vehicles

MDPI

Article

Object Awareness for Intelligent Vehicle Systems: A Software Framework for Selective and Image-Aligned Presentation of YOLO Perception Outputs

Mihai Duguleană * and Tiberiu Ghiță

Department of Automotive and Transportation, Faculty of Mechanical Engineering, Transilvania University of Brașov, 500036 Brașov, Romania
* Correspondence: mihai.duguleana@unitbv.ro

Abstract

PRISM (Perception-to-Risk Interface for Situational Mediation) couples object detection, monocular geometry, heuristic priority scoring, and bounded image-aligned cues to investigate how vehicle perception outputs can be translated into selective visual information. The framework combines cue-count and representation-cost limits with temporal persistence, adaptive detail, and spatial attenuation to control display activity. This software study evaluates allocation consistency using 554 COCO images, 2580 synthetic frames from eight scenes, 360 allocation configurations, analytic stress tests, and a 1200-frame real-video demonstration. Against an unthresholded risk ranking (risk budget), clear-scene PRISM reduces cue count by 54.7%, switching by 81.6%, and painted area by 30.1%, while displaying 213 of 325 score-defined reference frames, including 213 of 227 detected references. Simpler dwell ranking retains the same references with fewer transitions, identifying persistence as a major contributor to display stability. A paired geometry-to-selection diagnostic shows that box-height noise changes selected identities in 12.3% or 64.3% of observations with one or five cue slots, respectively, under specified competing-candidate conditions. Logged state and allocation time averages 0.117 ms per frame, excluding decoding, rendering, and display. The real recording demonstrates executable integration and false-cue propagation. Together, these results characterize the trade-offs among information retention, display activity, and geometric sensitivity, supported by publicly available code, experimental records, and reproducible analyses.

Keywords: YOLO; augmented reality; driver assistance; human–machine interface; time-to-collision; selective visualization; vehicle software

Check for updates

Academic Editor: Pak Kin Wong

Received: 16 September 2026
Revised: 19 October 2026
Accepted: 20 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

A vehicle interface must decide not only what is present in the environment, but which observations warrant the driver's attention. Detection confidence alone does not capture that decision: a confidently detected distant vehicle may be less relevant than a nearby pedestrian whose trajectory is uncertain. Conversely, displaying every observation can occupy substantial image space and repeatedly change the visible cue set.

Augmented-reality driving interfaces connect sensed objects with displayed cues, as demonstrated by DAARIA [1], while studies of hazard perception, workload, and display opacity evaluate the consequences for drivers [2–4]. These studies leave open a design problem: how should a system select and maintain a limited set of cues as observations

change? Risk-ranked selection under fixed capacities, without activation or persistence, provides the baseline for quantifying selective presentation. PRISM combines detection confidence with object geometry and track history to assign presentation priority. PRISM investigates an explicit software layer between perception and presentation.

Situation awareness encompasses perception, comprehension, and projection [5]. Appropriate reliance on automation depends on how a system communicates its capabilities and limitations [6]. These concepts motivate selective presentation that connects observations with interpretable cues. Driving-explanation studies show that explanation content and modality influence driver understanding and performance [7,8]. Evaluation therefore connects the information displayed and omitted with the driver's response. Context-aware computing adapts application behavior to information about a situation [9]. In PRISM, observed class, estimated proximity, approximate corridor position, and track history supply the context for cue allocation.

Object detection provides increasingly efficient components for such a system. Surveys describe the development of detection architectures, the YOLO lineage, and their application to autonomous driving [10–12], while YOLOv7, YOLOv9, and YOLOv10 demonstrate advances in training, gradient information, and end-to-end inference [13–15]. A parallel approach reduces the cost of running such models away from server hardware, through compression for near-edge processors [16], cooperative edge–cloud partitioning in vehicles [17], and lightweight architectures intended for mobile augmented reality [18]. Both lines of work address the supply of observations. Detector selection remains an engineering trade-off among accuracy, execution time, and available compute. Cue allocation adds a downstream decision: selecting how many observations to display and balancing temporal stability with newly urgent candidates. This decision is the subject of this study.

This paper presents PRISM, the Perception-to-Risk Interface for Situational Mediation. The framework connects tracked objects to monocular state estimates, an interpretable priority score, an evidence-quality indicator, and a bounded set of image-aligned cues. Its contribution is a reproducible systems investigation of temporal cue allocation: controlled comparisons isolate thresholding and persistence, painted area distinguishes numerical budgets from occupied pixels, and analytic stress tests expose geometric failure mechanisms.

Three research questions guide the study. RQ1: How do thresholding, hysteresis, dwell, and central attenuation change score-defined reference retention, painted area, and cue switching? RQ2: How do parameter and geometric perturbations affect estimates and final selection? RQ3: What execution costs and visible failures occur in the recorded pipeline? The evaluation characterizes software behavior and component execution on the recorded research hosts; Section 5.3 defines the scope of the evidence and Section 5.4 develops the next steps for driver and target-platform validation.

2. Related Work

We organize the related work along the pipeline the proposed system follows—from sensing, through selection, to display. Section 2.1 reviews detection and execution, which supply the raw observations. Section 2.2 covers geometry and criticality, which rank those observations into candidate priorities. Section 2.3 surveys augmented-reality presentation and how it has been evaluated with drivers.

2.1. Object Detection and Edge Execution

This section reviews object detection and edge execution as the upstream source of PRISM's observations, with attention to the conditions under which reported accuracy and latency transfer to a deployed platform. Schlosser et al. describe benchmarking across execution targets [19], Jacob et al. study integer quantization [20], and Al Amin et al. demonstrate

FPGA-based detection [21]. Setyanto et al. organize near-edge compression methods by the detector component being modified [16]. Liang et al. investigate vehicle–cloud partitioning through Edge YOLO [17]. Together, these studies show why model accuracy and execution cost must be reported under a specified platform and timing boundary.

Recent YOLO system studies reinforce the need for task-specific model selection. Chaman et al. benchmark YOLOv8–YOLOv12 on a unified traffic dataset and report accuracy and execution trade-offs for ADAS and intelligent transportation functions [22]. Vaghela et al. compare YOLO variants using depth-camera observations collected by a robot on a campus [23]. Their evaluation connects model choice to the sensor and operating environment. In a different application domain, Srijiranon et al. combine separate drowning and surface-vehicle detectors with a monitoring station and rescue vehicle, and test the integrated Robobuoy system in a controlled pool [24]. PRISM uses YOLO detections associated through the Ultralytics ByteTrack interface. ByteTrack links observations across frames and uses lower-confidence detections to recover associations [25].

The objects that are hardest to detect are also those whose image geometry is least well-conditioned. Small objects carry few pixels and limited appearance information; reviews of small-object detection identify multi-scale feature fusion, enrichment of available information, and positive-negative sample balance as the recurring remedies [26], and recent YOLO derivatives pursue the same objective through multi-scale fusion modules and scale-sensitive losses [27]. For an interface layer that derives range from bounding-box height, this weakness appears twice. The distant pedestrian or traffic light most likely to be missed also produces the box whose height is least reliable, so small image scale can affect both detection and geometric estimation.

RT-DETR offers a transformer-based alternative to YOLO for object detection [28]. Its published accuracy and throughput concern detector execution under its own dataset and hardware protocol. YOLO26 extends the set of candidate perception components. Official documentation describes a lighter detection head and an optional end-to-end path without a separate nonmaximum-suppression stage [29].

2.2. Geometric State and Criticality

Optical expansion provides a time-to-contact cue under constrained shape and motion assumptions [30]. This monocular temporal measure is independent of a constant absolute object-size scale. Collision assessment additionally involves lateral motion, object extent, road curvature, and braking. Criticality metrics are selected for a specified driving task [31]. PRISM combines temporal and spatial terms into a bounded heuristic priority score and evaluates preservation of the resulting reference through perception and allocation. Confidence, temporal stability, and geometric fit quality describe complementary aspects of an estimate. Calibration methods relate neural-network confidence to observed correctness [32]. PRISM combines confidence and temporal indicators into a heuristic evidence-quality index that controls rendering detail.

2.3. Augmented Reality and Selective Presentation

Visual clutter reflects the perceptual structure and arrangement of graphical elements [33]. Similarly, a visual-attention explanation of a driving model [34] and a rule-generated object label answer different questions. The former relates model behavior to image regions; the latter explains why an object receives a particular interface priority.

DAARIA already connected obstacle perception, dangerousness assessment, and augmented-reality assistance [1]. PRISM builds on this integration theme. Its focus is a reproducible allocation contract: urgent candidates take precedence over less urgent persistence preferences, every admitted cue must satisfy resource limits, and rendering

preserves the allocated level of detail. Comparisons with stateless rankings and component ablations examine the consequences of that contract. More recent systems place the detector on the display device itself. A YOLOv8 pipeline has been demonstrated running onboard a head-mounted display in real time, without offloading to an external server [35], and edge co-processing with a dedicated accelerator has been used to stream inferences back to a headset for spatial reprojection into the user's view [36]. Reviews of lightweight detectors for mobile augmented reality compare candidate architectures on accuracy, speed, small-object performance, and model size [18]. These contributions establish that detections can be produced and registered within the constraints of a wearable or vehicle-mounted display. Schall et al. evaluated AR cues with 20 older drivers in a simulator and measured hazard detection and interference with other tasks [2]. That study provides participant-level evidence for its cueing conditions.

Contemporary AR-HUD evaluation also examines how representation affects attention and workload. Maag et al. use a detection-response task to assess workload effects [3], while Lopez and Moacdieh examine opacity, visual attention, and situation awareness [4]. These studies motivate evaluating perceptual consequences separately from rendered area.

3. Materials and Methods

3.1. System Scope and Architecture

The framework accepts forward-scene camera images and timestamps. Its eight supported COCO categories are person, bicycle, car, motorcycle, bus, truck, traffic light, and stop sign [37]. The categories identify object types; rider association, signal-state recognition, and regulatory-sign interpretation require additional processing. Figure 1 shows the processing stages and the feedback of previously displayed identities into cue allocation. The design objectives are to bound simultaneous cues, reduce identity additions/removals, retain the specified reference where detected, and expose the reason for each selected cue. The prototype uses camera-image coordinates; deployment through a windshield or head-mounted display requires the optical registration described below. In Figure 1, B denotes the surrogate-cost budget and K the cue-count limit. Previously displayed identities feed back into allocation to support temporal persistence.

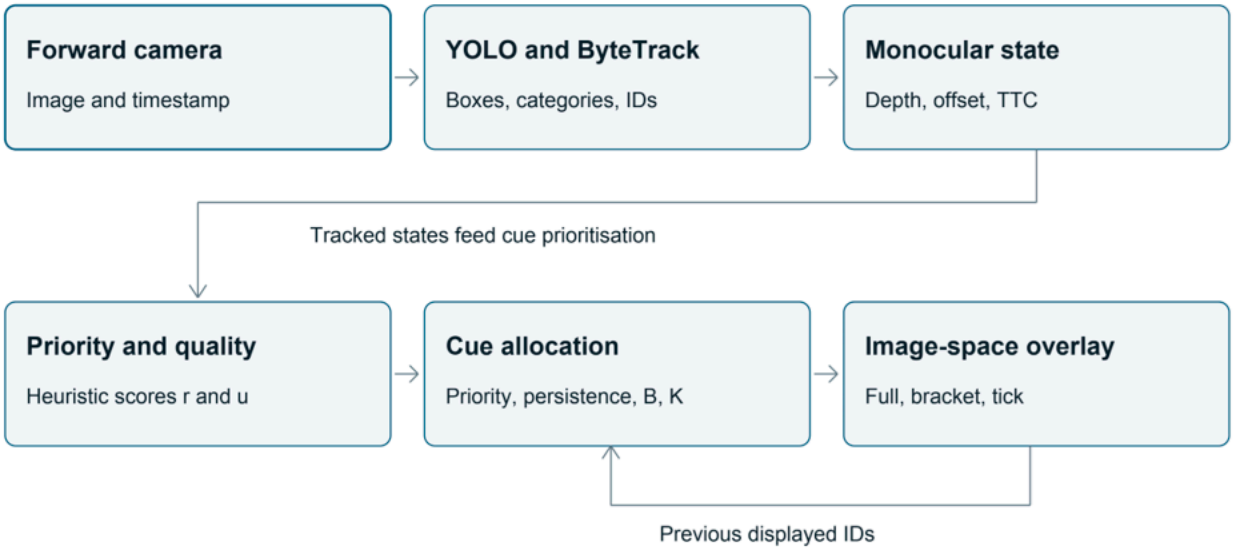


Figure 1. PRISM pipeline from camera observations to bounded cues, with feedback for temporal persistence.

The proposed vehicle integration is a calibrated forward camera, a local compute unit, and a display controller. The compute unit would execute frame acquisition and preprocessing, a compatible detector runtime, association, state reconstruction, scoring, and allocation. Detector inference may use a supported accelerator while tracking and allocation execute on the host CPU. The present implementation processes image/video inputs and draws camera-coordinate overlays. Even though we do not focus on the actual implementation in this paper, vehicle deployment will most likely need to extend our proposal with camera integration requiring timestamped frames and measured distortion, especially when resizing (to preserve the calibration used by geometry). A deployed implementation would also need freshness checks and a defined response to camera loss, processing overruns, and invalid geometry. A windshield HUD would require calibrated projection into the driver's view, accounting for display optics, eye position, and vehicle motion. The present corridor estimate uses camera-derived geometry. Incorporating steering, yaw rate, or vehicle speed would extend state estimation and would require synchronized units, timestamps, coordinate frames, and a validated read-only vehicle-bus interface.

3.2. Detection and Temporal State

The perception configuration uses YOLO11n at nominal input size 640 pixels, a detection-confidence threshold of 0.25, and bytetrack.yaml association. Cached observations contain boxes, categories, confidence, identities, and ages. Shared temporal reconstruction supplies the same candidate states to every policy. The tracker defaults include high/new-track thresholds of 0.25, a low-score threshold of 0.10, a 30-frame buffer, and a match threshold of 0.8. Because the detector discards scores below 0.25 upstream, the nominal 0.10–0.25 ByteTrack recovery band receives no observations in this configuration. For each identity, the implementation retains nine timestamp–height samples for time-to-collision (TTC) and the last eight confidence and consecutive-box intersection-over-union (IoU) observations for quality scoring. Histories expire after more than one second without observation. The pipeline accepts acquisition timestamps; the synthetic experiment uses frame index divided by the nominal 20 Hz sequence rate, while the real recording uses its supplied acquisition timestamps. Table A1 consolidates the fixed perception and temporal settings. YOLO11n generates the shared cache used for the allocation comparisons.

3.3. Monocular Geometry

Let W be image width in pixels, θ horizontal field of view, f focal length in pixels, and c_x the horizontal principal point (we assume square pixels and a horizontally centered principal point). The camera relations are displayed in Equation (1). Given box height h in pixels, horizontal box center x_c in pixels, and class-height prior H in meters, Equation (2) gives longitudinal depth Z and lateral offset X in meters.

$$f = W/[2\,tan(\theta/2)], \quad c_x = \frac{W}{2} \tag{1}$$

$$Z = fH/h, \quad X = (x_c - c_x)Z/f \tag{2}$$

Table A2 lists the assumed class-height priors H, dispersions σH, and priority weights $v(c)$. A bicycle prior refers to the detected bicycle box, with rider coverage dependent on the observation. Truncation, occlusion, pitch, pose, and variation in physical dimensions affect depth reconstruction.

First-order propagation uses a box-height noise scale σh and assumed physical-height dispersion σH. Equation (3) defines σh in pixels and the relative depth dispersion $\sigma Z/Z$. This model describes sensitivity to the specified geometric assumptions. With exact local-

ization, estimated-to-true depth is *H/H_true*; representing a 1.20 m person by H = 1.70 m overestimates depth by 41.7%.

$$\sigma h = 2 + 0.03h, \qquad \sigma Z/Z = \sqrt{\left[(\sigma H/H)^2 + (\sigma h/h)^2\right]} \tag{3}$$

For a rigid object under constant relative longitudinal closing speed v, define initial depth Z_0, time t, reciprocal height q, and fitted intercept a and slope b as in Equation (4). The least-squares fit uses at most nine observations and at least four strictly increasing timestamps. The implementation uses observation times in seconds; b therefore has units of pixel^{-1} s^{-1}. TTC is computed at the latest observation *t_latest* (Equation (5)).

$$h(t) = fH/(Z_0 - v\,t), \quad q(t) = 1/h(t), \quad \hat{q}(t) = a + bt \tag{4}$$

$$TTC(t_latest) = -\hat{q}(t_latest)/b, if\ b < -10^{-12}\ and\ \hat{q}(t_latest) > 0 \tag{5}$$

The slope guard of 10^{-12} pixel^{-1} s^{-1} stabilizes numerical division. Under the ideal model, $b = -v/(fH)$. A vehicle maintaining fixed relative depth has $v = 0$ and $b = 0$, as does a stationary camera–object configuration. The estimate describes relative longitudinal approach; lateral motion, camera rotation, pose change, and box noise affect its interpretation. Nonclosing and unsupported estimates are internally infinite and serialized as null; their temporal urgency contribution is zero, while distance remains active. Fit R^2 describes reciprocal-height fit quality.

3.4. Priority and Evidence Quality

The clipping function is presented in Equation (6):

$$clip(x) = min(1, max(0, x)) \tag{6}$$

Detection confidence p, observation count *age*, and category c define evidence E and class weight $v(c)$. U is combined urgency, C approximate corridor membership, and r the final bounded priority. TTC is in seconds, Z and X in meters, and all scores are dimensionless. Table A3 explains the constants in Equations (7)–(10) and the evidence-quality index.

$$U = max\{clip[(6 - TTC)/4.5], 0.55\ clip[(45 - Z)/37]\} \tag{7}$$

$$C = exp\{-max[\,|X| - (1.6 + 0.02Z), 0]^2/(2 \times 1.5^2)\} \tag{8}$$

$$E = clip(0.45 + 0.55p)\ clip[0.55 + 0.15\ min(age, 3)] \tag{9}$$

$$r = v(c)\ U\ (0.35 + 0.65C)\ E \tag{10}$$

The vulnerability weights $v(c)$ are 1.00 for person, 0.95 for bicycle, 0.90 for motorcycle, 0.66 for truck, 0.62 for car and bus, and 0.30 for traffic light and stop sign. They encode a fixed design preference for unprotected road users. The bound $r \leq v(c)$ keeps infrastructure categories below the 0.55 urgent threshold. Traffic-signal and regulatory-sign warnings would require separate state recognition and event rules. For unavailable or non-closing TTC, the temporal term is zero and distance can still activate urgency. Corridor membership denotes present spatial overlap with an approximate corridor, without predicting lateral motion. The numerical spacing fixes this ordering for reproducible comparisons.

The defaults specify a bounded heuristic reference policy. In Equation (7), 4.5 s and 37 m are the differences between the chosen onset and saturation limits (6–1.5 s and 45–8 m), ensuring linear ramps between zero and one. Taking the maximum avoids adding two correlated proximity indicators. The 0.55 distance multiplier makes distance-only

urgency weaker than saturated temporal urgency. These engineering horizons define the scoring range used in the experiments.

In Equation (8), the nominal corridor half-width grows from 1.6 m by 0.02 m per meter of depth; the 1.5 m Gaussian margin avoids a discontinuity at its boundary. This gives candidates near the nominal corridor boundary a smoothly varying spatial weight. Equation (9) retains nonzero evidence for a new or lower-confidence track instead of discarding it solely through multiplicative scoring. The age factor reaches one after three observations, while the confidence factor reaches one at $p = 1$. In Equation (10), $0.35 + 0.65C$ leaves a residual outside-corridor priority and preserves the upper bound $r \leq v(c)$. Together, these properties define the bounded reference policy used throughout the comparisons.

The interaction with the default activation threshold is consequential. For a mature, in-corridor candidate with $p = 0.9$, $E = 0.945$. With no usable TTC, the 0.24 activation threshold is reached at approximately 27.9 m for a person but 17.4 m for a car; with negligible distance contribution, the corresponding TTC cutoffs are about 4.86 and 4.16 s. These algebraic cutoffs describe the activation behavior of the rule. The class weights deliberately change activation as well as ranking. The sweeps evaluate allocator settings around this fixed scoring policy.

The quality indicator u combines six clipped components with weights (0.22, 0.14, 0.16, 0.18, 0.18, 0.12): one minus confidence; confidence standard deviation divided by 0.20; box instability $clip[(0.92 - mean\ IoU)/0.35]$; relative depth dispersion divided by 0.30; one minus TTC-fit R^2; and immaturity $clip[(5 - min(age,5))/5]$. Confidence variability defaults to 0.5 with fewer than three observations, box instability to 0.5 without history, and the TTC-fit component to 0.85 when TTC is unavailable. Larger u indicates weaker or less stable evidence. The six weights sum to one and represent a fixed design mixture. The weights and normalization constants define a heuristic index; the ablations evaluate its effect on allocation and detail selection.

3.5. Temporal Cue Allocation and Presentation

PRISM assigns one of three detail levels: full for $r \geq 0.55$, bracket for $r \geq 0.30$, and tick otherwise. When $u > 0.72$ and $r < 0.60$, the quality gate assigns a tick. Fixed surrogate costs are 0.0165, 0.0072, and 0.0022, respectively. The default cost limit B is 0.055 and cue-count limit K is five. Table A4 documents the allocation settings and their intended roles. Costs use fixed policy units; rendered pixel occupancy is measured separately. The count cap and persistence duration control display density and temporal behavior.

A new cue requires raw score $r \geq 0.24$. A previously displayed identity receives a 0.06 bonus when persistence is enabled, and its effective score $r_eff = min(1, r + 0.06)$ is compared with 0.15. Outside the dwell preference, this permits continuation down to *raw* $r = 0.09$. During the first eight displayed frames, the dwell preference can keep an identity eligible below that threshold, but admission still requires feasible capacities. A fixed central rectangle spans 34–66% of image width and 40–78% of image height. If its box IoU with a candidate exceeds 0.10 and the candidate has raw $r < 0.55$, effective ranking priority is multiplied by 0.45. The rectangle is a road-center heuristic. Candidates are ordered first by urgency (*raw* $r \geq 0.55$), then persistence preference, then effective priority, with track identity breaking ties. A candidate is admitted only if both capacities remain feasible and its box IoU with every admitted candidate is at most 0.55. Persistence can therefore be preempted and never overrides the resource constraints. The allocation is greedy—an urgent cue may still be infeasible when the configured budget cannot accommodate its representation.

Rendering preserves the selected detail level. Full cues can carry a short class–corridor–proximity label; increasing u introduces dashed strokes and a softer halo. Brackets show corners and a base ellipse, while ticks show the ellipse alone. Figure 2 connects the detection

and presentation stages using a frame from the real recording described in Section 3.8. Color and shape differences are implemented properties. Optional status and summary panels are separate from the surrogate cue budget.

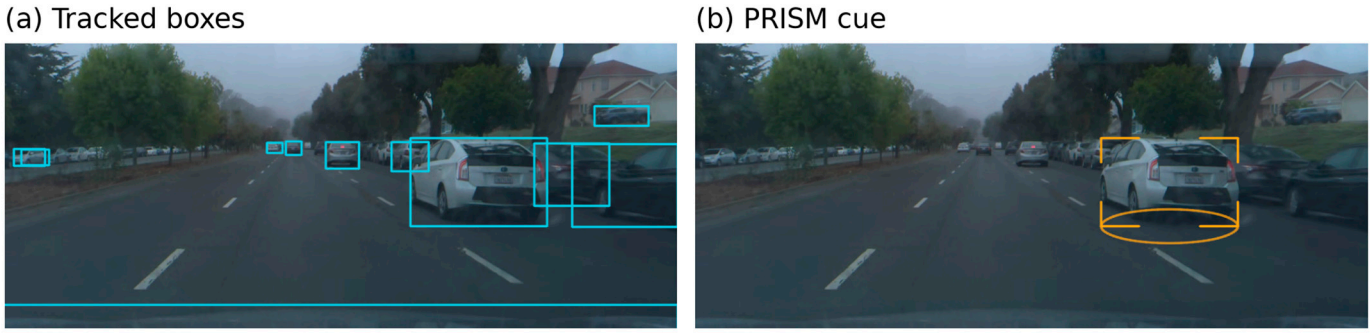


Figure 2. Real-video example at 10.00 s: (**a**) tracked-object boxes and (**b**) the bracket cue selected by PRISM. Panels share a crop for visibility. Source: comma.ai, comma2k19 [38], © 2018 comma.ai, MIT License; boxes and cues drawn by the authors.

For Figure 2, frame 200 contains thirteen tracked objects in the full 1164 × 874 image, from which PRISM selects one nearby-vehicle bracket. Both panels display x = 170–1164 and y = 230–650 pixels to enlarge road detail. Detection and allocation use the full frame, so some tracked boxes fall outside the crop. The demonstration video in the repository provides the complete uncropped sequence.

3.6. Data and Experimental Protocol

Detector benchmarking uses 554 images and 5644 traffic-category annotations selected from COCO val2017 [37]. The selection requires at least three traffic annotations, at least one annotation with normalized area $\geq$0.005, and at least two vehicle or infrastructure annotations. This selection concentrates the detector comparison on populated scenes. Filenames, labels, selection code, validation arguments, model hashes, and console records are supplied in the repository.

Temporal evaluation uses eight source photographs. Annotated patches are assigned depths from the same class-height priors used by the estimator, propagated along prescribed velocities, and composited over warped backgrounds. Each sequence is generated in clear, darkened, fog-like, and rain-like image conditions. There are 645 frames per condition and 2580 in total. Cropping, resizing, patch boundaries, and background inpainting determine the appearance of these constructed sequences. Table 1 records the scene configuration. The designated class in Table 1 sets the sequence duration; Equation (10) defines the instantaneous evaluation reference.

The evaluation reference at each frame is the visible generative object with the largest Equation (10) score under perfect confidence and mature-track evidence. A frame is eligible when this score is at least 0.35. This tests whether estimation and selection preserve the chosen priority rule. The reference mapping uses confidence-ordered one-to-one box matching with a strict IoU > 0.4 threshold and no category gating. Consequently, the reported state results are class-agnostic matched-object measures. The reference is endogenous to the tested priority formulation. Retention measures self-consistency under perception and allocation errors for one reference object per frame.

The detector experiments use eight model variants at 640 pixels, with additional resolution trials retained in the data package. Validation runs use confidence 0.001, NMS IoU 0.7, batch size one, and the eight selected classes. Backend timing uses 60 images after eight warm-up predictions. The recorded platform is a virtualized Intel

Xeon at nominal 2.80 GHz, with two available CPU cores, two PyTorch threads, approximately 7.8 GiB of memory, AVX-512 VNNI, and no discrete GPU. Library versions are Python 3.11.15, PyTorch 2.14.0, Ultralytics 8.4.146, ONNX Runtime 1.25.0, OpenVINO 2026.3.1, OpenCV 4.13.0, and NumPy 2.4.4.

Table 1. Configuration of the eight synthetic source scenes.

Scene	Objects	Ego Speed km/h	Frames	Designated Class
S1	14	19.8	90	bicycle
S2	12	18.0	62	person
S3	14	14.4	64	person
S4	8	90.0	69	car
S5	14	43.2	90	car
S6	5	18.0	90	car
S7	10	21.6	90	bicycle
S8	14	32.4	90	car

Table 2 compares a selected sample of COCO-pretrained models spanning YOLOv5nu, YOLOv8, YOLOv10, and YOLO11. Nano and selected small variants cover modest model sizes, while YOLO11m provides a larger-capacity contrast. All reported rows share the selected classes, input setting, validation subset, and recorded host.

Table 2. Detector accuracy and inference time on the selected COCO subset at 640 pixels.

Model	Parameters M	AP50%	AP50–95%	Inference ms
yolov5nu	2.65	54.53	35.87	94.99
yolov8n	3.16	56.84	38.45	104.67
yolov8s	11.17	66.29	45.99	271.97
yolov10n	2.78	58.73	39.94	96.40
yolov10s	8.13	67.13	46.75	239.62
yolo11n	2.62	58.08	39.31	106.98
yolo11s	9.46	66.96	46.93	237.07
yolo11m	20.11	72.58	52.36	711.91

Inference timing within detector validation excludes preprocessing and postprocessing; backend wall-clock timing covers the prediction call. The two experiments use different samples and timing procedures, so their latency values cannot be subtracted as nested measurements. Backend p95 is the sorted observation at zero-based index floor($0.95(n-1)$); the video summary uses linearly interpolated percentiles. PyTorch minimum-rectangle padding and static square exports can also differ, so backend ratios describe execution paths rather than an isolated precision effect. Neither measures capture-to-display latency.

3.7. Comparative Allocation Protocol and Metrics

Five principal policies share the reconstructed observations, detail-level rules, quality gate, box-overlap limit, and default capacities $B = 0.055$ and $K = 5$. Risk budget ranks all candidates without activation or persistence. Risk + threshold ranks eligible candidates above $r_on = 0.24$. Risk + hysteresis additionally lowers the raw continuation threshold to $max(0, r_on - 0.15)$, without a ranking bonus or dwell preference. Risk + dwell adds an

eight-frame preemptible dwell preference and the 0.06 ranking bonus. PRISM additionally attenuates nonurgent candidates in the central rectangle. The three simpler selectors are implemented separately from PRISM. All retain urgent-candidate precedence and use track identity to break ties. Render-all and confidence, nearest, and TTC rankings are retained as secondary repository results.

The default comparison isolates mechanisms at common settings. A factorial ablation independently enables or disables persistence, central attenuation, and the quality gate, yielding eight combinations. One-factor perturbations use dwell durations {0, 4, 8, 16}, bonuses {0, 0.03, 0.06, 0.12}, central multipliers {0.25, 0.45, 0.75, 1}, and quality thresholds {0.50, 0.60, 0.72, 0.90}. Bonus perturbation holds the effective off-threshold at 0.15, and therefore changes both ranking and raw continuation eligibility; it is not an isolated ranking-effect test.

An exploratory grid evaluates four thresholded families over ten activation thresholds {0, 0.12, 0.18, 0.24, 0.30, 0.36, 0.42, 0.48, 0.54, 0.60}, three budgets {0.0165, 0.033, 0.055}, and three cue limits {1, 3, 5}: 360 configurations, each replayed on all 32 sequences. Raw hysteresis continuation is $max(0, r_on - 0.15)$; the corresponding PRISM effective off threshold adds the 0.06 bonus. The default remains a separately identified operating point. The protocol was recorded before running this grid, after development had already used the scene corpus. No new detector inference or tuning on the real video is performed.

For each of eight source-scene folds, all four variants of one scene are withheld from configuration selection. Each family selects its configuration using the other seven scenes, maximizing J = mean coverage $-\ \lambda$ mean cue count $-\ 0.005$ mean transitions/s, with λ in {0, 0.025, 0.05, 0.10}. Coverage averages eligible sequences; count and transitions include all sequences.

For each condition, H counts eligible score-defined reference frames, D those with a matched reference detection, and S those where that detection is selected. Detection coverage is D/H, conditional display coverage is S/D, and end-to-end reference coverage is S/H. Missing detections remain in the end-to-end denominator. Reference identity may change, and only one reference per frame is scored.

Cue counts and identity additions/removals per second are averaged within a sequence and then across sequences; initial admissions count as transitions. Coverage pools frame counts for the reported tables. The longest omission counts consecutive eligible frames with no displayed reference, even if reference identity changes. A detected-reference bout is a consecutive run of eligible frames with the same recorded matched track identity. Missing detection, ineligibility, or identity change ends the bout. First-display delay is conditional on display within the bout; never-displayed bouts are reported separately. A selected cue below raw 0.24 is called below-activation. Bouts end when detection is missed, confidence filtering or association removes a matched candidate, the chosen reference falls below its eligibility score, or the reference/recorded identity changes. A detected but unshown candidate can fail activation, lose ranking competition, exceed cue or cost capacity, or overlap an admitted box.

Renderer support is measured at native sequence resolution by drawing each admitted cue onto a black canvas, with optional panel and ribbon disabled. The union of nonzero per-cue pixels measures painted area; pixels covered by at least two individual masks measure overlap. Anti-aliased nonzero support is included regardless of perceptual salience.

Paired scene-cluster bootstrap intervals use 10,000 resamples. The original default comparisons use seed 20260911; exploratory fold comparisons use 20260914 and pool each scene's four variants before pairing. Only seven source scenes have eligible references. State fidelity separately reports matched-observation depth error, TTC availability, and error conditional on a finite estimate.

3.8. Real Driving Demonstration and Execution Record

A separate qualitative demonstration uses the official comma2k19 example segment b0c9d2329ad1606b|2018-08-02--08-34-47/40 [38]. The 60-s recording contains 1200 frames at 1164 × 874 pixels and shows traffic, wet-road reflections, and windshield-wiper motion. It was selected for accessible source data, camera metadata, and visible traffic before running PRISM. No detector training, threshold tuning, or manual replacement of detections was performed for this sequence. This identifier denotes one source driving segment, not a benchmark split or experimental parameter; the pinned source and checksum are listed in the repository. The segment provides a concrete integration case for the recorded pipeline. Algorithm 1 makes the per-frame application explicit.

The supplied camera matrix has $f_x = f_\gamma = 910$ pixels and principal point (582, 437), corresponding to a 65.20° horizontal field of view. Actual frame timestamps, with median spacing of 0.049999 s, drive temporal estimation. Playback is 20 FPS; the elementary HEVC stream's default 25 FPS interpretation would shorten the recording incorrectly. The full image is processed without an ego-vehicle exclusion mask. Minimum-rectangle padding is used by the PyTorch prediction path; 640 denotes the nominal inference size.

Algorithm 1. Frame-wise PRISM allocation

1. Read current tracked boxes, identities, categories, confidence, age, and acquisition timestamp. Expire state histories absent for more than one second; update reciprocal-height, confidence, and IoU histories.
2. Estimate Z, X, TTC and fit quality; compute r and u. For each candidate, recover its previous displayed state. Add the 0.06 bonus only for a previously shown identity when persistence is enabled.
3. Test effective score against 0.24 for new identities or 0.15 for previously shown identities. Allow the preemptible eight-frame dwell preference to retain eligibility. Assign full, bracket or tick detail using raw r and the quality gate.
4. Multiply effective ranking score by 0.45 for nonurgent boxes overlapping the central rectangle by IoU > 0.10. Sort by raw urgency first, dwell preference second, decreasing effective score third, and identity last.
5. Traverse the ordered candidates. Admit a cue only if its cost fits the remaining B = 0.055 budget, fewer than K cues are selected, and IoU with every admitted box is ≤ 0.55. Skip infeasible candidates. Urgency and dwell never override these constraints.
6. Render admitted cues at their assigned detail. Replace the displayed-identity map with the selected set, preserving each continuously displayed identity's start frame. Archive states, selections and stage timings.

The pipeline runs YOLO11n and ByteTrack, shared state reconstruction, scoring, and PRISM allocation on each frame. Risk-budget ranking receives identical reconstructed states. The all-tracked-object video panel shows detector/tracker box outlines and is distinct from the full-cue render-all policy in the synthetic experiment. Per-frame observations, policy decisions, and three timing fields are retained. Saved timings describe an Apple M5 host with 16 GiB memory, macOS 26.6.2, Python 3.12.14, PyTorch 2.14.0, Ultralytics 8.4.146, OpenCV 5.0.0, NumPy 2.5.3, two PyTorch threads, and explicit CPU execution. The host specifications were captured separately from the run.

3.9. Analytic Geometry Stress Fixtures

Controlled fixtures decouple physical object height from the estimator's person prior of 1.70 m. True heights are 1.20, 1.50, 1.70, and 2.00 m, with a 960 × 720 pinhole camera and

60° field of view. Each trajectory has 100 samples at 20 Hz, initial depth 50 m, and initial closing speed 5 m/s. Apparent height is projected from the independently specified physical height and trajectory. The reference TTC is instantaneous depth divided by closing speed.

Eight conditions comprise noiseless constant speed; Gaussian box-height noise with standard deviation 2 + 0.03h pixels; closing acceleration of 1.5 m/s^2; independent 30% observation dropout; observed height reduced to 65% on frames 30–59; identity replacement every ten frames; receding motion at 2 m/s with noise; and stationary depth with noise. Fifty seeded repeats per height and condition use seed 20260914 plus replicate index. Noiseless repeats are identical and are not counted as independent evidence. Availability denominators include startup observations but exclude dropped observations. Errors include all finite closing estimates, without clipping large errors; nonclosing conditions report the fraction spuriously estimated as closing.

3.10. Propagation of Geometry Error into Selection

A paired diagnostic propagates the trajectories through the unchanged priority function and allocator. One focal person candidate uses either true depth and instantaneous TTC or their archived estimates. Five fixed person competitors have (Z in m, TTC in s) pairs: (12, 2.5), (20, 4), (28, 5.5), (36, ∞), and (44, ∞). Both branches use confidence 0.9, age 10, lateral offset X = 0, evidence-quality index u = 0.2, and identical disjoint boxes. Competitor i = 0,. . .,4 has box (20 + 180i, 100, 60 + 180i, 180); the focal box is (450, 400, 490, 480), in 960 × 720 pixel coordinates. These constructed, disjoint boxes isolate score propagation from overlap and localization effects.

Each branch maintains its own persistent selection history. The cost budget is 0.055, with separate cue-count conditions K = 1 and K = 5; all remaining settings retain their defaults. The comparison reports a changed raw-score ordering, a changed set of displayed identities, focal extra admission, and focal omission relative to the true-geometry branch. A set comparison ignores within-set order and representation changes.

3.11. Real-Video Geometry Sensitivity

A deterministic replay tests how prescribed geometric changes affect allocation on the 1200 recorded real-video frames. The unmodified branch recomputes risk and exactly reproduces all saved selected identities and detail levels. Separate branches multiply every candidate's depth and lateral offset together by 0.8 or 1.2, multiply finite TTC by 0.8 or 1.2, apply both changes together, or mark all TTC values unavailable. Scaling Z and X together represents a common physical-height scale change while preserving the observed bearing. Finite-TTC scaling is a separate input stress, not a consequence inferred from constant height scaling.

Boxes, classes, confidence, identities, ages, and quality indicators remain at their recorded values; each branch starts empty and evolves its own persistent allocation history with default capacities. The analysis reports disagreement in full raw-score ordering and selected-identity sets relative to the recorded branch, plus mean cue count. All 1200 frames remain in the denominator, including empty selections. The ±20% factors probe common geometric scaling across candidates while reusing the recorded perception outputs and quality indicators. The repository provides the script, input checksum, and paired decisions.

3.12. Use of Generative AI Tools

Two generative AI tools were used in this study: Claude (Opus 5.5, Anthropic) and Codex (GPT6-Astra, OpenAI). They were used for three tasks.

Software. The tools assisted in writing and refactoring parts of the PRISM implementation, the experiment, replay, and figure-generation scripts, and the software tests. The authors specified all algorithms, parameters, and experimental designs. The authors reviewed the

generated code; its behavior is checked by software tests and by the deterministic replay of the recorded decisions (Section 4.4), and the code is public in the repository.

Manuscript editing. The tools assisted with language editing, restructuring, and consistency checks of the text. They did not produce data or results: every reported number is computed by the published scripts from the archived records. The authors checked all references against the original publications.

Machine-assisted visual labelling. In the real-video audit (Section 4.6), the six sampled observations flagged by the geometric screen were classified with the image-understanding capability of Codex, which identified each as the recording vehicle's own body.

No figure content was generated by an AI image model; all figures are rendered by the repository scripts from recorded data or, drawn over dataset frames.

4. Results

4.1. Detector and Backend Observations

YOLO11n achieves 39.31% AP50–95 with 2.62 million parameters. YOLOv10n achieves 39.94% AP50–95 and 96.40 ms inference latency, compared with 106.98 ms for YOLO11n. The downstream cache and demonstration use YOLO11n as a fixed perception component, allowing every allocator to receive identical inputs. Evaluating detector choice at the interface level additionally requires rerunning detection and association and measuring the resulting changes in candidates, state, and allocation.

The table makes the size–accuracy–latency trade-off explicit. Relative to YOLO11n, YOLO11s increases AP50–95 by 7.62 percentage points, parameters from 2.62 to 9.46 million, and inference time from 106.98 to 237.07 ms. YOLO11m reaches 52.36% AP50–95 at 20.11 million parameters and 711.91 ms. Conversely, YOLOv10n gains 0.63 percentage points over YOLO11n and saves 10.58 ms, with 2.78 rather than 2.62 million parameters. These operating points make model selection dependent on the application's accuracy and execution budget.

OpenVINO INT8 prediction averages 35.77 ms versus 96.55 ms for PyTorch FP32, a ratio of approximately 2.70 (see Table 3). This compares execution paths, including possible padding differences. Overlapping calibration and evaluation makes INT8 accuracy exploratory.

Table 3. YOLO11n backend prediction time and accuracy at 640 pixels.

Backend	Mean ms	p95 ms	AP50–95%
PyTorch FP32	96.55	109.24	39.31
ONNX Runtime FP32	72.36	82.95	38.97
OpenVINO FP32	57.56	65.42	38.97
OpenVINO INT8	35.77	43.04	38.67

4.2. Default Allocation and Painted Area

Thresholding reduces mean cues from 4.55 to 1.73 but also reduces displayed references from 218 to 174. Adding eligibility hysteresis restores 213 displayed reference frames with 2.03 cues. Risk + dwell and PRISM both show 213 references and 2.06 cues. Dwell ranking has fewer transitions than PRISM (3.24 versus 3.46 per second), whereas PRISM paints slightly less image area (0.846% versus 0.868%). Table 4 reports sequence means for count, transitions, and painted area, and pooled frame coverage. All policies share capacities and presentation rules.

Table 4. Clear-scene allocation results (H = 325; D = 227).

Policy	Cues	Transitions/s	Shown S	S/H %	Painted %
Risk budget	4.55	18.80	218	67.1	1.210
Risk + threshold	1.73	4.97	174	53.5	0.814
Risk + hysteresis	2.03	3.68	213	65.5	0.867
Risk + dwell	2.06	3.24	213	65.5	0.868
PRISM	2.06	3.46	213	65.5	0.846

4.3. Operating Curves and Scene-Grouped Selection

Figure 3a,b connects ten activation thresholds per family in threshold order, using B = 0.055 and K = 5 on clear sequences. Panels (c,d) aggregate all conditions after source-scene-grouped configuration selection at each of the four utility preferences defined in Section 3.7.

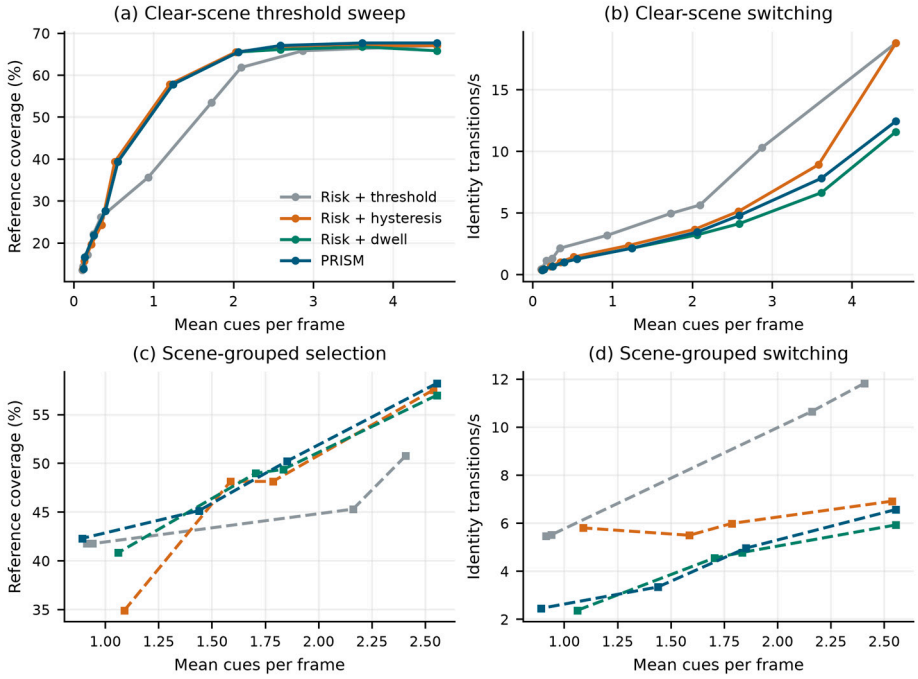


Figure 3. Allocation trade-offs: (**a**,**b**) activation-threshold sweeps and (**c**,**d**) outcomes after source-scene-grouped configuration selection.

Against risk budget, PRISM's mean painted area is reduced by approximately 30.1%, substantially less than its cue-count reduction. Against risk + dwell, the reduction is approximately 2.5%, with a switching cost. The 95th percentile of frame occupancy is 2.474% for PRISM and 2.507% for risk budget. PRISM's clear-scene mean overlap is 0.0097% of image pixels.

Overall, across the four conditions, PRISM displays 717 of 1300 eligible references (55.2%); risk + hysteresis displays 712, risk + dwell 710, and risk budget 763. Clear, darkened, fog-like, and rain-like PRISM coverage is 65.5%, 54.5%, 54.2%, and 46.5%, respectively. In clear replay, 98 reference frames are missed by perception, and 14 detected references are omitted by selection. The shared geometric and score-defined reference limits the interpretation of every coverage value.

Configuration selection changes the comparison. At $\lambda = 0$, PRISM retains 58.2% with 2.55 cues, compared with 57.0% and 2.55 cues for risk + dwell, but PRISM has more transitions (6.57 versus 5.93 per second). At $\lambda = 0.05$, PRISM retains 45.2% with 1.44 cues versus 49.0% with 1.70 cues for risk + dwell. At $\lambda = 0.10$, the corresponding values are

42.3% with 0.89 cues and 40.8% with 1.06 cues, while dwell again switches slightly less. For $\lambda = 0.05$, the scene-average PRISM minus dwell coverage difference is -2.17 percentage points, with an exploratory paired bootstrap interval of -5.51 to $+0.27$ points. At $\lambda = 0.10$, the difference is +1.13 points with an interval of +0.26 to +2.20. These are scene-averaged differences; Table 5 reports pooled frame coverage, so its differences are not identical. Table 5 pools 1300 eligible frames after selecting each configuration on the other seven source scenes, with corruption variants grouped by origin.

Table 5. Scene-grouped results: coverage (%)/mean cues/transitions per second.

Policy	λ = 0	λ = 0.025	λ = 0.05	λ = 0.10
Risk + threshold	50.8/2.41/11.83	45.3/2.16/10.66	41.8/0.94/5.52	41.8/0.91/5.47
Risk + hysteresis	57.5/2.54/6.92	48.2/1.78/5.99	48.2/1.59/5.50	34.9/1.09/5.81
Risk + dwell	57.0/2.55/5.93	49.4/1.83/4.79	49.0/1.70/4.56	40.8/1.06/2.38
PRISM	58.2/2.55/6.57	50.2/1.85/4.97	45.2/1.44/3.35	42.3/0.89/2.46

4.4. Temporal Omissions and Mechanism Sensitivity

Figure 4a uses the native-resolution union of nonzero cue pixels, excluding status panels and ribbons, as defined in Section 3.7. Panel (b) shows the sequence with the longest default-PRISM eligible omission across all conditions; selection is mechanical, with alphabetical tie-breaking.

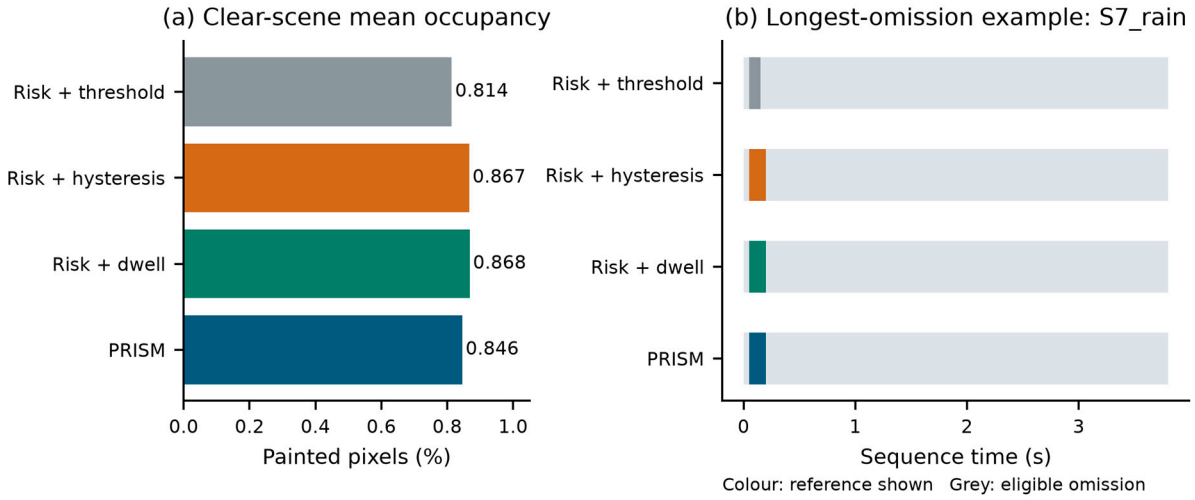


Figure 4. (**a**) Mean painted cue area in clear scenes and (**b**) the longest-omission example. Color indicates reference display, grey eligible omission, and blank intervals ineligibility.

Default PRISM has a longest consecutive eligible omission of 2.10 s in clear sequences and 3.60 s over all conditions. Across all conditions, 26 of 68 detected-reference bouts are never displayed, compared with 26 for simple hysteresis, 27 for dwell ranking, and 12 for risk budget. These bouts are defined by recorded matched track identities and can be fragmented by perception. In clear replay, PRISM displays 11 of 16 bouts. Below-activation cues constitute 18.6% of PRISM's clear-scene cue-frames.

Every persistence-enabled factorial variant displays 213 clear-scene references, versus 174 without persistence. The quality gate changes neither retention nor count at these settings. The gate changes how a cue is drawn, not whether it is selected. Any benefit of that representation must be measured with drivers in future work. Central attenuation changes switching and painted area, but does not increase clear-scene retention. With the

other defaults fixed, changing dwell from 0 to 16 frames leaves S = 213 while mean count increases from 2.03 to 2.10. Changing the bonus from 0 to 0.12 changes displayed references from 200 to 214 and count from 1.96 to 2.19. In Table 6, persistence comprises eligibility hysteresis, ranking bonus, and dwell; the gate controls the quality-based detail rule.

Table 6. Clear-scene factorial ablation of persistence, central attenuation, and quality gating.

Persistence	Center	Gate	Cues	Shown S	Transitions/s
Off	Off	Off	1.73	174	4.97
Off	Off	On	1.73	174	4.97
Off	On	Off	1.73	174	5.36
Off	On	On	1.73	174	5.36
On	Off	Off	2.06	213	3.24
On	Off	On	2.06	213	3.24
On	On	Off	2.06	213	3.46
On	On	On	2.06	213	3.46

The default implementation remains fixed for reproducibility. The simpler dwell policy is a parsimonious comparator for the measured clear-scene retention and stability outcomes; replacing the default globally would require independent data because central attenuation has a different painted-area trade-off and fold results vary. All 22 software tests pass, including capacity, preemption, timestamps, shared state, comparator behavior, and temporal metric fixtures. Replaying all 2580 synthetic frames confirms that the extended analysis preserves the original PRISM identities and detail levels.

4.5. State Fidelity and Independent Analytic Perturbations

The composited sequences yield 7.51% median depth error and finite TTC estimates for 840 of 1449 eligible matched observations (58.0%), with conditional median absolute TTC error 0.455 s. These figures describe shared-prior reconstruction and omit unmatched objects from state-error denominators. Table 7 defines TTC eligibility by finite generative TTC in (0, 4] s and conditions error on a finite estimate.

Table 7. State-estimation accuracy and TTC availability in matched synthetic observations.

Condition	Matched	Depth MdAPE %	Finite/Eligible TTC	TTC MdAE s
Pooled	12,222	7.51	840/1449	0.455
Clear	3349	6.88	226/378	0.451
Darkened	2989	7.86	202/352	0.481
Fog-like	2991	7.97	207/363	0.406
Rain-like	2893	7.49	205/356	0.462

Figure 5 uses physical heights specified independently of the 1.70 m estimator prior. Panel (a) concerns noiseless constant-speed projection. Availability includes startup and excludes dropped observations; finite-error percentiles retain large errors. The fixtures start at 50 m and 5 m/s closing speed, unlike the TTC $\leq$ 4 s population in Table 7.

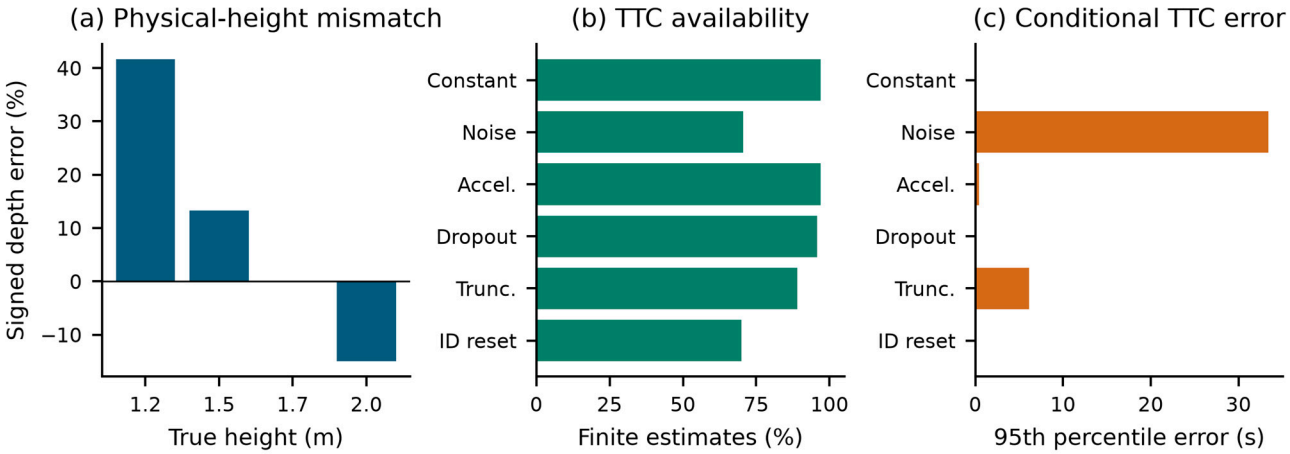


Figure 5. Analytic geometry stress tests: (**a**) physical-height-induced depth bias, (**b**) finite TTC availability, and (**c**) conditional 95th-percentile absolute TTC error.

Changing physical height to 1.20 m produces +41.7% depth bias, whereas 2.00 m produces −15.0%, despite exact image localization. Noiseless constant-speed TTC remains accurate to numerical precision because constant physical scale cancels in reciprocal-height fitting. With injected height noise, availability falls to 70.5%, median finite error is 4.10 s and the 95th percentile is 33.41 s. The noiseless truncation fixture has near-zero median TTC error but a 6.16 s 95th percentile, exposing transition errors concealed by the median. Identity replacement every ten frames reduces availability to 70.0% through repeated startup. Table 8 pools four physical heights and fifty seeds; finite-estimate errors are measured against instantaneous depth divided by closing speed.

Table 8. TTC accuracy and availability under analytic stress conditions.

Condition	Available %	Median Error s	p95 Error s
constant	97.0	$<1 \times 10^{-10}$	$<1 \times 10^{-10}$
noise	70.5	4.104	33.408
acceleration	97.0	0.131	0.437
dropout	95.7	$<1 \times 10^{-10}$	$<1 \times 10^{-10}$
truncation	89.0	$<1 \times 10^{-10}$	6.163
identity reset	70.0	$<1 \times 10^{-10}$	$<1 \times 10^{-10}$

Noisy receding and stationary fixtures produce finite positive closing estimates on 44.0% and 48.5% of observations, respectively. These estimator-level results identify the need to combine fitted slope and evidence quality with physical motion-validity checks.

4.6. Real-Video Integration and Visible Failures

PRISM selects 0.78 cues from 6.62 tracked objects per frame on average, compared with 4.66 for risk budget. The sequence has no independently labelled relevance reference; 509 frames have no PRISM cue. The logged state/scoring and allocation intervals together average 0.117 ms, approximately 0.4% of detection plus tracking time. Decoding, drawing, encoding, and physical display are excluded. The logged intervals locate computational cost within the pipeline; acquisition-to-display performance is a target of the deployment evaluation. Table 9 uses all 1200 frames for counts and frames 20–1199 for timings on the Apple M5 CPU.

Table 9. Cue activity and component timings in the 60 s real-video demonstration.

Measure	Observed Value
Tracked objects, mean/frame	6.62
Risk-budget cues, mean/frame	4.66
PRISM cues, mean/frame	0.78
PRISM frames without cues	509/1200 (42.4%)
PRISM/risk budget identity transitions/s	2.72/23.32
Detection + tracking, mean/p95 ms	28.903/32.633
State + scoring, mean/p95 ms	0.104/0.145
Cue allocation, mean/p95 ms	0.013/0.019
Combined logged stages, mean/p95 ms	29.020/32.764

At frame 405 (20.25 s), the detector labels the recording vehicle's lower-image body/interior as a car with confidence 0.483. The class-height model estimates 5.51 m and the heuristic score reaches 0.244, just above activation, so PRISM displays a tick. Figure 6 preserves this observed failure. At frame zero, a similar false detection has confidence 0.422 but score 0.163 and is not selected. Selection can suppress or propagate the same type of semantic error depending on temporal and priority state. Figure 6 shows the full frame and reproduces the saved detector/tracker boxes and selected base ellipse.



Figure 6. Recording-vehicle false detection at 20.25 s: (**a**) tracked boxes and (**b**) selected PRISM cue. Source: comma.ai, comma2k19 [38], © 2018 comma.ai, MIT License; boxes and cues drawn by the authors.

A fixed one-second inspection schedule samples frames 0, 20, . . ., 1180. A geometric screen flags boxes wider than 75% of the image, starting below 55% of image height and extending below 95%. It produces six sampled candidate observations; machine-assisted visual inspection identifies all six as the visible recording-vehicle region, and none is selected at those sampled instants. This sparse audit misses the selected failure at frame 405 and illustrates the limitations of periodic screenshots. The machine-assisted labels and geometric screen describe this specific recording-vehicle failure mode.

4.7. Geometry-to-Selection Propagation

Table 10 distinguishes score-order changes from final selection changes. Under constant-speed projection, differences in physical height change the raw ranking in 48.5% of observations but leave the selected set unchanged at K = 1 and change it in 10.5% at K = 5.

With box-height noise, ranking changes in 70.6%; displayed identities change in 12.3% at K = 1 and 64.3% at K = 5. The five-slot condition admits a low-priority focal candidate more readily and carries that difference through persistence. These outcomes depend on the specified competitors, threshold crossings, capacities, and selection histories. Table 10 compares estimated-geometry and true-geometry branches with other candidate attributes fixed. N counts observations per capacity condition, pooled over four heights and 50 seeds; noiseless repeats reuse identical trajectories.

Table 10. Geometry-induced changes in priority order and selected identities at K = 1 and K = 5.

Condition	N	Rank Changed %	Set Changed % K = 1	Set Changed % K = 5
constant	20,000	48.5	0.0	10.5
noise	20,000	70.6	12.3	64.3
acceleration	20,000	17.0	1.0	2.0
dropout	14,028	48.7	0.0	10.5
truncation	20,000	64.0	8.0	26.5
identity reset	20,000	48.5	0.0	10.5
receding	20,000	30.0	7.3	54.0
stationary	20,000	39.6	7.7	55.9

Large numerical TTC errors need not affect the selected set. In the noise condition, 1718 observations have finite absolute TTC error greater than 10 s. None change the selected set at K = 1, whereas 970/1718 (56.5%) do so at K = 5. The clipped urgency score, distance contribution, and competitor ordering can absorb some errors, while activation and persistence can propagate others. Noisy receding and stationary trajectories still alter the selected set in 54.0% and 55.9% at K = 5. The paired branches quantify how geometric differences propagate through the configured allocation policy.

4.8. Sensitivity of Real-Video Cue Decisions

Table 11 applies the Section 3.11 probes to the actual recorded candidate sequences. Scaling depth and lateral offset by 0.8 changes selected identities on 61.8% of frames; scaling by 1.2 changes them on 19.8%. Finite-TTC scaling alone changes selections on 26.0% and 19.4%, respectively. Joint scaling changes 76.6% or 30.5%, and removing the temporal term changes 51.1%.

Table 11. Real-video selection sensitivity relative to the recorded geometric estimates.

Input Condition	Rank Changed %	Set Changed %	Mean Cues
Recorded inputs	0.0	0.0	0.780
Z and X × 0.8	60.9	61.8	1.630
Z and X × 1.2	57.0	19.8	0.572
Finite TTC × 0.8	43.8	26.0	1.046
Finite TTC × 1.2	38.9	19.4	0.578
Both × 0.8	65.5	76.6	1.957
Both × 1.2	63.0	30.5	0.435
TTC unavailable	85.7	51.1	0.132

The real-video replay shows that geometric inputs materially influence cue selection through threshold crossings and persistent allocation histories. Disagreement is measured relative to recorded estimates, including the false cue documented in Figure 6.

5. Discussion

5.1. What the Allocation Comparison Establishes

The principal finding is that the display benefit comes from two simple mechanisms: activation thresholding reduces cue count and switching, while eligibility hysteresis and dwell restore the reference retention that thresholding loses. Comparing PRISM only with risk budget would attribute a 54.7% count reduction to the complete framework even though simpler persistence reproduces its clear-scene retention and count. Adding central attenuation trades a small reduction in painted area against more switching at that operating point. The 360-configuration exploration further shows that the preferred family depends on the selected cost of display activity. The utility weights specify exploratory preferences for comparing operating points. An application that prioritizes temporal stability may favor simple dwell ranking; another could value the measured area reduction. Independent sequences and an application-specific omission cost are required before choosing a deployment policy.

PRISM integrates capacity-constrained cue allocation with preemptible persistence, detail rules, and central attenuation, supported by replayable decisions and failure analysis. Against risk + dwell, its additional effect in clear scenes is a 0.022 percentage-point reduction in painted image area, approximately 2.5% relative, at the cost of 0.22 more identity transitions/s and unchanged reference retention.

5.2. Relationship to Existing Solutions

YOLO and RT-DETR address perception accuracy and execution [13–15,28]; ByteTrack addresses association [25]. They are potential upstream components rather than competing cue allocators. DAARIA already integrates perception and AR assistance [1], and participant studies examine hazard perception [2], workload [3], and display opacity [4]. PRISM addresses the step between perception and presentation: which observations occupy a limited display, and for how long. PRISM contributes inspectable software decisions and a controlled comparison of allocation mechanisms. Its comparison therefore centers on allocation mechanisms and their display consequences.

5.3. Validity and Limitations

The most important limitation concerns the evaluation reference. The main allocation reference is built from the same priority rule (Equation (10)) that the method uses. The comparisons measure how well that rule survives perception and allocation error. Cue count, identity transitions, painted area and overlap, and execution time do not depend on a score-derived label. They describe display activity and computational cost. The real-video perturbation in Section 4.8 has the same property: it measures sensitivity, not correctness. An independent test of effectiveness would need object and track annotations, together with encounter-level relevance judgements made without reference to Equation (10), ideally by several annotators with reported agreement. Geometric ground truth alone would not be enough, because it would validate range and motion estimates but not whether a cue matters to the driver.

The simulated errors are simpler than real detector errors. In the depth experiment, the same class-height assumptions are used both to generate the data and to estimate range. The physical-height stress fixtures break this equality but still rely on analytic projection. The pixel noise, with a scale of 2 + 0.03h pixels, depends only on apparent height. Small-

object studies [26,27] cannot show how large they are or which way they push. Likewise, the ±20% geometric replay factors are prescribed common perturbations, not empirical error quantiles or independent per-object draws. Section 4.7 isolates how geometric error propagates into selection while all other variables are held fixed. Because matching is class-agnostic, a misclassified detection can still count as a match, so an independent evaluation should use category-aware matching and scenes with several relevant objects. The joint error distribution of detection and geometric estimation has yet to be measured. The scoring weights, including the class weights v(c), were held fixed; calibrating them requires independent relevance labels, which could be the subject of future work.

The data limit external validity. The temporal corpus comes from eight source photographs. These eight independent visual origins are expanded into 32 corruption variants and 2580 correlated frames. The 554-image detector subset addresses a separate, static task, and the COCO selection favors populated images. Grouping variants by source prevents leakage between variants within a fold, but prior developmental reuse of the sources, overlapping training folds, correlated frames and exploratory multiple comparisons remain. Results also change materially with the utility weight. The bootstrap intervals should therefore be read as descriptive: the study supports no confirmatory claim of superiority, equivalence or non-inferiority. A broader evaluation should cover different routes, camera installations, day and night, weather, and encounters with vulnerable road users, with sampling and validation organized by independent journey.

The presentation layer is a prototype that draws overlays on camera images. The painted area depends on the renderer and does not measure salience, distraction, or comprehension. A low-cost ellipse can still cover a large region, and non-overlapping boxes do not guarantee non-overlapping text. The interface design draws on literature. Driver interviews, participatory design and iterative user testing were outside the scope of this study.

Finally, all timing measurements reflect software execution on the research hosts. Automotive hardware and driver outcomes are left to the validation program in Section 5.4. The real-video geometric screen and audit have one further limitation. The audit relies on machine-assisted labels that have not been independently checked by humans. Its sparse sample documents a single, narrowly observed failure mode and gives no denominator for detector precision or recall.

5.4. Validation Priorities

The next evaluation should freeze the allocation policy and use independently annotated driving encounters with geometry and relevance defined outside equation (10). A corpus such as nuScenes supplies metric scene information [39], while CARLA can support controlled trajectories [40]; relevance and human usefulness still require explicit operational definitions. Matched comparators should retain comparable activation and persistence rules, with utility preferences fixed before testing. Event-level omission, switching delay, false-cue duration, multiple simultaneous objects, and budget saturation should supplement frame averages. Geometry improvements should address apparent-height changes from truncation, pose and camera motion, and explicitly reject unsupported closing estimates. A calibrated own-vehicle mask and geometric plausibility checks should be validated against nearby real objects before use. Simply removing the bottom image region risks hiding relevant observations. INT8 evaluation needs separate calibration and test manifests; the complete pipeline needs repeated capture-to-display measurements on its intended device. That measurement must also fix the deployment partition, because the cost of detection depends on where inference runs: compressed models on near-edge processors [16], cooperative vehicle–cloud execution [17], and dedicated accelerators feeding a display device [36] occupy different operating points, and

none is interchangeable with a server-CPU figure. Capture-to-display latency, rather than detector latency alone, is the quantity a presentation layer depends on. A subsequent controlled simulator study should measure a predefined driver endpoint, such as missed hazards or response time, with counterbalanced conditions and justified sample size. Detection-response workload measures [3], gaze, NASA-TLX [41], and appropriate reliance [6] can complement that endpoint.

A broader road evaluation should select independent journeys before inspecting PRISM outputs and report route, lighting, weather, road type, camera mounting, and encounter composition. Development and test splits should be separated by journey and location. Synchronous metric sensing can support range/relative-motion labels, while separately specified multi-object relevance annotation is needed for cue selection.

6. Conclusions

PRISM supplies a reproducible software framework for bounded, temporally persistent cue allocation. In clear replay, it reduces simultaneous cues by 54.7%, switching by 81.6%, and painted area by 30.1% relative to unthresholded risk ranking (risk budget). It displays 213 of 325 eligible score-defined reference frames (65.5%), including 213 of 227 detected references (93.8%). These denominators quantify consistency with the specified heuristic reference. While we acknowledge they characterize software behavior and rule consistency, they do not establish real-world effectiveness, driver awareness or safety benefits. The recorded state and allocation stages average 0.117 ms per frame on the demonstration host, excluding rendering and capture-to-display costs.

Controlled comparison also identifies which mechanism produces that result. Activation thresholding accounts for most of the reduction in cue count and switching, and persistence restores the retention that thresholding loses, without the more elaborate ranking and gating rules: in clear sequences, PRISM and simpler dwell ranking both retain 65.5% of score-defined references with 2.06 cues, PRISM painting slightly less area and switching more. Scene-grouped exploration over 360 configurations maps preference-dependent trade-offs rather than a consistent advantage for any one family. The framework and mechanism-level evidence support explicit choices among retention, display area, and temporal stability.

Analytic perturbations expose class-height bias and sensitivity to noisy or discontinuous observations. The paired diagnostic changes displayed identities in 12.3% of noisy observations at $K = 1$ and 64.3% at $K = 5$, under the specified competition and fixed evidence quality. The real-video recording demonstrates integration and false-cue propagation, while replay with prescribed geometric perturbations quantifies selection sensitivity on recorded traffic observations. These findings connect estimation behavior to presentation decisions and guide further development. The next validation stage combines independently annotated driving encounters, a controlled participant study, and repeated capture-to-display measurements on the intended platform to evaluate relevance, driver benefit, and deployment performance.

Author Contributions: Conceptualization, M.D. and T.G.; methodology, M.D.; software, M.D. and T.G.; validation, T.G.; formal analysis, M.D.; investigation, M.D. and T.G.; data curation, T.G.; writing—original draft preparation, M.D.; writing—review and editing, M.D. and T.G.; visualization, M.D.; supervision, M.D.; project administration, M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the iDrive Project (https://idriveproject.ro/ accessed on 19 October 2026) through a grant of the Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project number PN-IV-P2-2.1-TE-2023-1434, within PNCDI IV.

Institutional Review Board Statement: Not applicable. This study analyses software outputs and publicly released images/video. No participants were recruited, and no driver behavior or identifiable participant outcomes were measured.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting this study are available in the public GitHub repository https://github.com/dugu/prism. Synthetic experiment records and replay results are in data/experiment and results; real-video tracks, states, policy decisions, summaries, timestamps, and verification scripts are in demos/real driving. Demonstration videos are provided in the real-driving-v1 GitHub release. Original COCO images, the comma2k19 source clip, and public pretrained weights are identified by pinned source locations and checksums; download instructions and applicable third-party notices are included. The supplied derived model exports are also retained in the repository. Extended-study data and protocol are presented in results/extended; the comparison, occupancy, stress-test, and audit scripts are placed in scripts. Publication figures are generated from these records without modifying recorded observations. The additional real-video sensitivity records are in results/, with the executable script in scripts/. The repository also provides the demonstration videos, paired geometry-to-selection decisions, input checksums, and replay instructions. Figures 2 and 6 reproduce the archived video decisions.

Acknowledgments: Claude (Anthropic) and Codex (OpenAI) assisted with manuscript editing, code, and machine-assisted visual labelling. The narrow video audit has not been independently human-validated. The human authors are responsible for verifying scientific claims, references, implementation, and the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Fixed Parameters and Their Interpretation

Table A1. Perception, timing, and numerical settings.

Parameter	Value and Role
Detector and tracker	YOLO11n; nominal 640 px; confidence ≥ 0.25; bytetrack.yaml. Tracker high/new 0.25, low 0.10, buffer 30 frames, match 0.8. Upstream filtering removes the nominal low-score recovery band.
Temporal windows	TTC: at most 9 observations, minimum 4. Quality: last 8 confidence and consecutive-box IoUs. History expires after > 1 s absence. Synthetic rate 20 Hz; real acquisition timestamps used.
Camera	Synthetic 960 × 720, 60° horizontal FOV. Real 1164 × 874, $f_x = f_\gamma = 910$ px, principal point (582,437), ≈65.20° horizontal FOV.
TTC numerical guards	$\Sigma(t - \text{mean } t)^2 > 10^{-9}$ s^2; slope b < -10^{-12} pixel^{-1} s^{-1}; fitted latest reciprocal height > 0. If reciprocal-height total squared deviation $\leq 10^{-12}$ pixel^{-2}, R^2 defaults to 0.
Metric definitions	IoU = intersection area/union area. NMS = nonmaximum suppression. AP50–95 = average precision over IoU thresholds 0.50:0.05:0.95. p95 = 95th percentile.

Table A2. Assumed category heights, dispersions and priority weights.

Category	H (m)	σH (m)	v(c)
Person	1.70	0.10	1.00
Bicycle	1.70	0.12	0.95
Motorcycle	1.60	0.15	0.90
Car	1.50	0.15	0.62

Table A2. *Cont.*

Category	H (m)	σH (m)	v(c)
Bus	3.20	0.35	0.62
Truck	3.40	0.55	0.66
Traffic light	0.90	0.20	0.30
Stop sign	0.75	0.10	0.30

Table A3. Priority and evidence-quality constants.

Component	Definition and Interpretation
Temporal and distance urgency	TTC contribution saturates at 1.5 s and is zero at $\geq$ 6 s; denominator 4.5 s. Distance contribution saturates at 8 m and is zero at $\geq$ 45 m; denominator 37 m. Distance multiplier 0.55.
Corridor C	Half-width 1.6 m + 0.02Z; Gaussian soft margin 1.5 m. The score multiplier 0.35 + 0.65C leaves residual priority outside the corridor.
Evidence E	Confidence factor 0.45 + 0.55p; age factor 0.55 + 0.15 min(age,3), clipped to [0, 1].
Quality u	Weighted clipped components: 0.22(1 − *p*); 0.14 confidence variability; 0.16 box instability; 0.18 relative range dispersion; 0.18 TTC fit deficit; 0.12 immaturity.
Quality normalization	Sample confidence SD/0.20; (0.92 − mean consecutive IoU)/0.35; (σZ/Z)/0.30; 1 − R^2; (5 − min(age,5))/5. Every component is clipped to [0, 1].
Missing evidence	Confidence variability = 0.5 for < 3 observations; box instability = 0.5 without IoU history; TTC-fit deficit = 0.85 when TTC is unavailable.

Table A4. Allocation and rendering controls.

Control	Setting and Interpretation
Capacities	B = 0.055 surrogate units; K = 5 cues. Full/bracket/tick costs 0.0165/0.0072/0.0022.
Activation and continuation	New raw r $\geq$ 0.24. Previous effective min(1, r + 0.06) $\geq$ 0.15, equivalent to raw r $\geq$ 0.09 outside dwell.
Persistence	Eight-frame preemptible dwell (0.4 s at 20 Hz); bonus 0.06.
Detail and quality gate	Full at r $\geq$ 0.55; bracket at r $\geq$ 0.30; tick otherwise. If u > 0.72 and r < 0.60, use tick. Raw r $\geq$ 0.55 defines urgent ordering.
Central attenuation	Normalized rectangle (0.34, 0.40, 0.66, 0.78); candidate IoU > 0.10 and r < 0.55 triggers multiplier 0.45.
Overlap and feasibility	Reject a box if IoU > 0.55 with any selected box. Cost comparison tolerance 10^{-12}. Greedy ordering: urgent, dwell, decreasing attenuated effective score, then ascending identity.
Rationale label	Corridor factor > 0.7: “in corridor”; otherwise |X| < 4 m: “near corridor”; otherwise “off path”. Show finite TTC < 6 s, else depth.

References

1. George, P.; Thouvenin, I.; Frémont, V.; Cherfaoui, V. DAARIA: Driver Assistance by Augmented Reality for Intelligent Automobile. In *Proceedings of the IEEE Intelligent Vehicles Symposium*; IEEE: New York, NY, USA, 2012; pp. 1043–1048. [CrossRef]
2. Schall, M.C., Jr.; Rusch, M.L.; Lee, J.D.; Dawson, J.D.; Thomas, G.; Aksan, N.; Rizzo, M. Augmented Reality Cues and Elderly Driver Hazard Perception. *Hum. Factors* **2013**, *55*, 643–658. [CrossRef] [PubMed]
3. Maag, C.; Schömig, N.; Naujoks, F.; Karl, I.; Keinath, A.; Neukum, A. Measuring workload effects of augmented reality head-up displays using detection response task. *Transp. Res. Part F Traffic Psychol. Behav.* **2023**, *92*, 201–219. [CrossRef]

4. Lopez, J.C.; Moacdieh, N.M. Opacity in car augmented reality head-up displays: Users' preferences, visual attention, and situation awareness. *Appl. Ergon.* **2025**, *129*, 104610. [CrossRef] [PubMed]
5. Endsley, M.R. Toward a Theory of Situation Awareness in Dynamic Systems. *Hum. Factors* **1995**, *37*, 32–64. [CrossRef]
6. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]
7. Koo, J.; Kwac, J.; Ju, W.; Steinert, M.; Leifer, L.; Nass, C. Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *Int. J. Interact. Des. Manuf.* **2015**, *9*, 269–275. [CrossRef]
8. Kaufman, R.; Costa, J.; Kimani, E. Effects of multimodal explanations for autonomous driving on driving performance, cognitive load, expertise, confidence, and trust. *Sci. Rep.* **2024**, *14*, 13061. [CrossRef] [PubMed]
9. Dey, A.K. Understanding and Using Context. *Pers. Ubiquitous Comput.* **2001**, *5*, 4–7. [CrossRef]
10. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
11. Wei, J.; As'arry, A.; Md Rezali, K.A.; Mohamed Yusoff, M.Z.; Ma, H.; Zhang, K. A Review of YOLO Algorithm and Its Applications in Autonomous Driving Object Detection. *IEEE Access* **2025**, *13*, 93688–93711. [CrossRef]
12. Vijayakumar, A.; Vairavasundaram, S. YOLO-based Object Detection Models: A Review and its Applications. *Multimed. Tools Appl.* **2024**, *83*, 83535–83574. [CrossRef]
13. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the CVPR*; IEEE: New York, NY, USA, 2023; pp. 7464–7475. Available online: https://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html (accessed on 19 October 2026).
14. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 1–21. [CrossRef]
15. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 1–28. Available online: https://proceedings.neurips.cc/paper_files/paper/2024/hash/c34ddd05eb089991f06f3c5dc36836e0-Abstract-Conference.html (accessed on 19 October 2026). [CrossRef]
16. Setyanto, A.; Sasongko, T.B.; Fikri, M.A.; Kim, I.K. Near-Edge Computing Aware Object Detection: A Review. *IEEE Access* **2024**, *12*, 2989–3011. [CrossRef]
17. Liang, S.; Wu, H.; Zhen, L.; Hua, Q.; Garg, S.; Kaddoum, G.; Hassan, M.M.; Yu, K. Edge YOLO: Real-Time Intelligent Object Detection System Based on Edge-Cloud Cooperation in Autonomous Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 25345–25360. [CrossRef]
18. Nafea, M.M.; Tan, S.Y.; Jubair, M.A.; Abd, M.T. A Review of Lightweight Object Detection Algorithms for Mobile Augmented Reality. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 536–546. [CrossRef]
19. Schlosser, M.; König, D.; Teutsch, M. A Framework for Benchmarking Real-Time Embedded Object Detection. In *Pattern Recognition*; Springer: Cham, Switzerland, 2022; pp. 545–559. [CrossRef]
20. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In *Proceedings of the CVPR*; IEEE: New York, NY, USA, 2018; pp. 2704–2713. [CrossRef]
21. Al Amin, R.; Hasan, M.; Wiese, V.; Obermaisser, R. FPGA-Based Real-Time Object Detection and Classification System Using YOLO for Edge Computing. *IEEE Access* **2024**, *12*, 73268–73278. [CrossRef]
22. Chaman, M.; El Maliki, A.; Dahou, H.; Hadjoudja, A. Benchmarking YOLO-based deep learning models for real-time object detection in hybrid ADAS and intelligent transportation systems. *Results Eng.* **2026**, *29*, 108942. [CrossRef]
23. Vaghela, R.; Vaishnani, D.; Sarda, J.; Thakkar, A.; Nasit, Y.; Brahma, B.; Bhoi, A.K. Optimizing object detection for autonomous robots: A comparative analysis of YOLO models. *Measurement* **2026**, *257*, 118676. [CrossRef]
24. Srijiranon, K.; Varisthanist, N.; Tardtong, T.; Pumthurean, C.; Tanantong, T. Towards Intelligent Water Safety: Robobuoy, a Deep Learning-Based Drowning Detection and Autonomous Surface Vehicle Rescue System. *Appl. Syst. Innov.* **2026**, *9*, 12. [CrossRef]
25. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *Computer Vision—ECCV 2022*; Springer: Cham, Switzerland, 2022; pp. 1–21. [CrossRef]
26. Wei, W.; Cheng, Y.; He, J.; Zhu, X. A review of small object detection based on deep learning. *Neural Comput. Appl.* **2024**, *36*, 6283–6303. [CrossRef]
27. Zhang, H.; Liu, R.; Zhu, Z.; Feng, Y. RCS-HFPN-YOLOV11: A New Small Target Detection Model. *Algorithms* **2026**, *19*, 306. [CrossRef]
28. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the CVPR*; IEEE: New York, NY, USA, 2024; pp. 16965–16974. [CrossRef]
29. Ultralytics. Ultralytics YOLO26. Official Model Documentation. Available online: https://docs.ultralytics.com/models/yolo26/ (accessed on 8 October 2026).
30. Lee, D.N. A Theory of Visual Control of Braking Based on Information about Time-to-Collision. *Perception* **1976**, *5*, 437–459. [CrossRef] [PubMed]

31. Westhofen, L.; Neurohr, C.; Koopmann, T.; Butz, M.; Schütt, B.; Utesch, F.; Neurohr, B.; Gutenkunst, C.; Böde, E. Criticality Metrics for Automated Driving: A Review and Suitability Analysis of the State of the Art. *Arch. Comput. Methods Eng.* **2023**, *30*, 1–35. [CrossRef]
32. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In *Proceedings of the ICML*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330. Available online: https://proceedings.mlr.press/v70/guo17a.html (accessed on 19 October 2026).
33. Rosenholtz, R.; Li, Y.; Nakano, L. Measuring visual clutter. *J. Vis.* **2007**, *7*, 17. [CrossRef] [PubMed]
34. Cultrera, L.; Seidenari, L.; Becattini, F.; Pala, P.; Del Bimbo, A. Explaining Autonomous Driving by Learning End-to-End Visual Attention. In *Proceedings of the CVPR Workshops*; IEEE: New York, NY, USA, 2020; pp. 340–341. Available online: https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Cultrera_Explaining_Autonomous_Driving_by_Learning_End-to-End_Visual_Attention_CVPRW_2020_paper.html (accessed on 19 October 2026).
35. Łysakowski, M.; Żywanowski, K.; Banaszczyk, A.; Nowicki, M.R.; Skrzypczyński, P.; Tadeja, S.K. Real-Time Onboard Object Detection for Augmented Reality: Enhancing Head-Mounted Display with YOLOv8. In *Proceedings of the IEEE International Conference on Edge Computing and Communications (EDGE)*; IEEE: New York, NY, USA, 2023; pp. 364–371. [CrossRef]
36. Orsholits, A.; Tsukada, M. Edge Vision AI Co-Processing for Dynamic Context Awareness in Mixed Reality. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*; IEEE: New York, NY, USA, 2025; pp. 1294–1295. [CrossRef]
37. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755. [CrossRef]
38. comma.ai. comma2k19 Dataset and Official Example Segment. GitHub Repository. Available online: https://github.com/commaai/comma2k19/tree/4c7f1a6e1957745beadc1def0e7225f559b09a2a (accessed on 13 September 2026).
39. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the CVPR*; IEEE: New York, NY, USA, 2020; pp. 11621–11631. Available online: https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html (accessed on 19 October 2026).
40. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An Open Urban Driving Simulator. In *Proceedings of the CoRL*; PMLR: Cambridge, MA, USA, 2017; Volume 78, pp. 1–16. Available online: https://proceedings.mlr.press/v78/dosovitskiy17a.html (accessed on 19 October 2026).
41. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.