Continuous Extrinsic-Miscalibration Monitoring Above a Fleet-Visible EyeCAN Flag Baseline

A Mixed-Regime Study of NADIR Pulse, Meridian, Apex, and Echo with Noisy Cross-Modal Residuals and Bounded Software Compensation

Dhruv Hegde[1,*] Srivatsan Balaji[1]

[1]NADIR Technologies, Ann Arbor, Michigan, USA

*Correspondence: `founders@nadirai.net`

July 2026 Study ID `ME-LDW-OVERLAY-2026-07`

Abstract

Forward-facing advanced driver-assistance cameras require a stable relationship between the camera coordinate frame, vehicle frame, and any radar or lidar frames used by downstream fusion. A shop calibration establishes that relationship at one time, while vibration, repair variation, mount disturbance, glass replacement, temperature cycling, and mechanical aging create a continuing monitoring problem. This paper studies whether a fleet software overlay can identify emerging extrinsic yaw disagreement before or in the absence of a useful fleet-visible binary warning. The comparison is intentionally narrow. The baseline is an observable EyeCAN-style flag monitor composed of lane-departure, tamper, maintenance, and vehicle-state fields. It is not a reconstruction of Mobileye's proprietary internal online-calibration logic, and the study does not claim that Mobileye lacks internal safeguards or correction mechanisms.

We evaluate four NADIR service tiers using a detector-independent mixture of micro gradual drift, moderate gradual drift, severe events, and harsh noisy episodes. Unlike an earlier stress-only fixture, the generator does not cap every drift below a chosen flag threshold. Flag probability rises smoothly with yaw, and the overlay residual channel includes heteroscedastic noise, persistent bias, missing observations, isolated outliers, weather effects, vibration, and variable transport cadence. The primary study contains 800 drifted episodes, 300 clean episodes, and 185,664 simulated frames, followed by factor sweeps. Pulse uses robust instantaneous residual evidence and persistence. Meridian adds motion-gated cumulative change detection. Apex adds a lightweight change-posterior confirmation path. Echo performs lower-cadence batch confirmation. A bounded advisory transform is evaluated against latent yaw so that estimation error can reduce or worsen correction quality.

Across the mixed drift set, the EyeCAN-Flag baseline reached 48.9% recall. Pulse reached 82.3%, Meridian 89.3%, Apex 89.5%, and Echo 87.3%. The largest separation appeared in micro gradual drift, where recall was 18.5% for the baseline, 57.0% for Pulse, 77.0% for Meridian, 78.0% for Apex, and 55.5% for Echo. In severe events the baseline reached 90.5%, while every NADIR arm reached 100%, showing that the baseline remains competitive when faults become large and explicit. Clean-episode false-alert rates were 8.3% for the probabilistic flag fixture, 0% for Pulse and Echo, and 7.0% for Meridian and Apex. These values characterize the simulation, not a field estimate of Mobileye false-warning behavior. Median algorithm compute remained below 0.17 ms per scored frame, but end-to-end alert time is governed mainly by ingest cadence. Bounded software compensation reduced

latent yaw in 92.8% to 98.6% of detected episodes, depending on tier. The evidence supports a complementary overlay hypothesis, not a general superiority claim. Field traces with synchronized calibration ground truth remain necessary before safety, OEM, or production-effectiveness claims are justified.

Keywords: ADAS calibration; camera yaw drift; EyeCAN; Mobileye; online extrinsic monitoring; robust residuals; CUSUM; change-point detection; software compensation; fleet telemetry

1 Introduction

Modern ADAS performance depends on geometric assumptions that are easy to state and difficult to preserve. A forward-facing camera is mounted at an estimated pose relative to the vehicle body. Radar and lidar observations are transformed into the same reference frame before association or fusion. Lane models, object bearings, time-to-collision calculations, and evidence visualizations inherit those transformations. A small rotational error can therefore create a range-dependent lateral discrepancy even when each sensor remains individually healthy. A one-time calibration result is important, but it does not answer whether the pose remains trustworthy after the vehicle leaves the bay.

The operational problem is not simply fault detection. Fleet operators need a sequence of decisions with different costs. A sensitive monitor can find weak drift but create unnecessary inspections. A conservative binary flag can preserve driver-warning specificity but reveal little about gradual geometric disagreement. A correction estimate can reduce residual error in a shadow transform, but an incorrect estimate can increase it. A system that reports only a final alert hides the evidence needed to decide whether a vehicle should remain in service, be scheduled for inspection, or be escalated to a calibration facility.

NADIR is designed as a software overlay for that intermediate monitoring layer. It consumes an explicit residual contract from available sensors and telematics, computes tiered evidence, and emits advisory fleet outputs. It does not replace the perception system, alter the Mobileye ECU, suppress driver warnings, or certify calibration. The distinction matters because exported flags and internal algorithms are different objects. A fleet integration can test what it can observe through documented or contracted outputs. It cannot infer the performance of proprietary logic that is not exposed.

This paper addresses one question: when the fleet-visible comparator is an EyeCAN-style binary flag monitor, how much additional detection coverage can a noisy continuous residual channel provide, and what costs accompany that coverage? The study is structured to avoid an answer that is predetermined by the generator. Some episodes remain small. Some cross a probabilistic flag transition. Some become severe enough that the binary baseline should perform well. Harsh episodes degrade both residual quality and flag reliability. Clean episodes contain the same forms of nuisance variation so that false alerts are not trivially zero.

The resulting conclusion is more restrained and more useful than the earlier stress-only result. NADIR is not uniformly faster. When an explicit flag fires, its parsing path is computationally simpler and its transport cadence can be much shorter. NADIR is not uniformly more sensitive under every data contract. Pulse and Echo miss nearly half of micro episodes in the current fixture, while Meridian and Apex buy additional recall at a 7.0% clean-episode false-alert rate. The practical advantage is concentrated in gradual and moderate disagreement where repeated residual evidence exists but a binary warning may never become persistent. That is an overlay-shaped advantage rather than a claim of complete-stack superiority.

The contributions are fivefold. First, we define a comparison boundary that separates public or contracted outputs from proprietary internal behavior. Second, we provide a mixed-regime generator with 42 operational factors and explicit nuisance processes. Third, we evaluate all four NADIR tiers on the same 1,100 primary episodes and publish regime-level recall, false alerts, precision, balanced accuracy, detection survival, compute timing, and correction quality. Fourth, we state formal properties for robust influence, conformal ranking, persistent accumulation, and bounded correction. Fifth, we provide code, SQLite results, tables, and programmatic 3D geometry so that the reported numbers can be reproduced and challenged.

2 Problem Review and Literature

2.1 Post-repair and lifetime calibration

Extrinsic calibration estimates the rigid transformation that maps one sensor coordinate frame into another or into the vehicle frame. The need is independent of whether the downstream system is a research platform, an OEM stack, or an aftermarket ADAS installation. Public Mobileye installation materials describe target acquisition calibration and a recalibration workflow for Mobileye 8 series units after changes such as windshield replacement [Mob, 2022, Mobileye, 2026]. SAE's aftermarket review identifies unresolved variation in repair-triggered procedures, validation, documentation, and liability [Waraniak, 2021]. These procedures establish an important service baseline, but a completed procedure is not itself a continuous estimate of pose quality.

The broader literature treats online calibration as a long-term reliability problem because vibration, environmental exposure, collision, maintenance, and mount variation can change the relationship between sensors. The targetless-calibration survey by Li et al. [2023] and learning-focused review by Tan et al. [2024] organize information, feature, motion, and learned approaches while emphasizing the difficulty of matching modalities with different resolution, field of view, and physical measurement principles. Borer et al. [2023] and Wei et al. [2024] focus specifically on continuous self-checking and recalibration, arguing that lifetime disturbances motivate updates from ordinary sensor data rather than periodic targets alone. NHTSA-sponsored degradation experiments further show that sensor degradation can produce outcomes ranging from feature shutdown to degraded behavior or no visible system impact, depending on system redundancy and fusion [Stowe et al., 2025].

This paper does not treat software monitoring as a substitute for OEM procedures. The intended relationship is sequential. Shop or OEM calibration supplies an accepted reference. A shadow monitor estimates whether operating data remain consistent with that reference. A governed workflow then decides whether the evidence is weak, requires more observations, justifies an advisory transform for a non-safety visualization, or requires physical recalibration. The functional-safety boundary follows the principle that monitoring evidence, intended-functionality analysis, standardized health interfaces, and safety-certified actuation are separate claims [International Organization for Standardization, 2018, 2022, 2026, SAE International, 2021].

2.2 Target-based, targetless, and online calibration

Target-based methods remain attractive because known geometry creates strong correspondences and traceable setup conditions. Their weakness for fleet monitoring is operational: targets, controlled space, trained labor, and service downtime are not continuously available. Targetless methods replace the

known target with natural-scene features, ego-motion consistency, mutual information, object tracks, or learned cross-modal representations.

Pandey et al. [2012] demonstrated targetless lidar-camera calibration by maximizing mutual information between sensor-measured surface intensities. The method provided an important proof that in-field calibration could be formulated without a dedicated target, while also exposing the dependence of information-based objectives on scene content. Levinson and Thrun [2013] addressed automatic online camera-laser calibration and correction of gradual drift, while de Paula et al. [2014] used road geometry and vehicle motion for onboard camera calibration. Schneider et al. [2017] introduced RegNet, which learned six-degree-of-freedom registration between lidar and camera and supported iterative online correction. Iyer et al. [2018] proposed CalibNet, using geometric and photometric consistency to supervise a 3D spatial-transformer network without directly regressing to ground-truth calibration parameters during training.

These approaches estimate a transformation. NADIR's primary role in this paper is narrower: monitor a residual stream, identify evidence of change, and govern an advisory correction. That distinction is important because a detector can operate with a lower-dimensional residual contract when the upstream platform cannot expose raw images or point clouds. It is also limiting because the detector cannot recover information that the contract does not contain. A scalar or low-dimensional residual is cheaper to ingest, but less identifiable than synchronized raw multimodal observations.

2.3 Radar-camera and multi-sensor evidence

Radar-camera calibration is particularly relevant because production vehicles often provide complementary radar range and velocity with camera semantic detail. Cheng and Cao [2023] proposed online targetless radar-camera calibration using learned common features, RANSAC, and nonlinear optimization. Their work illustrates both the value and the prerequisite of cross-modal correspondence: a useful calibration estimate requires objects or features that can be associated across sensors.

Multi-sensor methods broaden that concept. Peršić et al. [2021] used moving-object tracks from radar, lidar, and camera for lightweight decalibration detection and graph-based rotation calibration, reporting real-world evidence that existing tracking outputs can support online monitoring. Chen et al. [2021] used alignment-sensitive image-lidar embeddings to monitor cross-modal data faults, while the survey by Goelles et al. [2020] places calibration and attachment faults within the larger automotive fault-detection problem. This is close to the residual-contract idea used here. NADIR does not assume one universal residual. It accepts lane-geometry disagreement, associated-object bearing disagreement, projected depth error, track consistency, or an upstream calibration estimate, provided the contract carries quality, timing, and context.

The KITTI-360 dataset provides rich multimodal urban data and accurate localization for research at the intersection of vision, graphics, and robotics [Liao et al., 2023]. It is useful for later trace replay, but the present paper does not claim a KITTI-360 field validation. The primary results are generated from a protocol-oriented simulator so that yaw, onset, nuisance variables, and counterfactual correction error remain observable. This improves experimental control while reducing external validity. Both facts must be stated together.

2.4 Robust and sequential monitoring

Continuous monitoring converts geometric residuals into a statistical decision. Mahalanobis distance supplies covariance-aware multivariate scoring [Mahalanobis, 1936]. Huber's bounded-influence estimator reduces the impact of large isolated residuals [Huber, 1964]. Conformal prediction converts calibration scores into finite-sample ranks and can control marginal error under exchangeability without specifying a parametric score distribution [Vovk et al., 2005, Angelopoulos and Bates, 2023]. The guarantee is conditional on the validity of the calibration group and does not survive arbitrary seasonal, hardware, or policy shift without refresh.

Persistent drift requires more than an instantaneous score. Page's cumulative-sum procedure accumulates evidence for a sustained distributional change [Page, 1954]; later treatments formalize abrupt-change detection and operating-characteristic tradeoffs [Basseville and Nikiforov, 1993, Hinkley, 1971]. Bayesian online change-point detection maintains a posterior over run length and makes assumptions about hazard and predictive distributions explicit [Adams and MacKay, 2007]. These families motivate the Pulse, Meridian, and Apex ladder. Echo addresses a different constraint by reducing scoring cadence for batch fleet analytics.

2.5 Mobileye comparison boundary

Mobileye's public materials establish a production ADAS ecosystem, supported installation and calibration procedures, and fleet-facing manuals [Mobileye, 2026, Mob, 2022]. The public EyeCAN-like integration used by this repository exposes warning and vehicle-state fields that are suitable for an observable-output baseline. Public documentation does not disclose enough information to reconstruct every internal EyeQ calibration, plausibility, degradation, or fail-safe process. It would therefore be scientifically incorrect to call the simulated baseline "Mobileye detection" without qualification.

We use the term *EyeCAN-Flag baseline* for the implemented comparator. It decides from exported LDW, tamper, maintenance, fail-safe, blinker, speed, and streak fields. The simulator's flag-transition midpoint is a sensitivity variable, not a measured Mobileye threshold. The study can support a statement about an overlay relative to that observable flag abstraction. It cannot support claims that NADIR is better than Mobileye perception, object detection, driver HMI, internal calibration, or a specific production vehicle.

2.6 Gap addressed by this study

Prior NADIR experiments emphasized distribution-free false-alert control under sparse telemetry and found that static thresholds could outperform Pulse on short, strongly injected drift [Hegde and Balaji, 2026]. The stress-only predecessor to this paper then moved too far in the opposite direction: it capped yaw below an assumed flag visibility threshold, gave the overlay a nearly direct continuous view of yaw, and produced a 100% versus 2% headline. That fixture demonstrated a constructed blind zone, but it was not a balanced estimate of expected operation.

The current study closes that methodological gap. Endpoints are sampled by operating regime, not by detector response. Binary flags become more probable as yaw increases but can also fire from nuisances. Continuous residuals become informative as yaw increases but can be missing, biased, or corrupted. Severe episodes test the region where a binary baseline should recover. Clean episodes include weather and vibration. The study still uses synthetic data, but its central performance difference is no longer hard-coded by a universal yaw cap.

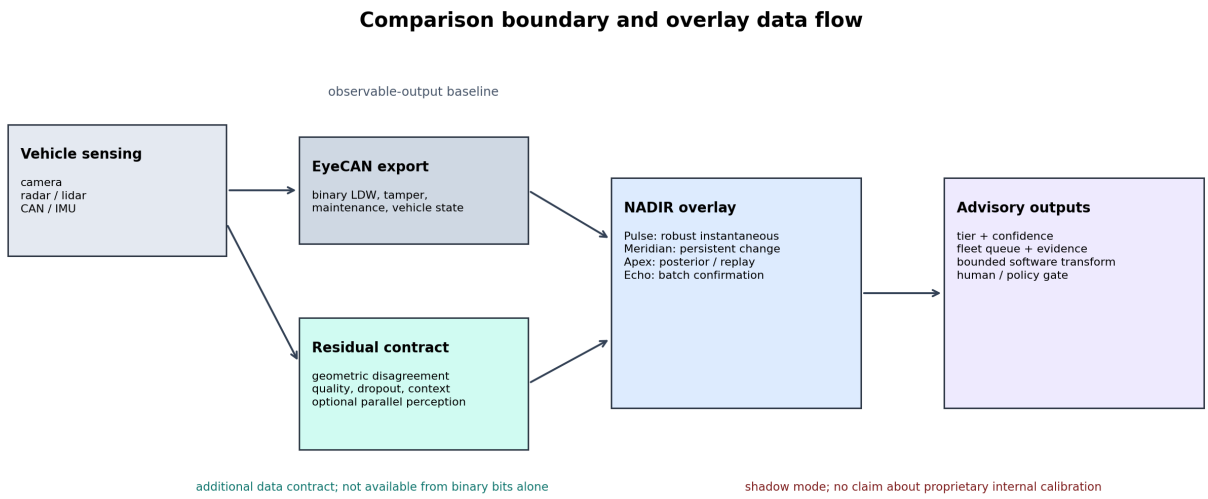

Figure 1: Comparison boundary and overlay data flow. The EyeCAN-Flag baseline uses exported binary warnings and vehicle state. NADIR also requires a continuous residual contract. The output is advisory and remains outside the Mobileye ECU and safety actuation path.

3 NADIR System and Tier Model

3.1 System role

NADIR is a fleet monitoring and evidence layer positioned outside the safety-critical perception and actuation path. A gateway, telematics service, or parallel-perception module supplies time-stamped observations. The edge scorer normalizes units, evaluates quality gates, produces a residual vector, and assigns a tier. The fleet service aggregates repeated events, signs or records evidence, and presents a human or policy-controlled action. No step in this study writes calibration parameters to EyeQ.

Figure 1 separates the two data paths. The upper path contains the binary fields available to the baseline. The lower path is an additional residual contract. The comparison therefore measures the combined value of richer data and tier algorithms. It is not an algorithm-only contest with identical features. That asymmetry is intentional because the proposed product is an overlay with an expanded data contract, but it must be disclosed whenever results are interpreted.

3.2 Residual contract

The residual contract is designed to represent heterogeneous deployments without pretending that all vehicles expose identical data. Each observation includes a geometric residual, timestamp, sensor-quality indicators, vehicle motion, weather or visibility context, source identifiers, and missingness. The reference implementation derives camera, radar, and lidar disagreement features and can accept an upstream yaw estimate. A real deployment must document how each feature is produced, synchronized, calibrated, retained, and quality-scored.

Table 1: Minimum residual contract assumed by the overlay. Optional modalities improve observability but are not silently imputed when absent.

Field group	Required content	Operational purpose
Identity and time	vehicle, sensor pair, source clock, event time	Prevents cross-vehicle mixing and supports replay, ordering, and clock-quality checks.
Geometric evidence	signed or vector residual with units and frame definition	Supplies the quantity monitored by Pulse and accumulated by higher tiers.
Observation quality	confidence, visibility, dropout, outlier or saturation markers	Separates lack of evidence from evidence of alignment and controls gating.
Vehicle context	speed, steering, yaw rate, blinker, wiper, road regime	Prevents known maneuvers and poor observability from being interpreted as calibration change.
Calibration reference	accepted pose version, completion time, service event	Anchors drift evidence to a traceable starting point.
Optional raw references	image, track, point-cloud, or object-association locator	Enables forensic replay and independent recomputation without requiring raw data on every frame.

3.3 Pulse

Pulse is the lowest-latency continuous scorer. It forms a robust residual score from the available modal differences, compares the current observation with an onboarding calibration set, and requires a short persistence streak before elevation. In the current implementation, the first sixteen valid residuals form a frozen baseline. A robust median and median absolute deviation provide location and scale. An observation must exceed both a score margin and a signed-yaw magnitude gate, and three consecutive exceedances are required for CAUTION. Five exceedances with the stronger magnitude condition produce CRITICAL.

Pulse is intended for deployments that value low callback volume and immediate local evidence. It deliberately leaves some micro drift undetected. The formal bounded-influence property is given in Proposition 1; the calibration-rank property is stated in Theorem 1. The implementation in this study uses a robust envelope in the hot path and retains the conformal interpretation as a calibration contract. A production claim of exact conformal coverage would require stored calibration ranks for each declared stratum.

3.4 Meridian

Meridian adds a persistent-change channel to Pulse. It motion-gates observations, standardizes the residual against the frozen robust baseline, and accumulates positive evidence with a one-sided cumulative sum. The accumulator resets toward zero when evidence is weak. Elevation requires both a threshold crossing and repeated strong observations, preventing a single large outlier from defining the episode.

Meridian is designed for slow change that may rarely exceed Pulse's instantaneous threshold. The additional memory explains the primary performance tradeoff: micro-drift recall rises from 57.0% to 77.0%, while the clean-episode false-alert rate rises from 0% to 7.0% in this simulation. The deterministic accumulation result in Lemma 1 explains why sustained weak evidence can cross a boundary even when

Table 2: NADIR tiers are deployment policies around a shared residual contract. They are not four independent perception systems.

Tier	Primary evidence	Typical cadence	Intended decision
Pulse	Robust instantaneous score plus short persistence	Streaming or edge poll	Low-noise CAUTION or CRITICAL evidence for fleet triage.
Meridian	Pulse plus motion-gated cumulative change	Streaming or frequent poll	Detect gradual persistent drift that weakly exceeds baseline.
Apex	Meridian plus change-posterior confirmation and replay context	Edge plus cloud evidence	Forensic escalation, changepoint narrative, and evidence retention.
Echo	Subsampled Pulse plus robust batch trend	Depot, nightly, or low-bandwidth	Fleet rollup and economical confirmation where live dispatch is unnecessary.

no single sample is decisive.

3.5 Apex

Apex combines the Meridian path with a lightweight Bayesian change-posterior approximation. It locks an initial baseline, evaluates whether the residual is more consistent with a new mean, and requires three consecutive change-supported exceedances before independent elevation. If Meridian or Pulse is already elevated, Apex can preserve that state as forensic confirmation rather than restarting the decision.

Apex is positioned for investigations that need a richer change narrative, replay, and posterior evidence. The benchmark implementation is intentionally called Apex-lite. It does not implement a full production BOCPD run-length matrix, learned likelihood family, or distributed evidence store. Its 89.5% overall recall is therefore a property of this implementation and fixture, not a universal Apex specification.

3.6 Echo

Echo is a lower-cadence confirmation and depot-analytics tier. It samples the Pulse path at a reduced frequency and includes a robust trend fallback over a nine-observation median with persistence. Echo cannot guarantee the earliest alert because it intentionally discards temporal resolution. It can, however, reduce compute and transport load when a fleet needs nightly or depot-level trend review rather than live dispatch.

Echo reached the highest balanced accuracy in the primary fixture because it combined 87.3% recall with no clean-episode alerts. That result does not imply that batch scoring is intrinsically superior. The equal weighting of drifted and clean performance in balanced accuracy rewards specificity, and the simulator may not contain every nuisance pattern that would challenge Echo's long window.

4 Experimental Design

4.1 Research questions

The experiment addresses four pre-registered questions. First, how does each arm's episode recall change across micro, moderate, severe, and harsh regimes? Second, what clean-episode false-alert cost accompanies that recall? Third, when an arm alerts, how often does a bounded software transform reduce latent yaw rather than increase it? Fourth, are algorithm compute costs operationally meaningful relative to CAN and telematics transport cadence?

The primary endpoint is episode recall on drifted data. An episode is counted as detected when an arm produces at least one elevated state. Clean-episode false-alert rate is the corresponding probability of any elevation on drift-free episodes. We also report precision under the study's fixed drift-to-clean sample ratio, F1, specificity, balanced accuracy, detection survival after onset, detection below the flag-transition midpoint, and cases where a NADIR arm detects while the baseline never fires. Precision is not portable to a fleet with a different prevalence and must be recalculated for deployment.

4.2 Mixed operating regimes

The drifted set contains four equally sized regimes with 200 episodes each. Micro gradual episodes end between 0.12 and 0.38 degrees and use linear or exponential growth. Moderate gradual episodes end between 0.38 and 0.85 degrees and include intermittent profiles. Severe events end between 0.85 and 1.55 degrees and include step, shock-recovery, and linear profiles. Harsh noisy episodes span 0.20 to 1.15 degrees while increasing vibration, wiper duty, low-visibility events, residual noise, dropout, and outliers. The 300 clean episodes draw the same platform, motion, weather, and nuisance families without latent yaw drift.

This mixture avoids using one detector's threshold to define all endpoints. The flag midpoint varies from 0.45 to 0.95 degrees, and the transition width varies from 0.08 to 0.22 degrees. A severe episode can therefore trigger repeated binary warnings. A micro episode can remain below the midpoint but is not guaranteed to do so for every parameter combination. The continuous residual is similarly imperfect. Its standard deviation ranges from 0.025 to 0.08 degrees in ordinary episodes and from 0.06 to 0.14 degrees in harsh episodes. Missingness reaches 35%, outlier probability reaches 10%, and persistent bias is sampled around zero.

4.3 EyeCAN-Flag baseline

The baseline consumes warning state and vehicle context only. It elevates on tamper, maintenance, fail-safe, or a qualifying unintentional LDW streak after blinker suppression. The simulator generates nuisance LDW probability from vibration and wiper state, then adds a smooth yaw-dependent probability term. The logistic midpoint and width are explicit sensitivity factors. Tamper requires the yaw to cross a separate threshold and remain there for a sampled delay.

These choices create a general observable-flag model, not a calibrated replica of a specific Mobileye product. The yaw-driven term is normalized to zero at zero yaw, while low-rate nuisance flags remain possible. The resulting 8.3% clean-episode false-alert rate reflects the generator, the episode-level OR rule, and windows of as many as 240 frames. It must not be quoted as Mobileye's real-world false-warning rate.

Table 3: Primary data volume. Factor sweeps are additional to the 1,100 primary episodes.

Dataset quantity	Value
Drifted episodes	800
Clean episodes	300
Total episodes	1,100
Total simulated frames	185,664
Primary random seeds	42–52
Factor-sweep rows	95

4.4 Residual and sensor model

For each frame, latent yaw defines the geometric disturbance. The camera channel receives the signed yaw estimate with Gaussian noise, persistent episode bias, and occasional heavy outliers. Radar and lidar residual magnitudes are coupled to yaw with modality-specific scale and additional vibration terms. Wiper and low-visibility states inflate uncertainty. Dropout removes the residual observation rather than replacing it with a nominal value. Vehicle speed, steering, yaw rate, blinker, and fail-safe fields provide motion and observability context.

The model intentionally gives NADIR richer information than the binary baseline. That is the product hypothesis: a residual contract can extract more information than a few flags. It also means the results decompose into data value and algorithm value only through ablation. The current camera-only and camera-radar profiles are coarse approximations. A future study should hold the data contract fixed and compare robust thresholds, Pulse, Meridian, and Apex against classical alternatives on identical residuals.

4.5 Sampling, ingestion scale, and reproducibility

Seeds 42 through 52 generate the primary blocks. Episode lengths are 96, 168, or 240 frames, with onset between frames 24 and 48. The first valid observations before onset provide onboarding calibration. Edge poll choices span 5, 15, 30, 60, and 300 seconds; CAN frame periods span 66 to 100 milliseconds. These time parameters affect wall-clock interpretation but not the frame-indexed recall metric.

Every trial and arm outcome is written to SQLite with the random seed, regime, complete factor JSON, alert frame, latent yaw, proxy yaw, correction outcome, and measured compute time. The scorecard is exported to JSON. Tables and figures are regenerated from those artifacts rather than copied by hand.

4.6 Operational factors

The design includes 42 factors covering platform labels, transport, drift, environment, warnings, statistical thresholds, episode scale, residual integrity, and correction policy. The platform labels are stratification labels and do not claim product-specific internal behavior. The full table is placed here rather than after the references so that the paper ends with its bibliography.

ID	Factor	Studied range or levels
1	platform	EyeQ5 Surround, SuperVision L2++, EyeQ6H MQB, ME8 Connect Fleet
2	can frame period ms	66 to 100 ms
3	edge poll interval s	5, 15, 30, 60, 300
4	uplink sparsity	continuous, hourly, daily
5	true yaw drift rate deg per day	0.01 to 0.5 deg/day
6	onset profile	step, linear, exponential, intermittent, shock_recovery
7	post repair age days	0 to 180 day
8	ambient temp c	-20 to 45 C
9	vibration index	0.0 to 1.0 0-1
10	speed band mph	0-15, 15-55, 55-75+
11	ldw streak cap	1 to 12 count
12	blinker contamination rate	0.0 to 0.4 fraction
13	wiper duty	0.0 to 1.0 fraction
14	low visibility failsafe rate	0.0 to 0.15 fraction
15	tamper onset delay h	0 to 72 h
16	maintenance false pos per week	0.0 to 0.02 rate
17	fcw headway stress rate	0.0 to 0.1 fraction
18	conformal alpha	0.01, 0.05, 0.1
19	mondrian stratum	platform_x_speed_x_weather
20	huber c H	1.0 to 2.5 unitless
21	me binary require streak	1 to 5 count
22	cusum k	0.1 to 0.5 z
23	cusum h	3.0 to 8.0 z
24	motion gate v min mph	5 to 25 mph
25	fleet size N	50, 200, 1000
26	episode length frames	96, 168, 240
27	monte carlo trials	800 drifted, 300 clean, 1100 primary total
28	sensor profile	camera_only_me8, camera_radar_overlay

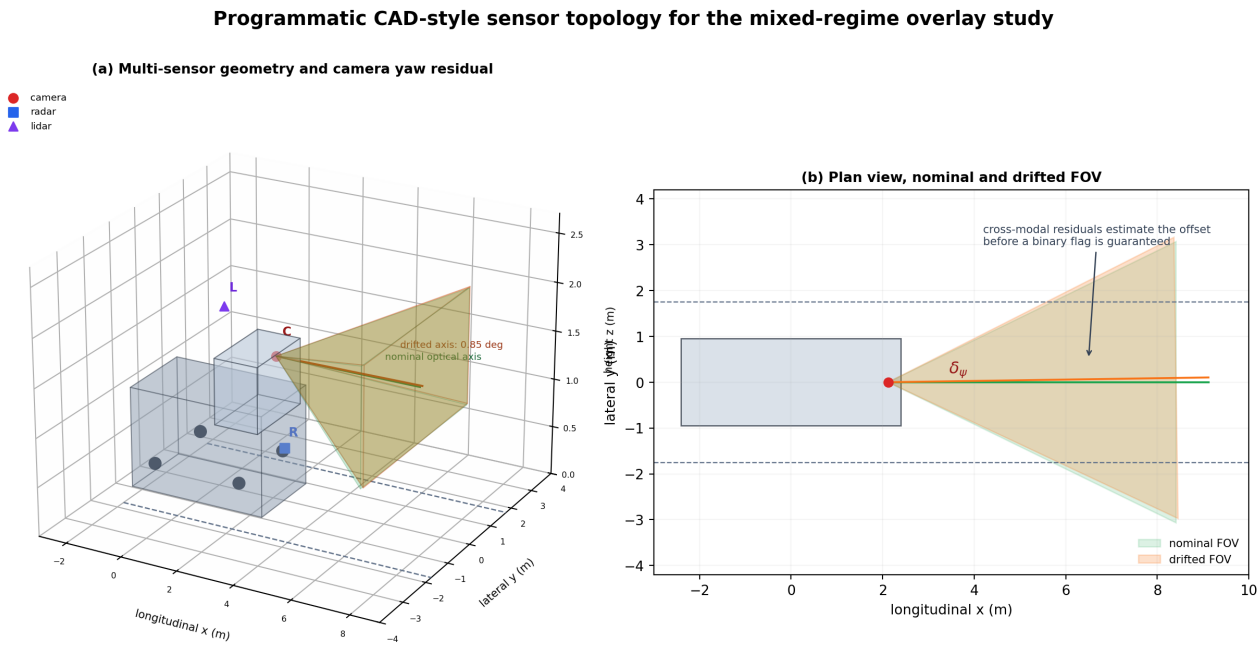

Figure 2: Programmatic 3D and plan-view geometry. The nominal and drifted frusta share an origin but differ in yaw. Cross-modal residuals estimate this disagreement from associated observations.

ID	Factor	Studied range or levels
29	remediation mid episode	False, True
30	ldw yaw proxy deg	0.05 to 0.12 deg
31	tamper yaw proxy deg	0.8 to 1.5 deg
32	true yaw alert threshold deg	0.18 to 0.35 deg
33	scenario regime	micro_gradual, moderate_gradual, severe_event, harsh_noisy, clean
34	flag midpoint deg	0.45 to 0.95 deg
35	flag transition width deg	0.08 to 0.22 deg
36	residual noise deg	0.025 to 0.14 deg
37	residual bias deg	-0.1 to 0.1 deg
38	residual dropout rate	0.0 to 0.35 fraction
39	residual outlier rate	0.0 to 0.1 fraction
40	correction efficiency	0.6 to 0.9 fraction
41	peak episode yaw deg	0.12 to 1.55 deg
42	tamper threshold deg	0.85 to 1.35 deg

4.7 Geometric visualization

Figure 2 is generated directly from the study code. It is a schematic, not a reverse-engineered vehicle CAD model. The oblique view labels camera, radar, and lidar locations, while the plan view shows how a yaw offset rotates the optical axis and creates a lateral discrepancy that grows with range. The illustrated 0.85-degree yaw is chosen for visibility in print and is not a universal detection threshold.

5 Methodology

5.1 End-to-end scoring procedure

Each episode begins by resetting adaptive state so that evidence cannot leak between vehicles. The generator produces raw EyeCAN-like frames and sensor observations from the same latent episode. The baseline is then evaluated using warning bits and vehicle context. NADIR arms process only frames for which the residual contract marks an observation available. Missingness therefore delays accumulation rather than appearing as zero drift.

Pulse scoring first invokes the shared production-oriented residual function. The benchmark then forms a scalar stress value from that output, the continuous yaw estimate, and camera drift evidence. The onboarding window supplies robust baseline location and dispersion. After onboarding, the current signed estimate is smoothed over five valid observations. Pulse requires magnitude, robust-score margin, and persistence. Meridian receives the same stress value and Pulse state, then applies motion gating and cumulative change. Apex receives the same inputs and evaluates both Meridian state and the change-posterior approximation. Echo observes a subsampled Pulse state and, if necessary, applies its longer robust trend rule.

The use of a shared stress value prevents each higher tier from receiving a hidden, tier-specific truth signal. The tiers differ through temporal memory and policy. They are still dependent arms because they share episodes, residuals, and parts of the implementation. Confidence intervals reported in this paper are per-arm binomial intervals, not independent-arm significance tests. A paired bootstrap or McNemar analysis is a better next step for formal arm-to-arm inference.

5.2 Baseline calibration and leakage prevention

An earlier implementation used the simulated onset frame to decide when the detector should begin thresholding. That introduces target leakage because a deployed monitor does not know the onset. The current implementation does not pass onset into tier logic. It freezes an onboarding baseline after at least eight and at most sixteen valid residual observations. The generator places onset after frame 24, leaving a calibration margin even with moderate dropout.

The onboarding assumption is operationally meaningful but optimistic. It presumes that the vehicle begins from an accepted calibration and that the first residual window is representative. If a vehicle enters monitoring already miscalibrated, the baseline can absorb part of the fault. A production system must anchor onboarding to a signed shop event, known-good service state, or fleet peer prior rather than silently trusting the first drive.

5.3 Robust instantaneous detection

Pulse uses a robust envelope rather than an unbounded Gaussian z-score. A minimum absolute score margin prevents a near-zero median absolute deviation from creating extreme sensitivity. A signed magnitude gate prevents a high dimensionless residual from automatically becoming a yaw claim. The short persistence rule suppresses isolated outliers. The combination is intentionally conservative and explains the zero clean-episode alerts in this fixture.

This zero is not proof of a zero false-alert probability. The Wilson upper confidence limit remains positive, and unmodeled nuisances can invalidate the result. The current implementation also samples

only a finite set of contexts and uses a frozen baseline. Seasonal drift, tire changes, road crown, repeated construction markings, sensor replacement, and clock failure require separate stress tests.

5.4 Persistent and change-point evidence

Meridian standardizes the stress value using robust baseline scale and updates a one-sided cumulative statistic only when vehicle speed exceeds a sampled motion gate. Weak or negative evidence reduces the statistic toward zero. A decision requires the cumulative threshold and two repeated strong observations. This rule converts persistent small deviations into a detectable event while limiting one-frame spikes.

Apex-lite maintains a running mean and variance after a twelve-observation lock. It computes a bounded approximation to change probability from residual surprise and a constant hazard. Independent Apex elevation requires a stronger residual margin, posterior support, and three consecutive observations. The implementation is designed for tractable Monte Carlo, not as the final production posterior. A full Apex implementation should retain the run-length distribution, use modality-aware likelihoods, and report posterior calibration on held-out sequences.

5.5 Advisory correction evaluation

At the first alert, the study uses the smoothed signed yaw estimate to form a bounded correction. The efficiency factor is sampled from 0.60 to 0.90, and the magnitude is capped at 0.80 degrees. Evaluation compares latent yaw before correction with latent residual after correction. This differs from the earlier implementation, which multiplied the observed estimate by a fixed 85% and therefore guaranteed improvement by construction.

The new evaluation can fail. If the estimate has the wrong sign or sufficient magnitude error, correction increases latent yaw. The correction-improvement rate is therefore reported beside mean reduction. This remains a shadow transform. The study does not evaluate whether a specific downstream perception or visualization stack accepts the transform, whether the transformed output improves object-level metrics, or whether any ECU permits such a write.

5.6 Metrics

Recall, specificity, precision, F1, and balanced accuracy follow their standard confusion-matrix definitions in Equations 18. Detection delay is the alert frame minus the simulated onset frame. Negative values indicate a nuisance alert before drift onset. Detection survival is the fraction of episodes without an alert as a function of frames since onset. “Below flag midpoint” records a NADIR alert while latent yaw remains below the episode’s sampled flag-transition midpoint. “Baseline missed” records a NADIR alert in an episode where the EyeCAN-Flag arm never elevates.

Compute timing is measured with a monotonic high-resolution clock inside the Python process and normalized per scored frame. These values compare implementation cost on one development environment. They are not hard real-time bounds. Wall-clock alert time is the sum of sensor availability, source buffering, gateway transport, poll quantization, scoring, persistence, uplink, and dispatch. In this study, transport cadence is orders of magnitude larger than algorithm compute.

5.7 Short implementation excerpt

Listing 1 shows the decision-critical portion of the hot path. It is intentionally short. Episode synthesis, CAN decoding, robust score construction, tier classes, SQLite writes, and export logic remain in referenced modules. Lines 1 to 4 reject missing observations and update the smoothed estimate. Lines 5 to 9 combine magnitude and robust-score evidence. Lines 10 to 12 require persistence before elevation.

Listing 1: Pulse persistence gate in the mixed-regime harness. Full implementation: `harness.py`.

```
if not frame["residual_available"]:
    continue
history.append(frame["continuous_yaw_deg"])
estimate = median(history[-5:])
margin = max(4.0 * baseline_mad, 0.45)
exceeds = (
    abs(estimate) >= 0.75 * alert_threshold
    and residual > baseline_median + margin
)
streak = streak + 1 if exceeds else 0
if streak >= 3:
    tier = CAUTION
```

6 Formal Model and Proofs

This section contains the mathematical statements used by the prose methodology. Displayed mathematics is concentrated here so that the experimental narrative remains readable. The results describe the idealized score and policy contracts. They do not prove that every implementation detail satisfies the assumptions.

6.1 Observation model and robust score

Definition 1 (Residual observation model)**.** *Let the latent extrinsic perturbation at frame t be $\delta_t \in \mathbb{R}^p$. The overlay receives*

$$z_t = A_t \delta_t + b_t + \varepsilon_t + o_t, \tag{1}$$

where A_t is a context-dependent sensitivity matrix, b_t is persistent bias, ε_t is zero-centered ordinary noise, and o_t is sparse contamination. Missing observations are represented by an availability indicator m_t, and no score update occurs when $m_t = 0$.

Definition 2 (Huberized covariance score)**.** *Given robust baseline location $\widehat{\mu}$, positive-definite covariance estimate $\widehat{\Sigma}_\lambda = \widehat{\Sigma} + \lambda I$, and componentwise Huber map*

$$\phi_c(u) = \begin{cases} u, & |u| \le c, \\ c\,\operatorname{sign}(u), & |u| > c, \end{cases} \qquad s_t = \phi_c(r_t)^{\mathsf{T}} \widehat{\Sigma}_\lambda^{-1} \phi_c(r_t), \qquad r_t = z_t - \widehat{\mu}, \tag{2}$$

the scalar s_t is the robust instantaneous evidence score.

Proposition 1 (Bounded single-observation influence)**.** *Assume $\lambda > 0$ and dimension $p < \infty$. For any residual vectors r and r', including an arbitrarily large contaminated residual r', the Huberized score in Equation 2 satisfies*

$$0 \le s(r') \le \frac{pc^2}{\lambda}, \qquad |s(r') - s(r)| \le \frac{pc^2}{\lambda}. \tag{3}$$

Consequently, one observation cannot make the instantaneous robust score diverge, although repeated bounded observations can still accumulate in a sequential detector.

Proof. The ridge term implies that every eigenvalue of $\widehat{\Sigma}_\lambda$ is at least λ. Therefore the largest eigenvalue of its inverse is at most $1/\lambda$. The Huber map clips every component of $\phi_c(r')$ to the closed interval from $-c$ to c, so its squared Euclidean norm is no larger than pc^2. Applying the Rayleigh quotient bound gives

$$s(r') = \phi_c(r')^\mathsf{T}\widehat{\Sigma}_\lambda^{-1}\phi_c(r') \le \lambda_{\max}\left(\widehat{\Sigma}_\lambda^{-1}\right)\left\|\phi_c(r')\right\|_2^2 \le \frac{pc^2}{\lambda}. \tag{4}$$

Positive definiteness gives the lower bound. Since both $s(r')$ and $s(r)$ lie in the same interval, their absolute difference is bounded by the interval width. The statement is intentionally about a single score. A sequential detector can still cross a boundary after many contaminated observations, and an adversary can exploit persistence. This is why Meridian adds quality and motion gates rather than relying on clipping alone. □

6.2 Calibration ranks

Definition 3 (Stratified calibration quantile)**.** *For context group g, let calibration scores $S_1, \dots, S_n$ and future score S_{n+1} be exchangeable. Define*

$$k_\alpha = \lceil (n+1)(1-\alpha) \rceil, \qquad \widehat{q}_{g,\alpha} = S_{(k_\alpha)}, \tag{5}$$

where $S_{(k)}$ is the k-th order statistic and the threshold is $+\infty$ if the requested rank exceeds n.

Theorem 1 (Finite-sample exceedance control)**.** *Under exchangeability within group g, the stratified threshold in Equation 5 obeys*

$$\Pr(S_{n+1} > \widehat{q}_{g,\alpha} \mid g) \le \frac{\lfloor (n+1)\alpha \rfloor}{n+1} \le \alpha. \tag{6}$$

Proof. Under exchangeability, the rank of S_{n+1} among the $n+1$ scores is uniform when ties are broken at random. The event that the future score exceeds the selected order statistic occurs only when its rank lies strictly above k_α. There are $n+1-k_\alpha$ such ranks. Therefore

$$\Pr(S_{n+1} > \widehat{q}_{g,\alpha} \mid g) \le \frac{n+1-k_\alpha}{n+1}. \tag{7}$$

Substituting the ceiling definition yields

$$n+1-k_\alpha \le n+1-(n+1)(1-\alpha) = (n+1)\alpha. \tag{8}$$

Because the numerator is an integer, it is no larger than $\lfloor (n+1)\alpha \rfloor$, proving the result. The proof does not require a Gaussian score distribution. It does require exchangeability within the declared group. Hardware replacement, software updates, seasonal shift, and changes in road or weather mix can violate that condition. A deployed monitor must therefore track group support and refresh calibration rather than treating Equation 6 as unconditional. □

6.3 Persistent evidence

Definition 4 (Motion-gated one-sided cumulative statistic)**.** *For standardized evidence x_t, allowance $k > 0$, threshold $h > 0$, and motion gate G_t, define*

$$C_0 = 0, \qquad C_t = \begin{cases} \max\{0, C_{t-1} + x_t - k\}, & G_t = 1, \\ C_{t-1}, & G_t = 0. \end{cases} \tag{9}$$

Lemma 1 (Deterministic crossing under persistent shift)**.** *Suppose the gate is open for m consecutive observations and each standardized observation satisfies $x_t \geq k + \epsilon$ for some $\epsilon > 0$. Then the statistic in Equation 9 crosses threshold h within at most*

$$m^\star = \left\lceil \frac{h - C_0}{\epsilon} \right\rceil \tag{10}$$

open-gate observations.

Proof. When the gate is open and $x_t \geq k + \epsilon$, the untruncated update is at least $C_{t-1} + \epsilon$. Since C_{t-1} is nonnegative, the maximum with zero cannot reduce that value. Induction gives

$$C_j \geq C_0 + j\epsilon \tag{11}$$

after j qualifying observations. Choosing the smallest integer j for which the right side is at least h gives Equation 10. Missing data and closed motion gates do not add evidence and therefore increase wall-clock delay without changing the count of qualifying observations. If evidence falls below the allowance, the bound no longer applies. This makes the lemma a persistence guarantee under stated conditions, not an expected-delay claim under arbitrary noise. □

6.4 Change posterior approximation

The Apex-lite implementation uses a bounded surprise-to-probability mapping with constant hazard. Its idealized update is

$$p_t = \frac{H\, L_{\text{new}}(z_t)}{H\, L_{\text{new}}(z_t) + (1 - H)\, L_{\text{run}}(z_t)}, \tag{12}$$

where H is the hazard and the two likelihoods represent a new segment and continuation of the current run. The benchmark uses an approximation rather than maintaining all possible run lengths. Equation 12 is therefore a design reference, not a claim that the lite implementation is an exact Bayesian filter.

6.5 Correction condition

Proposition 2 (Sufficient condition for correction improvement)**.** *Let latent scalar yaw be $\delta \neq 0$, signed estimate be $\widehat{\delta}$, and bounded efficiency satisfy $0 < \eta \leq 1$. Before correction the residual magnitude is $R_0 = |\delta|$; after an unsaturated correction it is*

$$R_1 = \left| \delta - \eta \widehat{\delta} \right|. \tag{13}$$

If the estimate has the correct sign and relative error satisfies

$$\left|\widehat{\delta} - \delta\right| \leq \rho|\delta|, \qquad 0 \leq \rho < 1, \tag{14}$$

then

$$R_1 \leq (1 - \eta + \eta\rho)\, R_0 < R_0. \tag{15}$$

Proof. Write the estimate as $\widehat{\delta} = \delta + e$. Substitution into Equation 13 gives

$$R_1 = \left|(1-\eta)\delta - \eta e\right|. \tag{16}$$

The triangle inequality and Equation 14 imply

$$R_1 \leq (1-\eta)|\delta| + \eta|e| \leq (1 - \eta + \eta\rho)\,|\delta|. \tag{17}$$

Since $\eta > 0$ and $\rho < 1$, the multiplicative factor is strictly below one. The correct-sign condition excludes a large opposite-direction estimate that may satisfy a weak absolute-error rule only near zero. If the correction saturates, the same conclusion can be checked directly using the clipped magnitude, but it is not automatic. The proposition explains why the empirical improvement rate is below 100% for several tiers: noisy early estimates can violate the relative-error or sign condition. □

6.6 Evaluation metrics

For true positives TP, false positives FP, true negatives TN, and false negatives FN, the reported classification metrics are

$$\begin{aligned} \text{Recall} &= \frac{TP}{TP + FN}, & \text{Specificity} &= \frac{TN}{TN + FP}, \\ \text{Precision} &= \frac{TP}{TP + FP}, & F_1 &= \frac{2\ \text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}, \\ \text{BalancedAccuracy} &= \frac{\text{Recall} + \text{Specificity}}{2}. \end{aligned} \tag{18}$$

Wilson score intervals are used for rate uncertainty. They quantify finite sample uncertainty under independent Bernoulli episodes but do not correct simulator misspecification, arm dependence, or tuning on related fixtures.

7 Results

7.1 Primary mixed-regime scorecard

Table 4 presents the main operating point. The EyeCAN-Flag baseline detected 391 of 800 drifted episodes for 48.9% recall. Pulse detected 658 episodes, Meridian 714, Apex 716, and Echo 698. Overall recall therefore increased by 33.4 percentage points for Pulse, 40.4 points for Meridian, and 40.6 points for Apex relative to the baseline. The result is a substantial difference in this fixture, but it is not the 98-point separation produced by the earlier hard-capped stress study.

Pulse and Echo produced no alerts in 300 clean episodes. Meridian and Apex each alerted in 7.0% of clean episodes. The baseline alerted in 8.3%, primarily because the episode-level rule treats any

Table 4: Primary scorecard across 800 mixed drift episodes and 300 clean episodes. Precision reflects the study prevalence. Delay is measured in simulation frames after latent onset.

Arm	Recall	FPR	Precision	F1	Bal. acc.	Delay (frames)
EyeCAN-Flag	48.9%	8.3%	94.0%	0.643	0.703	37.5
Pulse	82.2%	0.0%	100.0%	0.903	0.911	47.2
Meridian	89.2%	7.0%	97.1%	0.930	0.911	37.8
Apex	89.5%	7.0%	97.2%	0.932	0.912	37.0
Echo	87.2%	0.0%	100.0%	0.932	0.936	49.0

Table 5: Wilson 95% intervals for primary recall and clean-episode false-alert rate. These intervals address sampling error only.

Arm	Recall 95% CI	FPR 95% CI
EyeCAN-Flag	[45.4, 52.3]%	[5.7, 12.0]%
Pulse	[79.4, 84.7]%	[0.0, 1.3]%
Meridian	[86.9, 91.2]%	[4.6, 10.5]%
Apex	[87.2, 91.4]%	[4.6, 10.5]%
Echo	[84.8, 89.4]%	[0.0, 1.3]%

qualifying nuisance flag across a long episode as a positive. The correct interpretation is not that a production Mobileye system has an 8.3% false-warning rate. The result shows that a fleet monitor that OR-reduces exported warnings over a long window can create callbacks under the nuisance model.

Precision was 94.0% for the baseline, 100% for Pulse and Echo, and approximately 97.1% for Meridian and Apex. These values are conditional on 800 drifted and 300 clean episodes. At a field prevalence of one drift episode per hundred monitored vehicles, precision would be much lower for any nonzero false-alert rate. Deployment planning should use observed fleet prevalence, not the benchmark ratio.

7.2 Performance by drift regime

Regime stratification changes the story. In micro gradual drift, the baseline reached 18.5% recall. Pulse and Echo reached 57.0% and 55.5%, while Meridian and Apex reached 77.0% and 78.0%. The memory-bearing tiers therefore recovered many micro episodes missed by the baseline, but still missed more than one in five. This is a more realistic basis for a product claim than a perfect stress-fixture result.

Moderate gradual drift produced the widest practical separation. The baseline reached 33.0%, and every NADIR arm reached 100%. In this regime residual magnitude and duration were sufficient for every overlay arm, while the probabilistic flags remained absent or nonpersistent in many episodes. This is the clearest simulated evidence for the overlay use case.

Severe events reduced the gap. The baseline reached 90.5%, and every NADIR arm reached 100%. This confirms that the generator does not force the baseline to remain silent. In severe episodes, median frame delay was 14 for the baseline, 6 for Pulse, 5 for Meridian, 3 for Apex, and 7 for Echo because the residual channel responded directly to the simulated geometric change. Across all detected mixed episodes, baseline, Meridian, and Apex median delays were similar at 37.5, 37.8, and 37.0 frames. Transport must still be discussed separately because a native warning path can deliver an existing bit

Table 6: Episode recall by detector-independent operating regime. Each regime contains 200 drifted episodes.

Regime	EyeCAN-Flag	Pulse	Meridian	Apex	Echo
Micro gradual	18.5%	57.0%	77.0%	78.0%	55.5%
Moderate gradual	33.0%	100.0%	100.0%	100.0%	100.0%
Severe event	90.5%	100.0%	100.0%	100.0%	100.0%
Harsh / noisy	53.5%	72.0%	80.0%	80.0%	93.5%

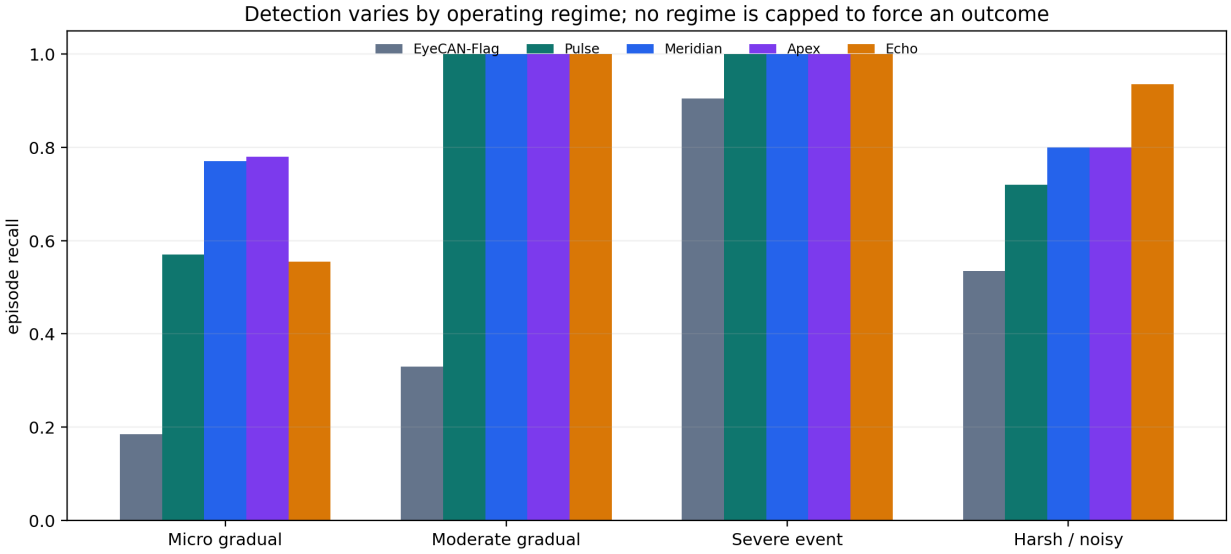

Figure 3: Recall varies with disturbance scale and observation quality. The EyeCAN-Flag baseline approaches the overlay arms in severe events, while persistent NADIR tiers separate most clearly in micro and moderate gradual drift.

much faster than a polled overlay.

Harsh noisy episodes produced intermediate results. Baseline recall was 53.5%, Pulse 72.0%, Meridian and Apex 80.0%, and Echo 93.5%. Echo's strong result follows from the long robust window, which suppresses individual outliers while retaining persistent trend. It should be replicated with additional nuisance families before being treated as a robust field ranking.

7.3 Detection survival

Figure 4 shows the fraction of episodes not yet detected after onset. In severe events, every curve falls rapidly and the remaining difference is mostly early delay. In moderate gradual drift, the baseline curve plateaus near 60% undetected while NADIR curves continue toward zero. In micro drift, Meridian and Apex continue accumulating evidence after Pulse and Echo plateau, but approximately 22% remain undetected at the episode horizon.

The survival curves also expose an issue hidden by endpoint recall. A monitor can eventually detect many episodes while requiring dozens or hundreds of frames. At a 5-second stream, 75 frames correspond to 6.25 minutes. At a 60-second poll, they correspond to 75 minutes. At a 300-second poll, they correspond to 6.25 hours. Product SLAs must therefore bind algorithmic frame delay to the actual ingest cadence.

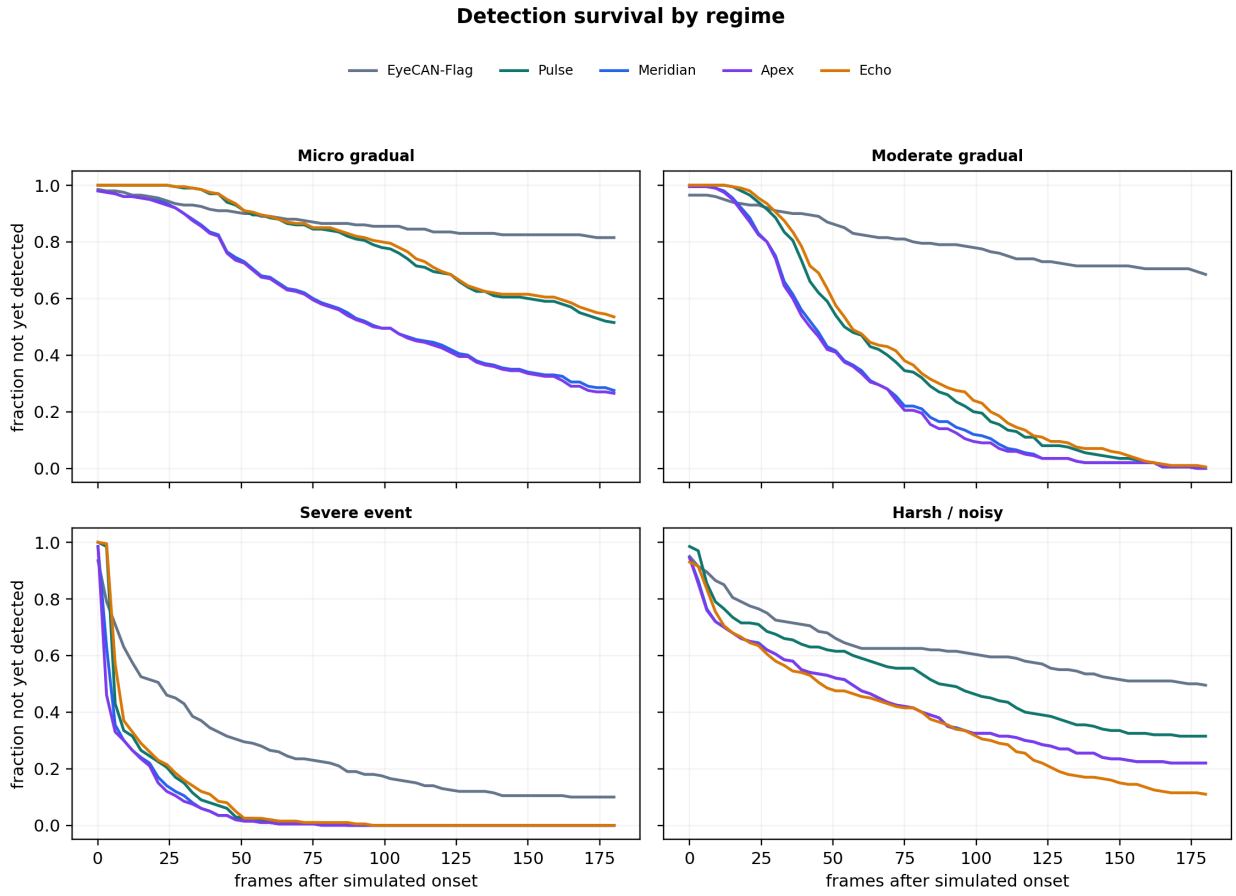

Figure 4: Detection survival after simulated onset. Lower is better. Curves combine detection probability and delay, and their wall-clock meaning depends on deployment cadence.

7.4 Recall and false-alert tradeoff

Figure 5 places each arm at its measured recall and clean false-alert operating point. Pulse offers the lowest-noise live path in the present fixture, but gives up approximately nine recall points to Meridian and Apex. Meridian and Apex occupy almost the same point because the Apex-lite independent confirmation rarely changes the Meridian endpoint. This near overlap is an important negative result: the current Apex-lite implementation adds little aggregate recall beyond Meridian on this benchmark.

Echo has the highest balanced accuracy, 0.936, because it combines zero clean alerts with 87.3% recall. It also has the highest F1, 0.932, by a negligible margin over Apex. No single metric establishes a universal winner. A fleet with expensive truck rolls may prefer Pulse or Echo. A safety-review team that tolerates more callbacks to recover micro drift may prefer Meridian or Apex. Threshold selection should be tied to a declared cost function.

7.5 Detection below the flag midpoint and baseline misses

Pulse detected below the sampled flag-transition midpoint in 62.5% of all drift episodes. Meridian reached 72.0%, Apex 72.9%, and Echo 66.1%. These metrics describe the simulated transition model. They do not identify an actual Mobileye blind zone. The more interpretable outcome is the paired episode result: Pulse detected 37.9% of all drift episodes in which the baseline never fired, Meridian

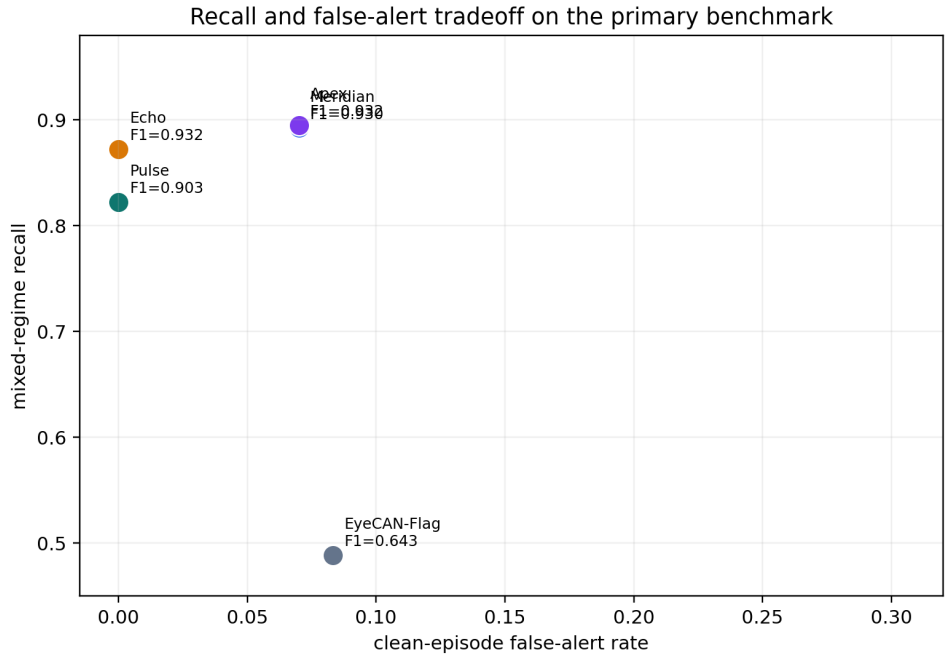

Figure 5: Primary operating points. The upper-left region is preferred. F1 values depend on the benchmark prevalence and should not be transferred directly to field operations.

Table 7: Bounded advisory correction. "Improved" means latent residual yaw after the transform is lower than before. The baseline does not estimate a correction.

Arm	Mean residual reduction (deg)	Improved corrections
EyeCAN-Flag	n/a	n/a
Pulse	0.311	98.6%
Meridian	0.189	92.8%
Apex	0.180	92.8%
Echo	0.311	97.2%

43.6%, Apex 43.9%, and Echo 40.0%.

The paired result is strongest in micro and moderate regimes. In micro drift, the baseline never fired in 81.5% of episodes; Meridian detected 61.5% of all micro episodes in which that occurred. In moderate drift, the baseline never fired in 67.0% and every NADIR arm detected those episodes. These are the episodes that motivate an overlay, provided the residual contract can be obtained in a real integration.

7.6 Advisory correction

Correction results are evaluated against latent yaw, not against the noisy estimate. Pulse reduced residual yaw by a mean 0.311 degrees and improved 98.6% of corrected episodes. Echo also reduced it by 0.311 degrees and improved 97.2%. Meridian and Apex produced smaller mean reductions, approximately 0.189 and 0.180 degrees, and improved about 92.8%. Their earlier alerts occur at smaller latent yaw and can rely on noisier early estimates, explaining part of the difference.

Figure 6 shows the distribution rather than only the mean. Negative values exist for every arm, confirming that the revised procedure does not guarantee success by construction. A production

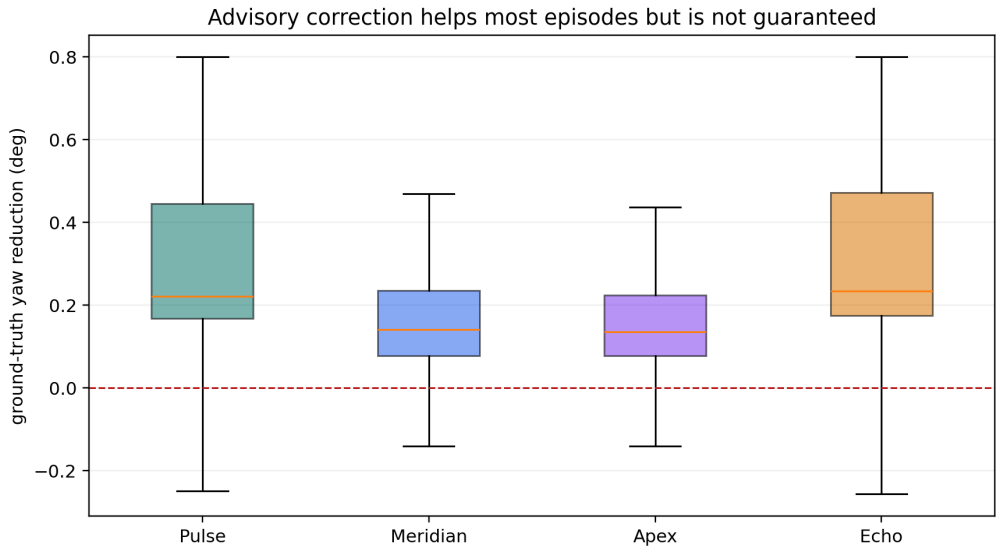

Figure 6: Distribution of latent yaw reduction after bounded advisory compensation. The dashed zero line separates improvement from degradation. Outliers are omitted from the boxplot display but retained in summary metrics.

Table 8: Measured algorithm compute. Shared decoding and transport are excluded, and higher-tier values should not be added without accounting for shared work.

Arm	Compute p50 (ms/frame)	Compute p95 (ms/frame)
EyeCAN-Flag	0.0010	0.0015
Pulse	0.1571	0.1663
Meridian	0.0230	0.0270
Apex	0.0487	0.0551
Echo	0.0550	0.0582

policy should withhold correction when sign confidence, estimate variance, observability, or cross-modal agreement fails. It should also evaluate object-level consequences, not only yaw residual.

7.7 Compute and transport timing

The binary decision path had median measured compute of approximately 0.001 ms per frame. Pulse required 0.157 ms, Meridian 0.023 ms beyond the shared input, Apex 0.049 ms, and Echo's amortized estimate was 0.055 ms. All p95 values remained below 0.17 ms per frame on the study machine. These are Python benchmark measurements, not certified worst-case execution times.

Transport dominates operational latency. CAN periods in the fixture span 66 to 100 milliseconds, while overlay polls span 5 seconds to 300 seconds. A 300-second poll is approximately three million times longer than the measured Pulse compute. Streaming ingest can remove poll quantization, but it cannot manufacture residuals that the upstream platform does not expose. For immediate driver HMI when a native bit exists, the native path remains the appropriate mechanism. The overlay is intended for fleet monitoring over minutes, hours, and service cycles.

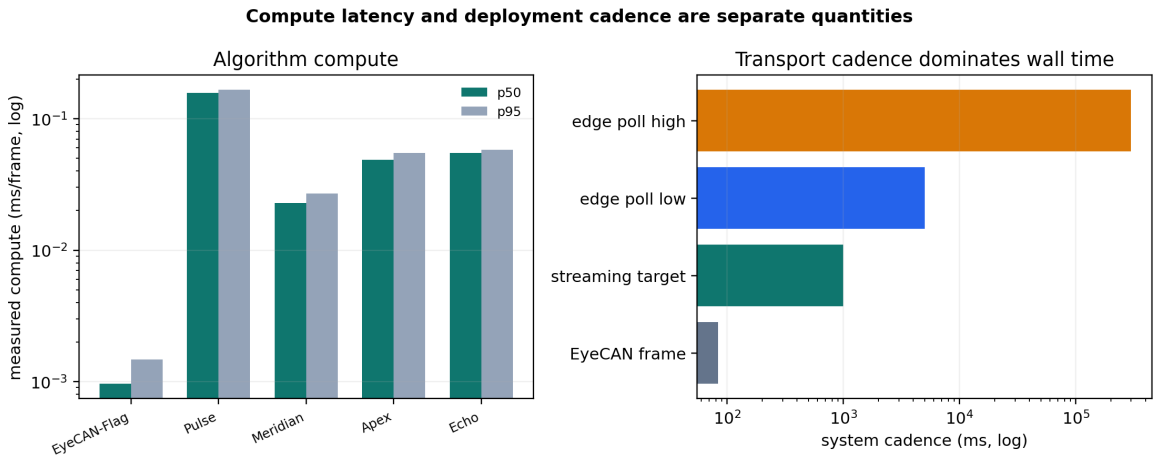

Figure 7: Algorithm compute and system cadence on logarithmic scales. End-to-end alert time is governed primarily by observation and transport cadence, not scorer arithmetic.

7.8 Sensitivity sweeps

One-factor sweeps varied peak yaw, flag midpoint, action threshold, CUSUM threshold, sensor profile, and onset profile with twenty episodes per level. Their purpose is diagnostic rather than inferential. Small cells create wide uncertainty, and changing one factor does not expose interactions. The table is included to make nonmonotonic behavior visible. For example, a higher yaw cap does not always produce a higher Pulse below-midpoint detection rate because it changes both detection timing and the condition defining the metric.

Factor	Value	EyeCAN	Pulse	Meridian	Apex	Echo
cusum_h	1.8	45%	75%	90%	90%	100%
cusum_h	2.5	55%	60%	85%	85%	85%
cusum_h	3.5	50%	75%	85%	85%	90%
me_visibility_yaw_deg	0.8	35%	65%	80%	80%	90%
me_visibility_yaw_deg	0.95	30%	85%	90%	90%	100%
me_visibility_yaw_deg	1.1	20%	60%	70%	75%	95%
onset_profile	exponential	30%	45%	60%	60%	90%
onset_profile	intermittent	45%	65%	70%	70%	95%
onset_profile	linear	40%	65%	80%	80%	100%
onset_profile	shock_recovery	55%	65%	80%	80%	95%
onset_profile	step	75%	85%	95%	95%	100%
peak_yaw_cap_deg	0.45	35%	55%	70%	70%	100%
peak_yaw_cap_deg	0.6	25%	85%	90%	90%	100%
peak_yaw_cap_deg	0.75	75%	90%	100%	100%	100%
sensor_profile	camera_only_me8	35%	65%	85%	85%	95%
sensor_profile	camera_radar_overlay	70%	75%	80%	80%	100%
true_yaw_alert_threshold_deg	0.18	65%	80%	80%	80%	100%
true_yaw_alert_threshold_deg	0.25	70%	65%	65%	65%	95%
true_yaw_alert_threshold_deg	0.35	45%	70%	85%	85%	85%

The sweeps show that sensor-profile labels alone do not guarantee a gain. The current synthetic camera-radar coupling is too coarse to establish the value of radar. A stronger next study should replay synchronized camera and radar associations, remove each modality in paired episodes, and report calibration error as a function of range, object class, weather, and association confidence.

8 Discussion

8.1 What the study supports

The results support a bounded statement: under the declared mixed simulator and additional residual contract, NADIR tiers provide more episode-level coverage than a fleet-visible EyeCAN-style binary flag monitor, especially for gradual micro and moderate yaw disagreement. The strongest practical operating points are Meridian and Apex when recall matters, and Pulse or Echo when nuisance callbacks matter. Bounded software compensation often reduces latent yaw, but it is not uniformly safe.

The study also supports a timing statement. The algorithmic cost of all arms is small relative to telematics cadence. That does not make NADIR faster than an on-vehicle warning. The binary baseline detects severe events earlier in frame count, and a native on-SoC path can reach driver HMI before a cloud or polled overlay. The overlay's value is additional observability and fleet evidence, not winning a millisecond race.

8.2 What the study does not support

The study does not prove that NADIR is more accurate than Mobileye as a company, EyeQ as a perception platform, or a production vehicle's internal calibration system. The comparator is an abstraction over observable flags. Its probabilistic thresholds and nuisance rates are experimental variables. No proprietary Mobileye source code, internal state, OEM acceptance criteria, or production trace was available.

The study does not validate physical software correction. It evaluates a bounded transform against simulated latent yaw. It does not demonstrate improved lane geometry, object association, AEB timing, or safety outcomes. It does not authorize writes to a production ECU. It does not replace a shop calibration or OEM sign-off.

The study does not establish field false-alert rates. Zero alerts in 300 clean episodes yield a finite upper confidence limit and can be overturned by a new nuisance family. The 7.0% Meridian and Apex rates may be too high for some fleets. The 8.3% baseline result is a property of an episode-level OR monitor under stress nuisances, not an estimate of driver warning quality.

8.3 Why the current differences are credible but provisional

Three redesign choices make the result more credible than the earlier paper. First, yaw endpoints span small to severe faults instead of remaining universally below a selected threshold. Second, flag probability changes smoothly and includes nuisance events rather than being forced silent. Third, residual observations contain bias, dropout, outliers, and weather-dependent noise, and correction is measured against truth.

The result remains provisional because both channels are synthetic. The continuous residual remains statistically coupled to latent yaw by design. Real associations can fail systematically rather than independently. Lane curvature, road crown, camera exposure, radar multipath, object motion, mounting flex, and clock skew can create structured errors that are more difficult than Gaussian noise plus outliers. Field evaluation must therefore focus on failure structure, not merely sample volume.

8.4 Tier interpretation

Pulse should be understood as the conservative live tier. Its 82.3% mixed recall and no clean alerts make it suitable for shadow pilots where operators must trust every callback. Its weakness is micro recall and long median delay. Meridian is the strongest general live detector in this implementation. It recovers an additional 7.0 percentage points overall and 20.0 points in micro drift relative to Pulse, at a 7.0% clean false-alert cost.

Apex does not yet justify a broad claim beyond Meridian. It adds 0.1 percentage points of overall recall and 0.5 points in micro drift while sharing the same clean false-alert rate. Its product value may instead be posterior evidence, replay, and provenance, which this benchmark does not score. Future Apex evaluation should include calibration of posterior probabilities, root-cause ranking, and analyst time.

Echo performs better than expected in harsh noise because its long window suppresses isolated corruption. It also has the best balanced accuracy. Its limitation is cadence. A nightly batch that detects a fault with excellent statistical accuracy may still be unacceptable for a fleet that needs same-route dispatch. The tier choice must include wall-clock service goals.

8.5 Deployment architecture

A defensible pilot begins with a signed calibration reference and a read-only residual contract. For the first phase, all outputs remain in shadow mode. Operators receive evidence but no automatic vehicle restriction or correction. The pilot records observation availability, clock health, residual distributions, warnings, service events, and operator dispositions. Thresholds are selected on a calibration fleet and frozen before validation.

The second phase compares alerts with independent ground truth. Static targets, dynamic OEM procedures, photogrammetry, or a calibrated reference rig can estimate pose before and after controlled disturbances. Naturalistic data should include ordinary repair variation, windshield replacement, pothole and vibration exposure, rain, snow, glare, lane-marking degradation, traffic density, and sensor blockage. Blinded reviewers should label whether an alert corresponds to measurable pose change.

The third phase evaluates advisory transforms in a sandboxed downstream consumer. Object bearing, lane projection, cross-modal association, and confidence calibration are measured before and after the transform. Any degradation causes automatic withholding. Production actuation requires a separate safety case, OEM authorization, cybersecurity review, rollback plan, and operational hazard analysis.

8.6 Recommended validation standard

A reasonable next-study standard is a preregistered, paired, vehicle-level evaluation. At least several vehicle and sensor configurations should be represented, with separate calibration and locked validation fleets. Each fault level should contain enough independent vehicles to estimate recall with useful confidence, rather than treating thousands of correlated frames as independent samples. Clean mileage should be large enough to measure callbacks per vehicle-day or per thousand miles.

The primary metrics should include yaw and translation estimation error, episode recall at declared fault magnitudes, false callbacks per vehicle-time, detection delay in wall-clock time and valid observations, calibration of confidence, and downstream geometric error. Results should be stratified by hardware, weather, speed, road geometry, repair type, and residual availability. Every comparison

should state whether arms receive identical inputs. If the overlay receives richer data, the study should report that as an architectural advantage, not hide it as an algorithmic win.

8.7 Threats to validity

Internal validity is limited by shared code and prior tuning. The tier thresholds were adjusted after observing earlier false-alert failures, so the present benchmark is closer to a development result than an untouched confirmatory test. A locked external validation set is required. The arm outcomes are paired and correlated, while the current confidence intervals are marginal. Formal comparisons should use paired tests or bootstrap vehicles.

Construct validity is limited by the EyeCAN-Flag abstraction and scalar yaw emphasis. Extrinsic calibration includes roll, pitch, translation, time offset, and potentially intrinsic effects. Binary warning behavior is influenced by perception confidence and road context, not yaw alone. The simulator's flag midpoint should be interpreted only as a sensitivity axis.

External validity is limited by synthetic data. Platform labels do not reproduce product-specific internals. The sensor profiles do not recreate raw camera, radar, or lidar phenomenology. Correction is not propagated through a real perception stack. Operational costs, bandwidth, retention, cybersecurity, and privacy are outside the numerical scorecard.

Statistical conclusion validity is limited by finite episodes, equal regime weighting, and exploratory factor sweeps. Precision depends on prevalence. Zero observed false alerts do not imply zero risk. The sweep cells contain only twenty episodes. Multiple comparisons are not corrected. These limitations are acceptable for a pilot-design paper, but not for a production certification claim.

9 Conclusion

This mixed-regime study replaces a benchmark that was structurally designed to make binary flags miss. The revised generator includes small, moderate, severe, and harsh disturbances; probabilistic flag onset; nuisance warnings; residual noise; bias; dropout; outliers; and imperfect bounded correction. Across 800 drifted and 300 clean episodes, the observable EyeCAN-Flag baseline reached 48.9% recall, compared with 82.3% for Pulse, 89.3% for Meridian, 89.5% for Apex, and 87.3% for Echo.

The result is not uniform superiority. The binary baseline reached 90.5% in severe events. NADIR's largest simulated advantage appears in gradual micro and moderate drift, where a continuous residual carries information that a flag-only monitor does not. Meridian and Apex provide the highest recall but incur a 7.0% clean-episode false-alert rate. Pulse and Echo are quieter but miss more micro drift. Apex-lite adds little endpoint recall beyond Meridian and needs a stronger forensic evaluation.

Bounded advisory compensation improved latent yaw in 92.8% to 98.6% of detected episodes, but failures remain and no physical downstream correction was validated. The paper therefore supports a shadow-pilot hypothesis: an explicit residual contract plus tiered monitoring can complement a fleet-visible EyeCAN flag path. It does not support claims of superiority over Mobileye's proprietary internal systems, safety certification, or replacement of OEM calibration.

Reproducibility and Artifact Availability

The benchmark is executed with `make mobileye-overlay-paper`. The runner, scenario model, arm harness, tier implementations, SQLite schema, table exporter, and 3D visualization code are located under `NADIR_SDK/python/nadir_sdk/benchmarks/mobileye_overlay`. The primary JSON scorecard is `out/mobileye_overlay_study/scorecard.json`; the trial database is `research/mobileye_overlay_paper/db/metrics.sqlite`. The claim boundary is maintained in `research/mobileye_overlay_paper/CLAIM_BOUNDARY.md`. Reproduction establishes software consistency, not field validity.

References

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Michèle Basseville and Igor V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

Jack Borer, Jeremy Tschirner, Florian Ölsner, and Stefan Milz. Continuous online extrinsic calibration of fisheye camera and LiDAR. *arXiv preprint arXiv:2306.13240*, 2023. URL `https://arxiv.org/abs/2306.13240`.

Yiqiang Chen, Feng Liu, and Ke Pei. Cross-modal matching CNN for autonomous driving sensor data monitoring. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2021. doi: 10.1109/ICCVW54120.2021.00346.

Lei Cheng and Siyang Cao. Online targetless radar-camera extrinsic calibration based on the common features of radar and camera. In *2023 IEEE National Aerospace and Electronics Conference*, pages 294–299, 2023. doi: 10.1109/NAECON58068.2023.10366051.

Maurício Braga de Paula, Cláudio R. Jung, and Luiz Gonzaga da Silveira Jr. Automatic on-the-fly extrinsic camera calibration of onboard vehicular cameras. *Expert Systems with Applications*, 41(4): 1997–2007, 2014. doi: 10.1016/j.eswa.2013.08.096.

Thomas Goelles, Birgit Schlager, and Stefan Muckenhuber. Fault detection, isolation, identification and recovery methods for automotive perception sensors including a detailed literature survey for LiDAR. *Sensors*, 20(13):3662, 2020. doi: 10.3390/s20133662.

Dhruv Hegde and Srivatsan Balaji. Distribution-free fleet monitoring of post-repair adas extrinsic drift under sparse telemetry proxies. NADIR Systems technical report, 2026.

D. V. Hinkley. Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523, 1971.

Peter J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1): 73–101, 1964.

International Organization for Standardization. Road vehicles: Functional safety. *ISO 26262*, 2018.

International Organization for Standardization. Road vehicles: Safety of the intended functionality. Technical Report ISO 21448:2022, ISO, 2022. URL `https://www.iso.org/standard/77490.html`.

International Organization for Standardization. Road vehicles: Logical interface between sensors and data fusion unit for automated driving functions, part 20. Technical Report ISO 23150-20:2026, ISO, 2026. URL `https://www.iso.org/standard/87976.html`.

Ganesh Iyer, R. Karnik Ram, J. Krishna Murthy, and K. Madhava Krishna. CalibNet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1110–1117, 2018. doi: 10.1109/IROS.2018.8593693.

Jesse Levinson and Sebastian Thrun. Automatic online calibration of cameras and lasers. In *Proceedings of Robotics: Science and Systems IX*, 2013. doi: 10.15607/RSS.2013.IX.029.

Xingchen Li, Yuxuan Xiao, Beibei Wang, Haojie Ren, Yanyong Zhang, and Jianmin Ji. Automatic targetless LiDAR-camera calibration: A survey. *Artificial Intelligence Review*, 56(9):9949–9987, 2023. doi: 10.1007/s10462-022-10317-y.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):3292–3310, 2023. doi: 10.1109/TPAMI.2022.3179507.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

Mobileye 8 Connect Technical Installation Guide. Mobileye, 2022. Installation Center, target acquisition calibration, and recalibration procedures.

Mobileye. Mobileye support guides and manuals. Official support portal, 2026. URL `https://www.mobileye.com/support/`. Accessed July 17, 2026.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Gaurav Pandey, James R. McBride, Silvio Savarese, and Ryan M. Eustice. Automatic targetless extrinsic calibration of a 3d LiDAR and camera by maximizing mutual information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 2053–2059, 2012. doi: 10.1609/aaai.v26i1.8379.

Juraj Peršić, Luka Petrović, Ivan Marković, and Ivan Petrović. Online multi-sensor calibration based on moving object tracking. *Advanced Robotics*, 35(3–4):130–140, 2021. doi: 10.1080/01691864.2020.1819874.

SAE International. Taxonomy and definitions for terms related to driving automation systems. *SAE J3016*, 2021.

Nick Schneider, Florian Piewak, Christoph Stiller, and Uwe Franke. RegNet: Multimodal sensor registration using deep neural networks. In *2017 IEEE Intelligent Vehicles Symposium*, pages 1803–1810, 2017. doi: 10.1109/IVS.2017.7995968.

Loren Stowe, Javier Alcazar, Steven Huggins, Gregory T. Beale, Matt Palmer, and Zachary R. Doerzaph. Safety implications of potential advanced driver assistance systems sensor degradation. Technical Report DOT HS 813 740, National Highway Traffic Safety Administration, 2025.

Zhiguo Tan, Xing Zhang, Shuhua Teng, Ling Wang, and Feng Gao. A review of deep learning-based LiDAR and camera extrinsic calibration. *Sensors*, 24(12):3878, 2024. doi: 10.3390/s24123878.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

John Waraniak. Unsettled issues on sensor calibration for automotive aftermarket advanced driver-assistance systems. Technical Report EPR2021008, SAE International, 2021.

Pengjin Wei, Guohang Yan, Xin You, Kun Fang, Tao Ma, Wei Liu, and Jie Yang. Online LiDAR-camera extrinsic parameters self-checking and recalibration. *Measurement Science and Technology*, 35(10): 105204, 2024. doi: 10.1088/1361-6501/ad6469.