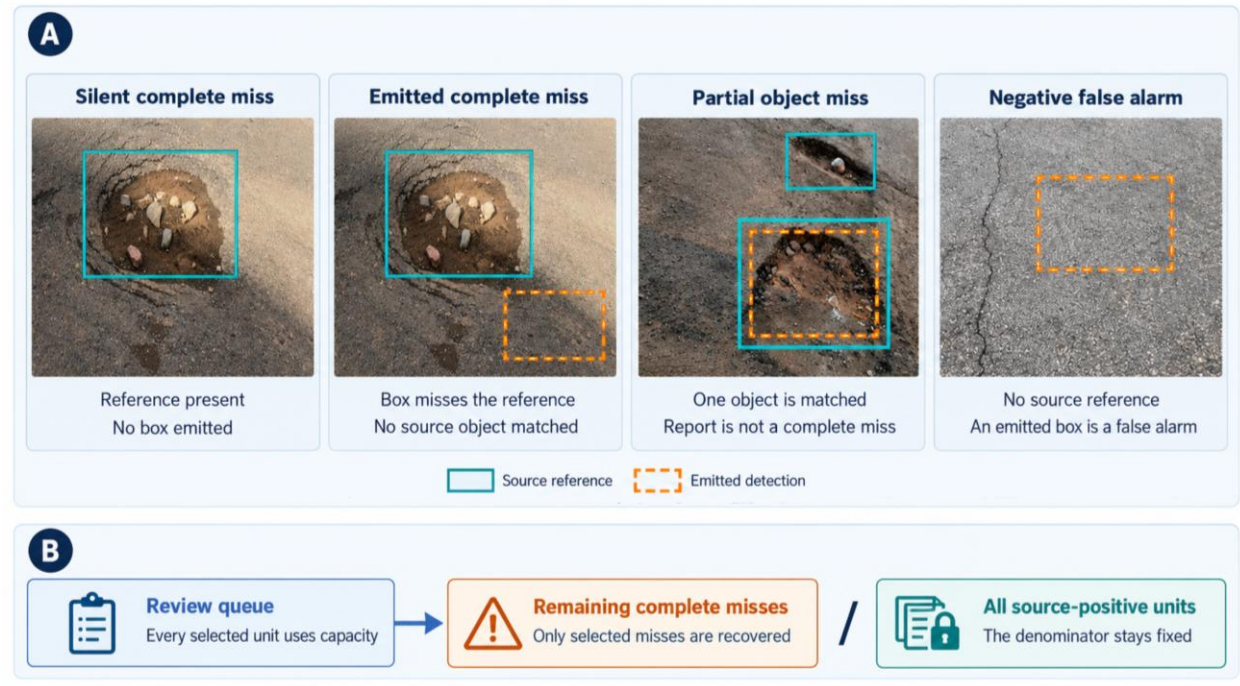


Figure 4. Complete report failure, partial object omission, and fixed-budget review accounting. Panel A distinguishes silent complete misses, emitted but incorrect complete misses, partial misses, and source-negative false alarms. Panel B identifies the fixed denominator used after oracle review: all source-positive units remain in the denominator, including those selected for review.

* All boxes are schematic. An overlapping reference and detection in the partial-miss panel indicates a match; the other source object remains undetected.

Regional representations were extracted from the first 128 native, post-nonmaximum-suppression RPN proposals in returned objectness order. Native multiscale RoIAlign used 7×7 outputs, spatially averaged to 256-channel descriptors (He et al., 2017); no mask-prediction task was introduced. For byte-unique views V_i , each image’s regional mean was averaged equally across views, while its channel-wise maxima were pooled by a second maximum:

$$\mu_i = \frac{1}{|V_i|} \sum_{v \in V_i} \mu_{iv}, \quad a_{ic} = \max_{v \in V_i} a_{ivc} \quad (3)$$

Here μ_{iv} and a_{ivc} are the within-image regional mean and channel maximum; empty-image vectors were set to zero before report pooling. Four count features described unique views, total proposals, proposals per view, and empty-view frequency. Concatenating these with regional means and maxima yielded 516 features. Adding six emitted-detection summaries yielded the 522-dimensional failure representation. Exact feature definitions, normalization, and sparse-supervision defaults appear in Table S3. Feature construction rejected source-reference fields; labels were joined separately for supervised monitor fitting and outcome calculation. Memory-mapped regional shards and report-wise streaming prevented a requirement to retain all normalized regional embeddings in memory.

For a standardized fitting representation $\tilde{x}_i$, report failure probabilities were estimated using a sigmoid logistic head with an unpenalized intercept and equal report weight:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^n [-M_i \log p_i - (1 - M_i) \log(1 - p_i)] + \frac{\lambda}{2} \|w\|^2 \quad (4)$$

$$p_i = \sigma(w^T \tilde{x}_i + b)$$

Optimization used limited-memory BFGS (Liu & Nocedal, 1989), with tolerances given in Table 2. Here w and b are coefficients and intercept, n is the number of fitting reports, and λ is the regularization value. Selection used monitor-tuning log loss, with stronger regularization preferred on an exact tie. This is selection for a specified supervised target, not a claim that the resulting probabilities were calibrated. The regional-max comparator instead predicted missed-object regions

and aggregated their probabilities by a report maximum. Its target and aggregation therefore differ from complete-report failure.

Comparison policies and ablations

All 21 policy definitions were frozen before main test release: 18 deployable rankings and three reference policies. The primary candidate was M19, the regional-plus-detector failure head with no-box-first ordering. It was compared with both the centre-crop and letterbox presence controls, M06 and M07; neither control could be discarded after observing results. For maximum own-view presence p_i and maximum emitted confidence d_i , with d_i set to zero when no box was emitted, the principal scores were:

$$\begin{aligned} S_{M19,i} &= 2Z_i + p_{M18,i} \\ S_{M06/M07,i} &= 2Z_i + p_{\text{presence},i} - d_i \end{aligned} \quad (5)$$

The no-box stratum precedes the emitted-box stratum for all three primary policies, allowing the regional representation to be assessed beyond that simple ordering rule. Table 3 summarizes the comparison logic; Table S2 gives the complete policy inventory. The ablations isolate count information, regional pooling, detector summaries, and the ordinal wrapper in sequence. Their interpretation is conditional on the fixed representations and supervision; differences do not establish equal-capacity comparisons or general architectural superiority.

Table 3. Comparison families and the methodological question isolated by each family.

Family	Policies	Comparison purpose
Observable cues	M02–M05	Detector confidence, image quality, and raw presence without learned report-failure supervision.
Strong presence controls	M06, M07	Common no-box-first ordering with centre-crop or letterbox target-presence evidence.
Failure and fusion controls	M08–M12	Detector summaries, presence fusion, a stratum prior, and separate silent/emitted failure branches.
Presence capacity check	M13, M14	Fine-tuned presence with ordinal ranking or a learned report-failure fusion head.
Regional missed-object alarm	M15	Maximum regional alarm under a distinct missed-object supervision target.

Representation ablations	M16, M17, M18	Counts only → counts plus regional pooling → added detector summaries.
Ordinal candidate ablation	M18→M19	The same failure probability with versus without no-box-first ordering.
Reference limits	M00, M01, M20	No review, uniform-random expected recovery, and unattainable perfect oracle allocation.

* Only M19 versus M06 and M19 versus M07 define the confirmatory comparison family.

Locked evaluation and statistical inference

After all model and threshold audits, the training completion freeze was recorded on 23 October 2026 at 18:48:21 UTC. Each of the nine matched configurations was applied to Saudi Arabia and the seven external acquisition settings, giving 72 canonical prediction runs. Prediction streams were sealed before outcomes were joined for analysis. External inference retained the source-trained weights, feature normalization, confidence, and policy definitions unchanged. Saudi report-level and external image-level risks were analyzed separately because their units, annotation processes, and acquisition distributions differ.

For a cohort of N units and review fraction B , the selected set Q contains the K highest priorities, with ascending unit ID breaking score ties. At the primary budget $B = 0.10$, the Saudi set therefore contained 417 reviewed reports. Source-negative selections consumed capacity. The unresolved complete-miss risk was:

$$K = \lfloor BN \rfloor, \quad R_{m,sk}(B) = \frac{\sum_{i=1}^N M_{i,sk} - \sum_{i \in Q_{m,sk}(B)} M_{i,sk}}{\sum_{i=1}^N Y_i} \quad (6)$$

The denominator is every source-positive unit. No-review risk was retained; uniform-random expected recovery was calculated analytically, and the oracle ceiling could recover at most the smaller of the review capacity and the number of complete misses. Additional fixed budgets were 0.05, 0.20, and 0.40. Object-level F1, report false alarms, silent/emitted recovery, failure discrimination, and Brier score or log loss for probability-valued heads were treated as diagnostics. Proper probability scoring was kept separate from ordinal ranking evaluation ([Gneiting & Raftery, 2007](#)).

For control c, the primary contrast was the equal fold mean within each seed, followed by the equal mean across seeds:

$$\Delta_c(B) = \frac{1}{9} \sum_{s=1}^3 \sum_{k=0}^2 [R_{M19,sk}(B) - R_{c,sk}(B)] \quad (7)$$

Lower values favour the candidate. All nine configuration results and the three seed means were retained; configurations were neither selected nor treated as independent test samples. Uncertainty was estimated using 10,000 paired protected-group bootstrap resamples, with random seed 20261031. Observed groups were sampled with replacement, keeping their units together and using the same group multiplicities for both controls and every configuration. Cohort size, integer review capacity, positive denominator, and ranked recovery were recomputed within each resample. This preserves the dependence recognized by the grouping construction, consistent with the rationale for cluster-aware resampling ([Field & Welsh, 2007](#)).

Two-sided 97.5% percentile intervals were computed for each of the two primary contrasts, providing the prespecified Bonferroni adjustment. Conditional superiority required both upper bounds to be below zero. Replicates without positive references were omitted and counted; interval reporting was suppressed if more than 5% were invalid. These intervals are conditional on the observed protected groups and fixed trained configurations.

Sensitivity analyses changed the evaluation IoU to 0.30 and 0.75 while retaining the original heads and ranking inputs, and separately imposed confidence 0.25 while recomputing emitted features and policy scores with the same fitted heads. The latter is an operating-point stress test, not recalibration. External evaluation was also repeated without exact-byte deduplication. None of these analyses supplied new fitting labels or altered the primary frozen comparison. Independent expert error labels were not available for an empirical human learning-to-defer comparison; optional annotation exercises were excluded from the main supervised targets and primary risk estimates.

Execution integrity and computational accounting

Training and evaluation were executed locally on an arm64 MacBook with 18 GiB physical memory and macOS 15.6, using CPU execution with four threads. The recorded environment comprised Python 3.12.14, PyTorch 2.14.1, TorchVision 0.29.1, NumPy 2.3.5, and Pillow 12.3.0. Python, NumPy, and PyTorch random states were recorded, deterministic algorithms were enabled, and optimizer states, batch order, learning rates, and checkpoint hashes were retained. Reporting the execution environment and recoverable computational record follows established reproducibility guidance ([Pineau et al., 2021](#)).

During detector fitting on 22 October 2026, an empty region-of-interest set in a batch of two negative images produced an undefined mean loss. Amendment A1 returned differentiable zero classification and box-regression losses only for that consistent empty set, while retaining RPN losses and rejecting other nonfinite states. Nonempty losses and gradients were unchanged in parity checks. Recovery resumed the committed checkpoint at update 8,128 and reproduced the next two journaled updates before continuing the prescribed schedule. No images, optimizer steps, or failed logs were removed, and the earlier four completed configurations were retained.

During the first India evaluation on 24 October 2026, an EXIF orientation check stopped prediction after 2,269 images. Amendment B1 enforced the pre-existing stored-raster convention through a hash-, size-, and orientation-checked allowance for exactly two India files; it did not rotate images, transform annotations, or omit cases. The completed prefix was retained byte-for-byte, geometry and prediction-prefix parity were checked, and a separately written suffix completed the canonical stream. Original failed-attempt records and the recovery registry were preserved.

Table 4. Measured computation intervals and their accounting boundaries.

Recorded interval	Elapsed hours	Accounting boundary
Training end to end	49.930	21–23 October 2026; includes interruption, recovery, and audit.
Training stage attempts	48.341	Nested within the training interval; includes failed

		attempts.
Evaluation end to end	26.569	23–24 October 2026; includes diagnosis, recovery, and source analysis.
Prediction attempts	20.544	Nested within evaluation; includes the interrupted India prefix attempt.
Independent arithmetic audit	0.026	Separately measured after evaluation; approximately 94 seconds.

** These are recorded wall-clock intervals, not CPU-hours or active researcher effort. Nested intervals overlap and must not be summed. Rounding is to three decimal places; original timestamps and durations remain in the timing ledgers.*

Integrity checks linked source manifests, coordinate conversion, role tables, model states, prediction streams, and analyses through SHA256 records. A separate arithmetic implementation reconstructed the comparison calculations from sealed predictions. A post-evaluation demonstration reran 16 fixed units through the frozen model and checked exact scientific-output agreement; its native-window video documents execution on recorded images. Display pauses were excluded from any interpretation as an inference-speed benchmark.

Code, frozen design documents, and numerical evidence are maintained in the private [PAVE repository](#), with an [archived repository commit](#) providing a fixed code and evidence reference. Raw image archives, full trained weights, prediction streams, and original videos are retained separately; the repository alone is therefore not a complete neural-retraining package. Source-specific licences and provenance remain applicable.

Results

Primary endpoint and failure burden

All 9 training configurations completed their audits, and all 72 prediction runs were sealed before outcome analysis. The primary Saudi evaluation comprised 4,179 original reports, including 2,802 source-positive reports. Before review, the mean number of complete misses was 1,030.56: 684.11 silent misses and 346.44 misses with emitted but incorrect boxes. Mean object F1 was 0.5894. Unless specified otherwise, results below are equal-weight means across the 3 seeds crossed with 3 held folds.

At the prespecified 10% budget, 417 reports could be reviewed. Residual complete-miss risk was 24.63% for the candidate M19, compared with 25.04% for the center control M06 and 25.01% for the letterbox control M07 (Table 5). The candidate recovered 11.56 and 10.78 additional complete misses, respectively. Its adjusted 97.5% intervals crossed zero in both comparisons. Superiority over both controls was therefore not established; the intervals also do not establish equivalence.

Table 5. Primary Saudi results at the fixed review budget

ID	Comparison	Risk (%)	Recovered reports	M19 minus control pp [97.5% CI]
M00	No review	36.78	0.00	—
M01	Random expectation	33.11	102.83	—
M06	Center control	25.04	329.00	−0.412 [−0.849, +0.107]
M07	Letterbox control	25.01	329.78	−0.385 [−0.864, +0.087]
M13	Fine presence	24.33	348.78	—
M19	Candidate	24.63	340.56	—
M20	Oracle reference	21.90	417.00	—

** Risk uses all 2,802 source-positive reports as denominator. Fractional counts are configuration means. M13 is descriptive; M20 is an unattainable outcome-informed reference.*

The candidate had lower point risk than both controls in 8 of the 9 configurations, and the mean differences were negative for each of the 3 seeds. These correlated repetitions did not alter the

inferential conclusion. Figure 5 places the primary contrasts alongside the fixed budget curve; per-configuration values are retained in Table S5. Although the candidate achieved lower mean unresolved complete-miss risk than both controls at the prespecified 10% review budget, the adjusted paired confidence intervals crossed zero (Figure S4).

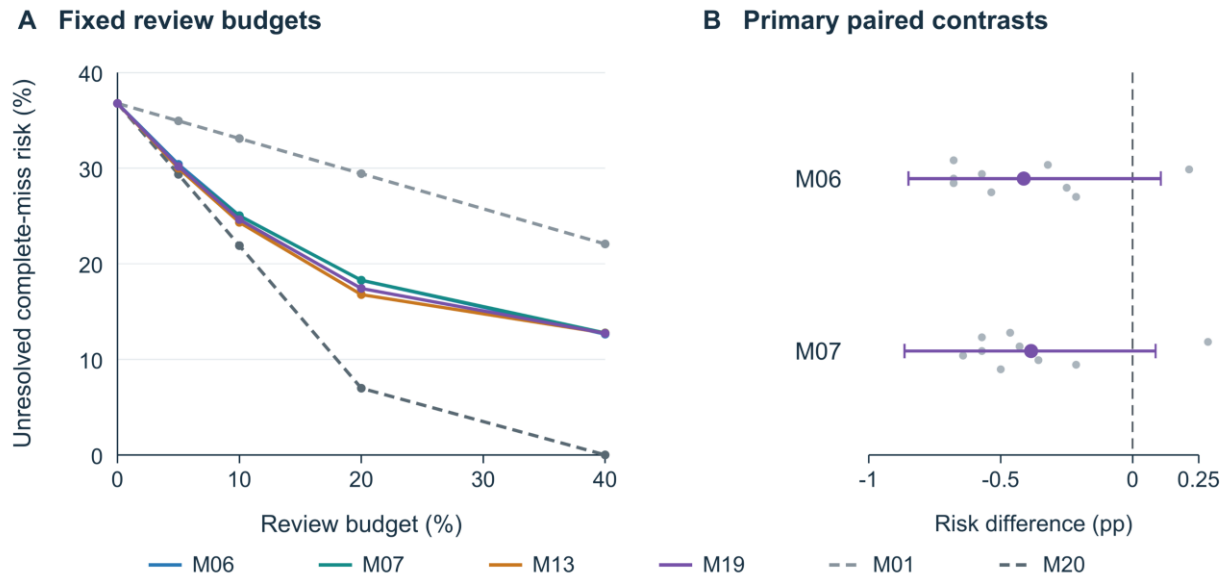


Figure 5. Review-budget dependence and primary paired effects. (A) Mean unresolved complete-miss risk at the 5 fixed budget points; lines connect evaluated points. M01 denotes random expectation and M20 the oracle reference. (B) M19-minus-control risk differences at 10% review. Purple points and bars show the mean and adjusted 97.5% interval; gray points show the 9 correlated configurations.

* Negative differences favor M19. Pp: percentage points.

Comparison policies and external transfer

In every configuration, M19, M06, and M07 filled all 417 slots with reports having no emitted box. All recovered complete misses were consequently silent; none was an emitted complete miss. For M19, the residual burden averaged 343.56 silent and 346.44 emitted complete misses, while 76.44 reviewed reports were source-negative. Figure 6 separates these allocations from the fixed feature sequence.

Risk decreased from 30.06% for counts alone to 27.57% with regional pooling, 24.75% after adding detector summaries, and 24.63% after no-box-first ordering. The fine presence policy M13 nevertheless had the lowest observed Saudi risk among the 18 deployable policies (24.33%); M14 followed at 24.38%. The regional maximum policy M15 remained substantially weaker (31.91%). M18 had mean failure-prediction AUROC/AP of 0.8209/0.6566, compared with 0.7788/0.6268 for ordinal M19; better allocation at one budget did not imply better global discrimination. Full diagnostics are given in Table S4.

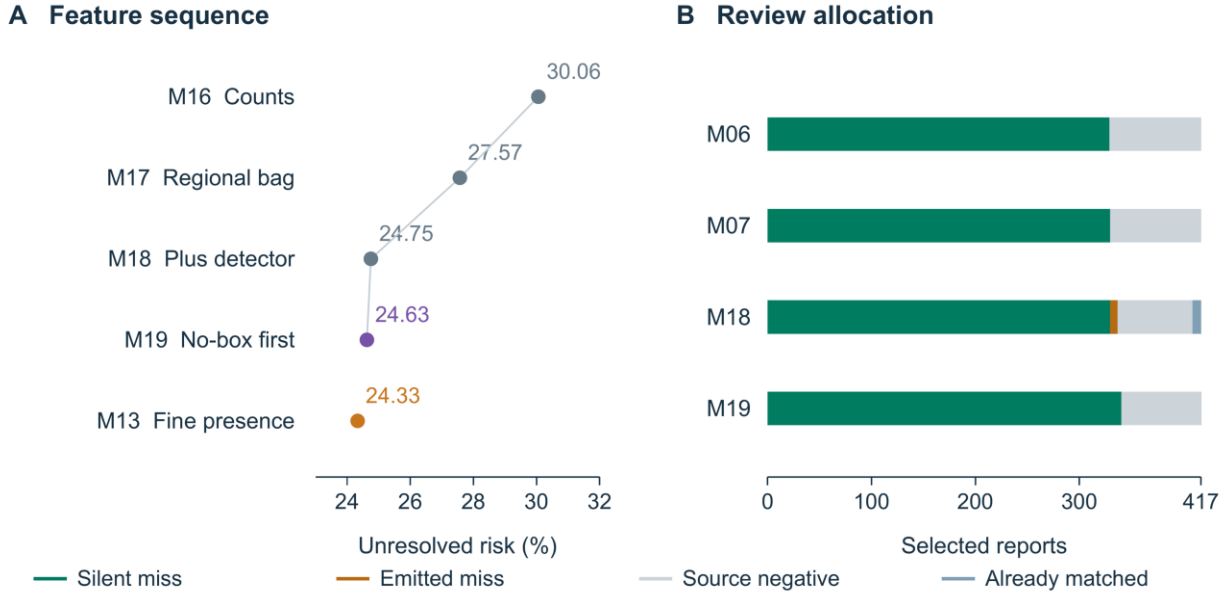


Figure 6. Feature comparisons and the composition of the review queue. (A) Saudi risk at 10% review for the fixed counts, regional-bag, detector-summary, and no-box-first sequence, with fine presence M13 shown separately. (B) Mean selected reports partitioned by source-reference outcome. Each bar totals 417 reviews; narrow segments are retained at their exact widths.

** Feature comparisons are descriptive: feature dimensionality and supervision differ. Recovered emitted misses average 7.33 for M18 and 0 for M19 and both primary controls.*

Transfer was heterogeneous (Table 6; Figure 7). In Japan, M19 reduced risk relative to both controls by 2.130 and 2.212 percentage points, with source-specific adjusted intervals entirely below zero. On China-drone images, risk was higher by 7.494 points versus M06; the interval excluded zero. The corresponding M07 interval reached zero and did not establish a strictly positive lower bound. India favored M19 in both point comparisons, but both intervals crossed zero. The United States favored M19 against M06 and disfavored it against M07. No external result changed the primary Saudi verdict. Exact external intervals are reported in Table S7.

Table 6. External source results under the fixed operating protocol.

Source	Positive images	Object F1	M19 risk (%)	M06 risk (%)	M07 risk (%)
China drone	43	0.0919	51.16	43.67	46.77
China motorbike	137	0.0775	86.94	86.94	86.46
Czechia	119	0.0000	84.97	79.74	78.34
India	1212	0.0185	81.08	82.57	82.42
Japan	1090	0.0411	77.50	79.63	79.71
Norway	206	0.0010	85.60	83.50	82.20
United States	95	0.0084	87.60	88.77	83.86

** External units are byte-unique images. Values are means across 9 configurations; risk uses 10% review and F1 precedes review; source-positive image counts are risk denominators, not numbers of detected objects. External analyses are secondary, with no multiplicity control across sources. The Saudi report endpoint is not pooled with these image endpoints.*

Mean external object F1 ranged from 0.0000 to 0.0919, substantially below the Saudi value. No object-level true positives were recorded in the Czech collection at the primary operating points in any configuration. The complete source-by-policy pattern is retained in Figure S1; these collection-specific measurements are not estimates of national pavement condition.

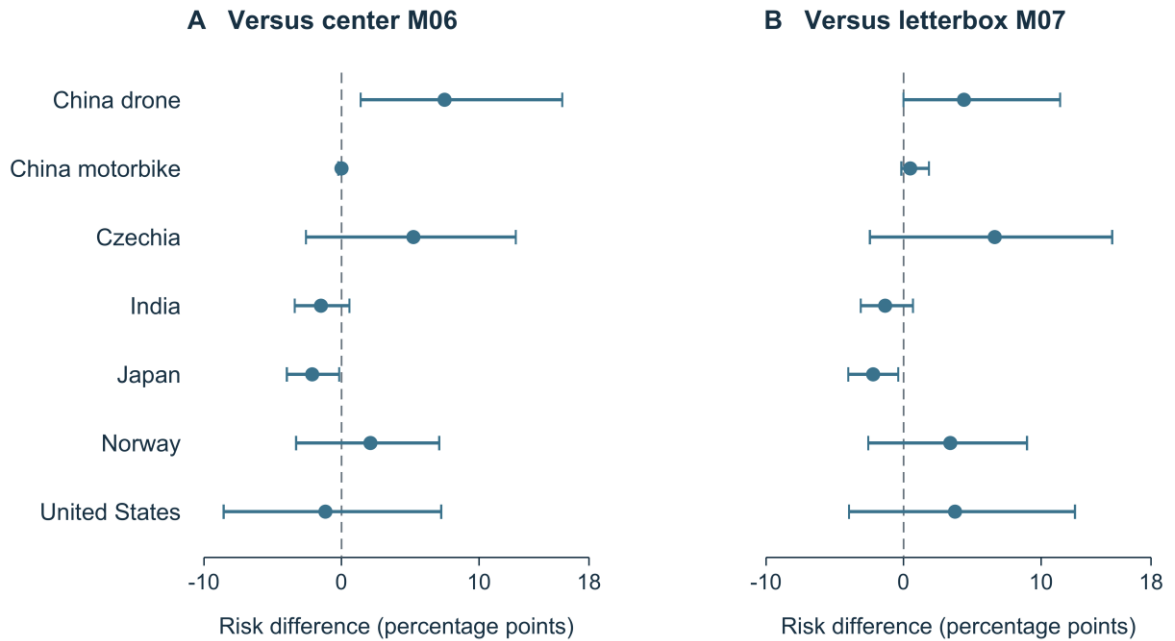


Figure 7. External candidate-minus-control effects. Points and bars show mean risk differences and source-specific 97.5% paired protected-group intervals at 10% review, relative to (A) M06 and (B) M07. Each acquisition source is analyzed separately.

** Negative differences favor M19. Adjustment covers the 2 contrasts within each source, not the family of external sources. The near-zero China-drone M07 lower limit is treated as zero within numerical precision.*

Sensitivity analyses and evidence checks

Changing the matching IoU from 0.50 to 0.30 reduced candidate risk to 20.10%; requiring 0.75 increased it to 45.45%. The primary candidate-minus-control differences and intervals remained numerically unchanged because all selected reports were no-box cases, whose positive references remained missed at every IoU. The confidence-0.25 stress analysis gave 20.28% candidate risk and negative intervals for both contrasts, but object F1 decreased to 0.5678. This secondary operating condition did not replace the primary thresholds (Figure S2; Table S6). Retaining the single United States duplicate changed interval quantiles slightly without changing the primary point estimates.

The independent arithmetic audit reconciled 36,855 curve records and 6,318 diagnostic records across 39 source/condition combinations. A separate implementation reproduced the Saudi paired bootstrap arrays to a maximum absolute discrepancy below 2.7×10^{-17} . The recorded 16-unit frozen-model demonstration reproduced all 16 scientific outputs exactly, excluding historical timing and memory measurements, with all 67 checked frozen files unchanged. Recorded end-to-end intervals were 49.930 h for main training and 26.569 h for evaluation; their nested attempt durations were not added together.

Figure 8 displays authentic matched, partially missed, silent, and emitted-failure examples from the fixed first model. Additional negative examples are shown in Figure S3, with per-image precision, recall, confidence, and matched overlap in Table S8. These previously outcome-selected images illustrate distinct errors and cannot estimate their frequency. Detector confidence is not empirical precision, and source-reference agreement does not establish physical pothole truth.



Figure 8. Frozen detections across four image-level outcomes. Solid green boxes denote supplied references; dashed amber boxes denote saved detections at confidence ≥ 0.40 . Box labels are detector confidence. At IoU 0.50, TP/FP/FN counts are (A) 1/0/0, (B) 1/1/1, (C) 0/0/1, and (D) 0/1/2. Corresponding precision is 1.00, 0.50, undefined, and 0.00. A partial object miss is compatible with a report-level hit.

* Fixed model *s20261020_f0*; stored-raster orientation retained. Examples were selected after outcomes were known. Images: (AlElaiwi, Ahmad, et al., 2023), CC BY 4.0; detection overlays added. Export resolution does not increase native image detail.

Adding regional features and detector summaries progressively reduced mean unresolved risk, while no-box-first ordering provided a smaller additional reduction; these differences are descriptive and do not establish causal effects (Figure S5).

Interpretation and limitations

What the locked comparison establishes

The candidate improved the mean allocation of review effort. At 10% review, unresolved complete-miss risk was 24.63%, versus 25.04% and 25.01% for the controls; both adjusted intervals crossed zero (Table 5). Moreover, fine presence M13 had the lowest observed deployable-policy risk, 24.33%, without a new adjusted superiority test. Richer regional representations therefore did not establish additional operational value over the strongest presence baseline in this experiment.

All 417 candidate review slots contained no-box reports in every configuration. The candidate could improve ordering within that stratum, but the 346.44 mean emitted complete misses remained outside review (Figure 6). Under the observed allocation, improving within-stratum ranking alone cannot recover those failures: the priority rule or allocation across strata must also change.

A cost-sensitive detector framework treats missing and false detections explicitly under user-defined error budgets ([Kassem-Sbeyti et al., 2024](#)). Here, the budget counted reviewed reports, and no-box priority effectively excluded an entire failure type. Global discrimination was also insufficient to describe allocation: M18 had higher AUROC than M19, yet slightly higher observed residual risk at the primary budget. A useful ranking over all reports need not provide the best first 417 selections; M19’s ordinal priority score should not be read as a calibrated failure probability.

Japan favored the candidate against both controls, but mean detector F1 there was only 0.0411; improved review ordering coexisted with weak detection. China-drone results favored the center control, with only 43 source-positive images contributing to risk estimation (Table 6). These patterns provide no basis for universal transfer. Cross-scene road-damage improvements have instead been reported after adversarial feature adaptation ([Li et al., 2026](#)), while false-negative prediction has

been used to guide target-domain acquisition and adaptation ([Nakamura et al., 2024](#)). Those interventions differ from unchanged Saudi-trained deployment.

Implications for inspection and system design

For a road authority, review priority ranks reports for potential correction of source-reference misses; it does not rank pavement severity or repair urgency. Abaddi's camel–vehicle collision framework connects Gulf road-hazard perception with tracking and warning ([Abaddi, 2026a](#)). The relevant connection is the transition from visual evidence to a response.

A practical next comparison is therefore a fixed-budget allocation that reserves some reviews for emitted-box reports, evaluated against the same presence baselines. Its value remains untested. Independent pavement assessment should record reviewer accuracy and elapsed review time, with blinded reference construction and a protected prospective evaluation. Human-only, automated, and assisted workflows need separate comparison: a meta-analysis found that human–AI combinations, on average, fell short of the better standalone constituent ([Vaccaro et al., 2024](#)). Perfect-review recovery cannot establish such complementarity.

Threshold selection belongs in the same operational specification as the review budget. At confidence 0.25, both secondary contrast intervals were negative, but detector F1 fell from 0.5894 to 0.5678 (Table S6). Changing emission thresholds changes the reports entering each queue stratum. Post-processing can also alter detector calibration ([Huseljic et al., 2025](#)); the present stress analysis did not test calibration as a causal explanation. The stress result defines a trade-off requiring new validation; it cannot replace the locked primary operating points.

Limitations

The endpoint measures agreement with supplied annotations. No independent pavement-expert ground truth was established, and a source-negative image does not prove physical absence of damage. A report with any matched target is counted as a hit even when other targets are missed; complete-miss risk therefore omits potentially important partial failures. Review was assumed to correct selected failures perfectly. Actual reviewer errors, inspection time, repair decisions, and safety outcomes were not measured.

Generality is further limited by one compact detector family, a bounded training schedule, and 3 optimization seeds crossed with overlapping folds. Bootstrap intervals condition on these fixed models; they do not capture every source of training variability. Protected grouping and exact-byte checks cannot guarantee independence between all nearby views or routes, and acquisition collections are not national samples. The main protocol was locally frozen after engineering development. Multiplicity control covered the 2 primary contrasts, without adjustment across the family of secondary sources, budgets, or sensitivities. Favorable secondary results remain insufficient to reverse the primary conclusion.

Conclusion

Regional failure prediction produced small average gains over both presence controls, but the locked comparison did not establish superiority. At 10% review, unresolved complete-miss risk was 24.63% for the candidate, compared with 25.04% and 25.01% for the controls; both adjusted intervals included zero. Fine-tuned presence achieved the lowest observed deployable-policy risk, 24.33%, as a descriptive result. These findings leave the incremental value of the regional candidate uncertain and support retaining strong presence baselines in subsequent comparisons.

The principal contribution is a reproducible assessment of which detector failures remain after a fixed amount of review. The allocation diagnostic was particularly informative: every candidate review concerned a no-box report, while a mean 346.44 emitted-but-incorrect complete misses remained unreviewed. Within this allocation, better ranking among no-box reports alone

cannot recover emitted failures. The preserved protocol, full policy comparisons and independently checked calculations make that constraint inspectable. Heterogeneous external results further show that favorable review ordering in one collection can coexist with weak detection, and does not establish transferable inspection performance.

The next empirical test should compare an explicit allocation between no-box and emitted-box reports at unchanged total review capacity, against the same presence controls. A prospective protocol should jointly fix the emission threshold and allocation rule before evaluation on new, protected acquisition groups. Independent pavement assessment should measure actual correction accuracy and review time, and record partial misses and physical severity as separate outcomes. Broader detector families and new routes or camera settings are also needed to test generality; target-domain adaptation should be evaluated as a separate intervention.

Declarations

Conflict of interest

The author declares no financial or non-financial competing interests.

Funding

The study received no external funding.

Data availability

Source images and annotations are available from the [Saudi Arabia Public Roads Visual Pollution Dataset version 4](#) and [RDD2022](#). Access routes, versions and processing rules are documented in the Methods and PAVE. Numerical inputs and aggregate results are held in the private repository; raw images, weights and full prediction streams are stored separately. Source-provider terms apply.

Code availability

Code, the frozen protocol, software specifications and numerical reproduction instructions are deposited in the private [PAVE GitHub repository](#) (Supplementary video 1). The snapshot verified on 25 October 2026 is commit [545485b57e8d50a41f4557b348895c0dff244341](#). Access requires author authorization; full neural replication also requires separately retained assets. Third-party data and software terms apply.

Ethics approval

The main benchmark analyzed public road images, source annotations and model outputs. No new field acquisition or intervention was undertaken. Exploratory annotation and an optional volunteer pilot were excluded from the main analysis; independent expert validation was not established. No institutional ethics approval or formal exemption is claimed.

Consent to participate and publication

Not applicable to the secondary-image benchmark. No participant records, reviewer identities or private correspondence are reported. Source-image reuse remains subject to provider permissions and attribution.

Clinical trial registration

Not applicable. This computational road-image study involved no clinical intervention.

AI assistance

AI was used for proofreading, improving the flow and correcting typos. Responsibility for the scientific content and final manuscript rests with the author.

Supplementary videos

Video 1. Private PAVE GitHub repository walkthrough: access requires author authorization.

Video 2. Live frozen-model inference on 20 recorded road-image units, showing presence classification, bounding boxes and review priority, including failures.

References

- Abaddi, S. (2026a). Quantum-omni: A quantum computing and GPT-4o solution for camel-vehicle collisions. *International Journal of Transportation Science and Technology*, 21, 306–321. <https://doi.org/10.1016/j.ijtst.2025.02.002>
- Abaddi, S. (2026b). Urban road traffic management system using proactive simulation models. *World Review of Intermodal Transportation Research*, 12(4). <https://doi.org/10.1504/WRITR.2026.10079969>
- AlElaiwi, M., Ahmad, H., Hussain, J., Al-antari, M., Almarri, B., & Azhar, A. (2023). Saudi Arabia Public Roads Visual Pollution Dataset (Version 4) [Data set]. Mendeley Data. <https://doi.org/10.17632/bb7b8vtwry.4>
- AlElaiwi, M., Al-antari, M. A., Ahmad, H. F., Azhar, A., Almarri, B., & Hussain, J. (2023). Visual pollution real images benchmark dataset on the public roads. *Data in Brief*, 50, Article 109491. <https://doi.org/10.1016/j.dib.2023.109491>
- Almalki, H., & Algethami, N. (2025). Ensemble deep learning with image captioning for visual pollution detection, classification, and reporting. *Scientific Reports*, 15, Article 31516. <https://doi.org/10.1038/s41598-025-17200-0>
- Arya, D., Maeda, H., Ghosh, S. K., Toshniwal, D., & Sekimoto, Y. (2024). RDD2022: A multi-national image dataset for automatic road damage detection. *Geoscience Data Journal*, 11(4), 846–862. <https://doi.org/10.1002/gdj3.260>
- Chen, Z., Zhang, J., Lin, Y., Wu, R., Zhou, H., & Wang, W. (2026). Multi-adversarial metric-constrained domain generalization for road damage detection. *Image and Vision Computing*, 175, Article 106173. <https://doi.org/10.1016/j.imavis.2026.106173>
- Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3), 369–390. <https://doi.org/10.1111/j.1467-9868.2007.00593.x>

- General Authority for Statistics. (2025). Road transport statistics 2024. https://www.stats.gov.sa/documents/20117/2435281/Road%2BTranspot%2BStatistics%2B2024_EN.pdf/2c6ede72-2f23-7af1-50a3-150e493caddf
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>
- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 2980–2988). IEEE. <https://doi.org/10.1109/ICCV.2017.322>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1314–1324). IEEE. <https://doi.org/10.1109/ICCV.2019.00140>
- Huseljic, D., Herde, M., Hahn, P., Mücke, M., & Sick, B. (2025). Systematic evaluation of uncertainty calibration in pretrained object detectors. *International Journal of Computer Vision*, 133(3), 1033–1047. <https://doi.org/10.1007/s11263-024-02219-z>
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), Article 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- Kassem-Sbeyti, M., Karg, M., Wirth, C., Klein, N., & Albayrak, S. (2024). Cost-sensitive uncertainty-based failure recognition for object detection. In N. Kiyavash & J. M. Mooij (Eds.), *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence* (Vol. 244, pp. 1890–1900). PMLR. <https://proceedings.mlr.press/v244/kassem-sbeyti24a.html>
- Li, J., Qu, Z., & Yin, X. (2026). Domain-invariant feature enhancement domain adaptation for cross-scene road damage detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(3), 3466–3480. <https://doi.org/10.1109/TCSVT.2025.3613484>

- Lin, C., Tian, D., Duan, X., Zhou, J., Zhao, D., & Cao, D. (2023). DA-RDD: Toward domain adaptive road damage detection across different countries. *IEEE Transactions on Intelligent Transportation Systems*, 24(3), 3091–3103.
<https://doi.org/10.1109/TITS.2022.3221067>
- Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 936–944). IEEE.
<https://doi.org/10.1109/CVPR.2017.106>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (Lecture Notes in Computer Science, Vol. 8693, pp. 740–755). Springer.
https://doi.org/10.1007/978-3-319-10602-1_48
- Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45, 503–528.
<https://doi.org/10.1007/BF01589116>
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bkg6RiCqY7>
- Montreuil, Y., Yeo, S. H., Carlier, A., Ng, L. X., & Ooi, W. T. (2025). A two-stage learning-to-defer approach for multi-task learning. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, & J. Zhu (Eds.), *Proceedings of the 42nd International Conference on Machine Learning* (Vol. 267, pp. 44726–44749). PMLR.
<https://proceedings.mlr.press/v267/montreuil25b.html>
- Nakamura, Y., Ishii, Y., & Yamashita, T. (2024). Active domain adaptation with false negative prediction for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 28782–28792). IEEE.
<https://doi.org/10.1109/CVPR52733.2024.02719>

- Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164), 1–20. <https://jmlr.org/papers/v22/20-303.html>
- Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- Saudi Press Agency. (2026, February 12). RGA deploys new technologies in distinguished safe roads campaign. <https://www.spa.gov.sa/en/N2511008>
- Traub, J., Bungert, T. J., Lüth, C. T., Baumgartner, M., Maier-Hein, K. H., Maier-Hein, L., & Jäger, P. F. (2024). Overcoming common flaws in the evaluation of selective classification systems. *Advances in Neural Information Processing Systems*, 37, 2323–2347. <https://doi.org/10.52202/079017-0076>
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- Zimmermann, J. P., Holzbach, G., & Lerch, D. (2026). Knowledge-guided failure prediction: Detecting when object detectors miss safety-critical objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 4312–4321). https://openaccess.thecvf.com/content/CVPR2026W/SAIAD/html/Zimmermann_Knowledge-Guided_Failure_Prediction_Detecting_When_Object_Detectors_Miss_Safety-Critical_Objects_CVPRW_2026_paper.html

Supplementary materials

Table S1. Outer Saudi roles and held-fold fitting counts.

Role or held fold	Source rows	Unique images	Reports	Protected groups
Original training	19,024	18,280	16,745	16,704
Source validation	4,740	4,535	4,159	4,145
Unused calibration	3,253	3,116	2,838	2,833
Locked test	4,778	4,593	4,179	4,171
Fold 0 detector fitting	12,711	12,214	11,157	11,129
Fold 0 monitor fitting	5,015	4,810	4,435	4,425
Fold 0 monitor tuning	1,298	1,256	1,153	1,150
Fold 1 detector fitting	12,751	12,262	11,256	11,233
Fold 1 monitor fitting	5,014	4,823	4,393	4,381
Fold 1 monitor tuning	1,259	1,195	1,096	1,090
Fold 2 detector fitting	12,586	12,084	11,077	11,046
Fold 2 monitor fitting	5,133	4,941	4,525	4,517
Fold 2 monitor tuning	1,305	1,255	1,143	1,141

** Each held-fold block partitions the original training data for one rotation; blocks overlap and must not be added. External reserve retained 7,712 labelled rows and was unused.*

Table S2. Complete frozen policy inventory.

Policy	Priority or reference	Information and target
M00	No review	Unresolved source-reference complete misses.
M01	Uniform-random expected recovery	Analytical reference at the same integer review capacity.
M02	1 – maximum emitted confidence	Default priority 1 when no box is emitted.
M03	Inverse image-quality proxy	$1/(1 + \text{mean own-view Laplacian variance})$.
M04	Maximum centre-crop presence	Source-presence probability across own views.
M05	Maximum letterbox presence	Source-presence probability across own views.
M06	No-box-first centre-crop control	$2Z + \text{maximum presence} - \text{maximum emitted confidence}$.
M07	No-box-first letterbox control	Same ordering as M06 with letterbox presence.
M08	Detector-summary logistic head	8 report features; complete-miss target.
M09	Centre-crop pooled fusion	11 report features; complete-miss target.
M10	Letterbox pooled fusion	11 report features; complete-miss target.
M11	Emission-stratum prior	Beta(1, 1) smoothed complete-miss frequency.
M12	Separate emission-stratum heads	Centre-crop fusion features; separate complete-miss heads, shared regularization.
M13	Fine-tuned presence ordering	No-box-first formula with fine-tuned letterbox presence.
M14	Fine-tuned presence fusion	11 report features; complete-miss target.
M15	Maximum regional alarm	Region missed-object target; report maximum.
M16	Count-only logistic head	4 report count features; complete-miss target.
M17	Regional-bag logistic head	516 features; complete-miss target.
M18	Regional-plus-detector head	522 features; complete-miss target.
M19	Primary candidate	$2Z + \text{M18 probability}$; ordinal no-box-first score.
M20	Perfect oracle allocation	Unattainable reference: recover $\min(K, \text{complete misses})$.

* *A maximum over emitted scores is zero when none are emitted.*

Table S3. Representation, normalization, and sparse-supervision rules.

Element	Exact rule
Four count features	$\log(1 + \text{unique views})$; $\log(1 + \text{total proposals})$; mean $\log(1 + \text{proposals per view})$; fraction of views with no proposals.
Six additional detector features	No-box flag; maximum, mean, and population standard deviation of per-view maximum emitted confidence; $\log(1 + \text{total emitted boxes})$; fraction of views without emitted boxes.
Eight-feature detector summary	Six detector features above, plus $\log(1 + \text{unique views})$ and mean $\log(1 + \text{own-view Laplacian variance})$. Presence fusion adds maximum, mean, and population standard deviation of own-view presence.
Presence pixel normalization	RGB values scaled to $[0, 1]$, then channel means (0.485, 0.456, 0.406) subtracted and divided by standard deviations (0.229, 0.224, 0.225).
Report and presence normalization	Fitting-only feature means and population standard deviations; values below 0.00000001 replaced by 1. No tuning or external normalization update.
Regional normalization	Fitting-only means and population standard deviations; values below 0.000001 replaced by 1; moments accumulated in float64, region batches in float32.
Logistic optimization	L-BFGS uses strong-Wolfe line search, at most 300 iterations, gradient tolerance 0.00000001, and change tolerance 0.000000000001. A nonfinite fit or maximum absolute final gradient above 0.00001 stops the job.
Global one-class logistic fitting	Explicit intercept-only Beta(1, 1) prior, $p = (a + 1)/(n + 2)$, for a positive targets among n fitting units. No claim of learned feature coefficients. The same prior applies to undersupported stratum branches.
Separate stratum heads	A branch requires at least 20 fitting reports and 5 observations of each class. Otherwise its declared Beta(1, 1) prior is used and flagged.
Regional-alarm supervision gate	At least 20 positive and 20 negative fitting regions, 5 of each tuning class, 10 positive fitting images, and 5 positive tuning images; otherwise a flagged, unfitted zero alarm is retained.
Perceptual-screen decoding	EXIF transpose and draft decoding toward 128 pixels were used only for duplicate screening. A 32×32 LANCZOS grayscale DCT hash with its DC bit zeroed and a 9×8 difference hash supplied candidate distances. Training/evaluation preserved stored-raster coordinates.

Table S4. All Saudi comparison policies and failure-score diagnostics

ID	Description	Risk (%)	AUROC	AP	Brier	Log loss
M00	No review	36.78	—	—	—	—
M01	Random expectation	33.11	—	—	—	—
M02	Low detector confidence	30.83	0.6963	0.3582	—	—
M03	Low quality proxy	32.43	0.5219	0.2670	—	—
M04	Raw center presence	34.27	0.5401	0.2433	—	—
M05	Raw letterbox presence	34.47	0.5436	0.2463	—	—
M06	Center no-box-first	25.04	0.7734	0.6130	—	—
M07	Letterbox no-box-first	25.01	0.7696	0.6150	—	—
M08	Detector summary	30.39	0.7045	0.3811	0.1674	0.5003
M09	Center pooled	25.06	0.8191	0.6432	0.1346	0.4260
M10	Letterbox pooled	25.09	0.8126	0.6466	0.1362	0.4305
M11	Emission prior	30.83	0.6432	0.3255	0.1745	0.5281
M12	Separate failure heads	24.99	0.8234	0.6449	0.1337	0.4214
M13	Fine presence no-box-first	24.33	0.7855	0.6495	—	—
M14	Fine presence pooled	24.38	0.8394	0.6917	0.1253	0.4041
M15	Regional max alarm	31.91	0.6218	0.3121	—	—
M16	Counts only	30.06	0.6746	0.3868	0.1735	0.5255
M17	Regional bag	27.57	0.7633	0.5147	0.1543	0.4749
M18	Regional plus detector	24.75	0.8209	0.6566	0.1323	0.4205
M19	Candidate no-box-first	24.63	0.7788	0.6268	—	—
M20	Oracle ceiling	21.90	—	—	—	—

** All values are means across the 9 primary configurations. Risk is measured at 10% review; AUROC and AP evaluate complete-failure discrimination across all reports. Brier and log loss apply only to probability-valued policies; a dash means not applicable. No post hoc policy selection replaces M19. Probability-valued outputs were not independently calibrated.*



Figure S1. The complete deployable policy family across acquisition sources. Cells show mean unresolved complete-miss risk (%) at the fixed 10% budget for all 18 deployable policies. The shared color scale encodes risk rather than a difference or statistical significance.

** Saudi units are reports; external units are images. Darker cells indicate higher source-reference risk. Source-specific prevalences and target references differ, so columns do not form a pooled national comparison.*

Table S5. Configuration-level Saudi outcomes at the primary operating points

Seed	Fold	Object F1	Silent misses	Emitted misses	M19 risk (%)	Δ vs M06 (pp)	Δ vs M07 (pp)
20261020	0	0.5850	692	361	25.27	−0.678	−0.464
20261029	0	0.5889	718	333	25.05	−0.321	−0.571
20261030	0	0.5921	680	344	24.95	+0.214	+0.286
20261020	1	0.5925	654	356	24.13	−0.571	−0.428
20261029	1	0.5930	684	355	24.63	−0.678	−0.571
20261030	1	0.5900	682	341	24.02	−0.678	−0.642
20261020	2	0.5918	708	333	24.98	−0.250	−0.357
20261029	2	0.5859	618	361	23.30	−0.535	−0.214
20261030	2	0.5857	721	334	25.30	−0.214	−0.500

** Miss counts precede review. Seeds and held folds are crossed; folds share fitting data and are not independent repetitions. Confidence was 0.40 except seed 20261029, fold 2, for which the validation-selected value was 0.35. Negative differences favor M19.*

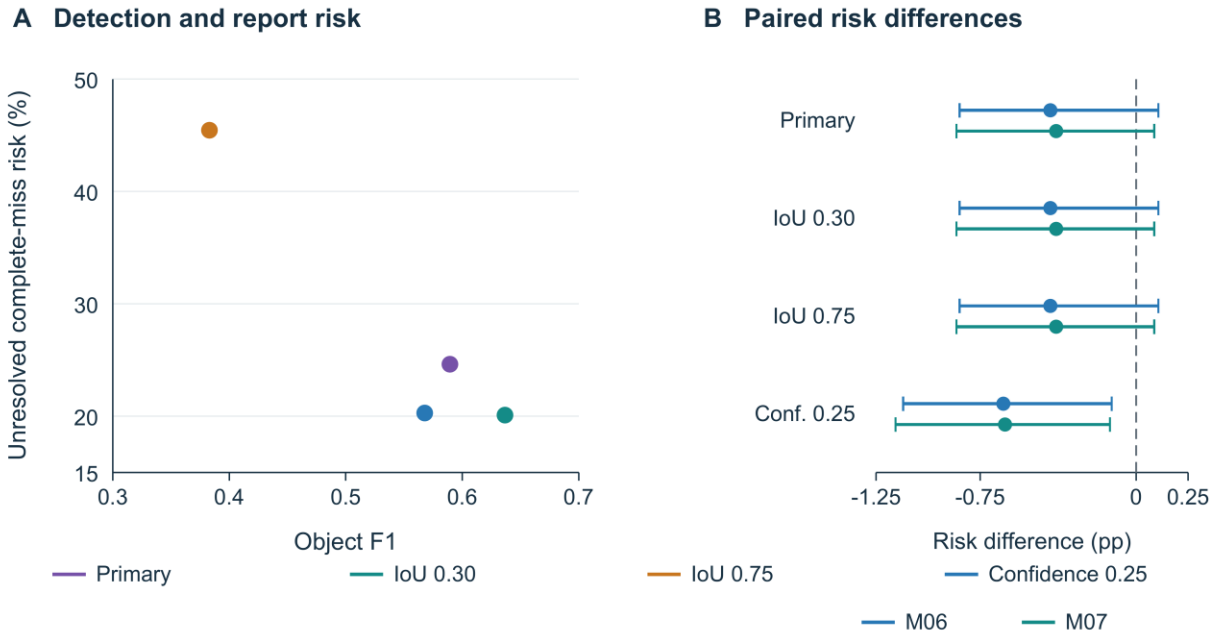


Figure S2. Sensitivity of detection quality, residual risk, and paired contrasts. (A) Mean object F1 and M19 risk. (B) M19-minus-control differences and 97.5% paired intervals. All conditions use 10% review.

** Heads remain fixed. IoU contrasts are structurally dependent. Confidence 0.25 is secondary. pp, percentage points.*

Table S6. Exact Saudi sensitivity estimates and adjusted intervals

Condition	Object F1	M19 risk (%)	Δ vs M06 pp [97.5% CI]	Δ vs M07 pp [97.5% CI]
Primary IoU 0.50	0.5894	24.63	-0.412 [-0.849, +0.107]	-0.385 [-0.864, +0.087]
IoU 0.30	0.6367	20.10	-0.412 [-0.849, +0.107]	-0.385 [-0.864, +0.087]
IoU 0.75	0.3830	45.45	-0.412 [-0.849, +0.107]	-0.385 [-0.864, +0.087]
Confidence 0.25	0.5678	20.28	-0.638 [-1.120, -0.117]	-0.631 [-1.157, -0.126]

** IoU sensitivities retain primary emitted boxes. Confidence 0.25 retains IoU 0.50. Denominators are unchanged; values are rounded for presentation.*

Table S7. Exact source-specific external contrasts

Source	Δ vs M06 pp [97.5% CI]	Δ vs M07 pp [97.5% CI]
China drone	+7.494 [+1.389, +16.049]	+4.393 [0.000, +11.382]
China motorbike	0.000 [-0.231, 0.000]	+0.487 [-0.160, +1.852]
Czechia	+5.229 [-2.586, +12.669]	+6.629 [-2.448, +15.168]
India	-1.494 [-3.397, +0.579]	-1.338 [-3.115, +0.684]
Japan	-2.130 [-3.974, -0.173]	-2.212 [-4.017, -0.392]
Norway	+2.104 [-3.302, +7.105]	+3.398 [-2.569, +8.974]
United States	-1.170 [-8.568, +7.252]	+3.743 [-3.969, +12.461]

* At 10% review; negative differences favor M19. Adjustment is within source only. The China-drone M07 lower bound is numerically zero.



Figure S3. Source-negative examples with and without emitted detections. The fixed model emitted 2 false-positive boxes in (E) and no boxes in (F), relative to the supplied source annotations. Panel F had a maximum raw detector score of 0.399968, below the 0.40 emission threshold; no rounded score was used for thresholding.

* Panel letters continue Figure 8.

Table S8. Per-image evidence for the six illustrated outcomes

Panel	Fine presence	Max emitted score	TP/FP/FN	Precision	Recall	Matched IoU
A	0.938	0.853	1/0/0	1.00	1.00	0.886
B	0.848	0.916	1/1/1	0.50	0.50	0.751
C	0.701	—	0/0/1	—	0.00	—
D	0.147	0.505	0/1/2	0.00	0.00	—
E	0.975	0.531	0/2/0	0.00	—	—
F	0.911	—	0/0/0	—	—	—

** Metrics use one-to-one matching at IoU 0.50 and confidence ≥ 0.40 . Precision = $\frac{TP}{TP + FP}$; recall = $\frac{TP}{TP + FN}$. A dash denotes an undefined denominator, no emitted box, or no match, as appropriate. Each matched panel contains exactly 1 true-positive box.*

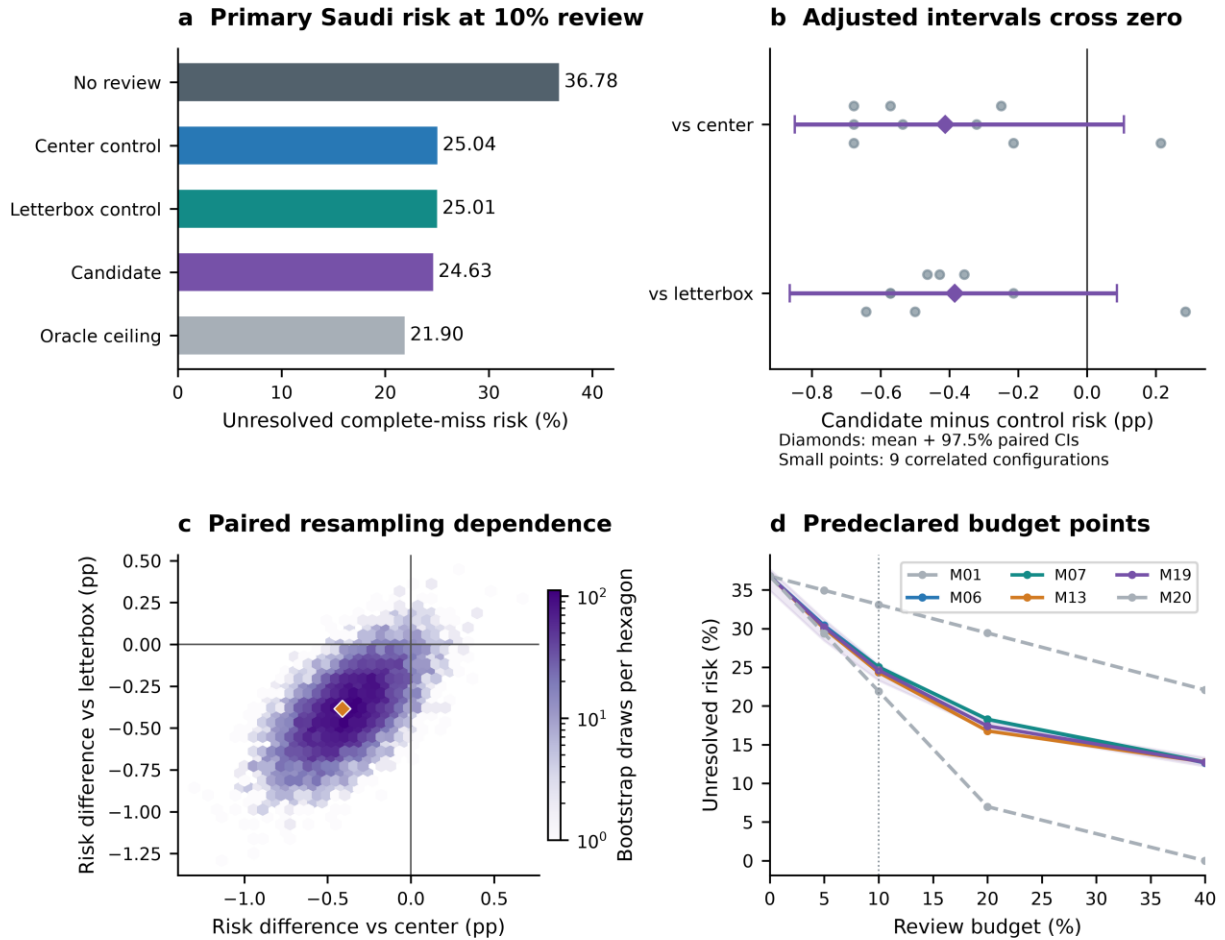


Figure S4. Saudi primary cohort: 4,179 reports, 2,802 source positives, 417 reviews, three optimization seeds crossed with three held folds. (a) Mean risk. (b) Candidate-minus-control effects with two Bonferroni-adjusted 97.5% paired protected-group bootstrap intervals; negative favors M19. Small points are correlated configurations, not independent replications. (c) Frequency of the same 10,000 paired draws, not a posterior or probability of superiority. (d) Means at five fixed budgets; shaded candidate range is the minimum–maximum across configurations, not a confidence band. Lines join observed points only.

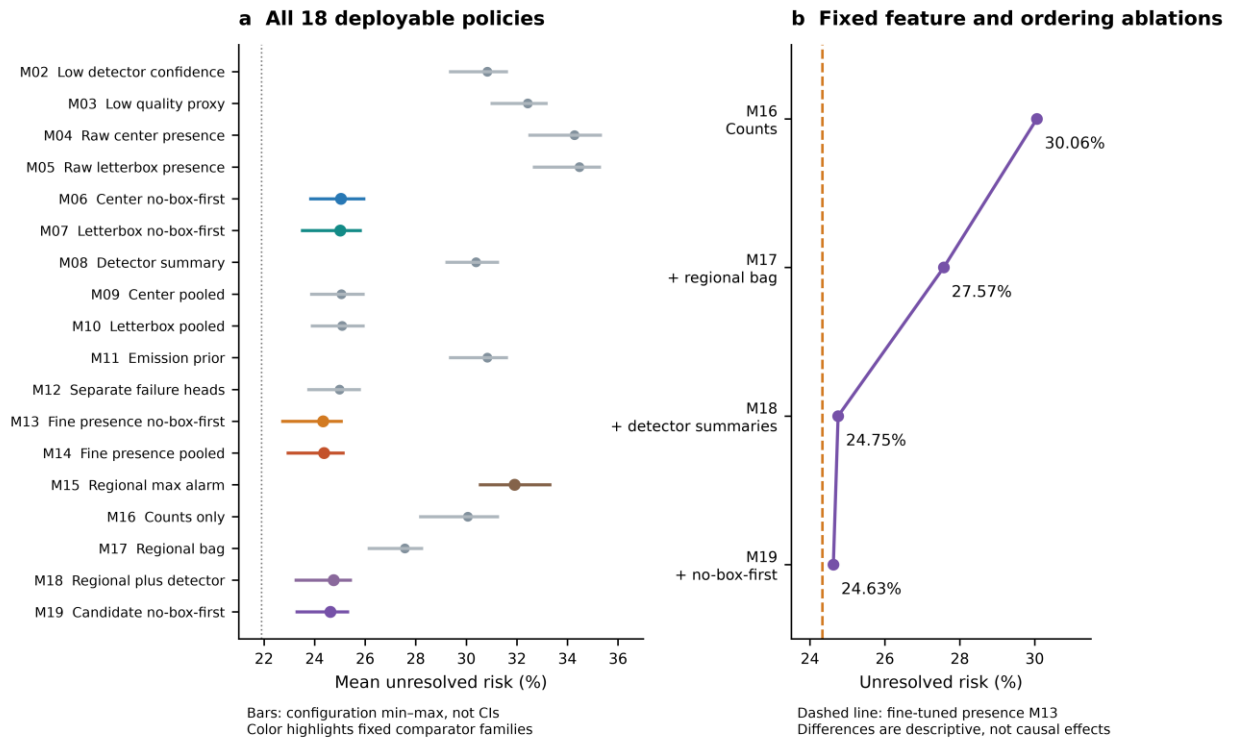


Figure S5. Saudi primary risk at the 10% report budget. Panel (a) retains the entire deployable family in identifier order. Horizontal bars describe the range across nine correlated configurations, not uncertainty intervals. Panel (b) follows the prespecified counts, regional-bag, detector-summary and no-box-first sequence; M13 is the fine-tuned presence reference. Different feature dimension, information and supervision prevent interpreting these differences as isolated causal effects. No best policy replaces the frozen M19 hypothesis.