Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evidence-Aware Adaptive Recognition for Intelligent Traffic Sign Systems with Calibrated Uncertainty under Distribution Shift

Abdulrahman S. Alturki

astrky@qu.edu.sa

Qassim University

Article

Keywords: Traffic-sign recognition, visual evidence, conformal prediction, selective classification, context-aware fusion, calibration.

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10950436/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evidence-Aware Adaptive Recognition for Intelligent Traffic Sign Systems with Calibrated Uncertainty under Distribution Shift

Abdulrahman S. Alturki
Department of Electrical Engineering, College of Engineering, Qassim University
Buraydah 51431, Saudi Arabia
Email: astrky@qu.edu.sa

***Abstract*—Traffic-sign recognition under image shift raises two related questions: which label is supported and whether the crop still contains the visual evidence on which that decision is based. Adaptive sign-conditioned recognition (ASCR) keeps these quantities separate. A temperature-calibrated classifier with regularized adaptive prediction sets (RAPS) returns semantic probabilities and label sets; a label-free monitor measures preserved color, area, shape, and component structure. Track-disjoint traffic-sign data and controlled corruptions reveal a high-confidence/low-evidence regime: the lowest-evidence group retains 0.913 confidence at 0.162 error. The evidence–error slope also changes sign across the evaluated environments, indicating that a single global evidence coefficient can be misleading under these shifts. ASCR therefore acts at the decision stage: semantic ambiguity expands the label set, while evidence departure triggers intervention. Within the evaluated settings, this separation provides a transparent decision interface for calibrated classifiers and visual-evidence monitors.**

***Index Terms*—Traffic-sign recognition, visual evidence, conformal prediction, selective classification, context-aware fusion, calibration.**

I. Introduction

TRAFFIC-SIGN recognition under image shift can be examined through two distinct judgments. The semantic judgment asks which traffic-sign label best matches a crop. The evidence judgment asks whether the crop still contains the contrast, color support, coherent components, and geometry that make that label credible. Modern convolutional networks, vision transformers, and calibrated predictors estimate the semantic judgment with increasing accuracy [1]–[5]. Their probability vector expresses preference among labels; a separate measurement describes the state of the visual evidence supporting the task.

Confidence, entropy, ensembles, representation distances, and conformal prediction quantify uncertainty from model outputs or learned features [6]–[11]. Image-quality measures, segmentation stability, reconstruction fidelity, and explicit structural checks measure properties of the input itself [12]–[16]. A single pooled score is useful when both signals keep the same risk meaning across operating conditions. Our analysis shows that the evidence–error relation changes with the corruption mechanism in the evaluated conditions. This motivates preserving semantic uncertainty and visual evidence as separate measurements until the action stage when their conditional associations differ.

We introduce adaptive sign-conditioned recognition (ASCR) to implement this separation. We use *semantic uncertainty* for ambiguity among class labels. A temperature-calibrated classifier produces probabilities $p_T(y \mid X)$ and a regularized adaptive prediction set $\Gamma_\alpha(X)$ containing the labels supported at target miscoverage α [6], [11]. We use *visual evidence* for agreement with an observable traffic-sign specification. A label-free score $\rho(X)$ measures preserved color, area, shape, and component structure. We call each clean or corrupted operating condition an *environment* C. Figure 1 shows the resulting workflow: semantic ambiguity controls the label set, visual evidence controls intervention, and the calibrated probabilities remain unchanged.

The experimental design separates mechanism from physical identity. A controlled benchmark with exact masks isolates illumination, contrast, blur, occlusion, hue, and scale effects. The German Traffic Sign Recognition Benchmark (GTSRB) supplies real crops and repeated frames of physical signs [17]. Our identity analysis finds 1,305 physical-sign tracks shared across the image-level fitting and calibration split used for comparison. We therefore assign whole tracks to model fitting and temperature tuning. Conformal calibration receives a separate set of physical identities.

Our main finding is that visual evidence and semantic risk change differently across environments. Under contrast reduction, the visual-evidence score changes much faster than semantic performance, revealing crops that retain high class confidence after substantial evidence loss. Across environments, the conditional evidence–error slope changes sign. The total-covariance decomposition shows why pooled analysis can hide this change, producing a Simpson-type aggregation effect [18]. This structure motivates context-aware routing when the evidence–error association changes by environment, rather than assuming one universal monotone coefficient for visual evidence.

The paper makes four contributions.

1) We formulate a two-channel recognition interface that preserves calibrated semantic probabilities and routes

explicit visual evidence at the decision stage.
2) We define evidence lead, a high-confidence/low-evidence region, and an environment-conditioned slope that reveal when visual evidence and semantic risk evolve at different rates.
3) We derive a context requirement for global monotone fusion and an error accounting inside and outside the retained evidence domain that connect the statistical analysis to commit, prediction-set, and intervention actions.
4) We provide a physical-track-disjoint GTSRB protocol and evaluate the framework through controlled mechanism tests, official-crop evaluation, calibration, conformal efficiency, selective risk, mask analysis, and compute measurements.

II. Related Work and Positioning

We organize prior work around the two questions ASCR separates: what label the model supports and what visual evidence the input preserves.

A. Modern recognition and explicit visual evidence

Traffic-sign recognition progressed from color segmentation, morphology, gradients, and discriminative classifiers to multi-scale and high-capacity convolutional systems [19]–[28]. Full-scene perception adopted proposal-based and single-stage detectors, including Faster R-CNN, SSD, YOLO, and focal-loss detectors [29]–[32]. Contemporary vision now includes scalable ConvNets, vision transformers, masked autoencoders, self-supervised foundation features, and promptable segmentation [1]–[3], [14]–[16]. These advances strengthen semantic representations and also provide label-free evidence measurements such as reconstruction fidelity, representation distance, patch consistency, and mask stability. ASCR places any such measurement beside, rather than inside, the calibrated semantic distribution.

GTSRB established cropped recognition, the German Traffic Sign Detection Benchmark (GTSDB) added full-scene detection, TT100K broadened scale and taxonomy, and the Challenging Unreal and Real Environments for Traffic Sign Recognition benchmark (CURE-TSR) introduced controlled adverse-condition sequences [17], [33]–[35]. Our evaluation combines exact-mask controlled interventions with real repeated physical identities. No-reference measures such as BRISQUE and NIQE measure departures from natural-image statistics [12], [13]; ASCR instead evaluates agreement with a declared task specification. Both explicit and learned evidence coordinates fit the same two-channel formulation.

B. Uncertainty, conformal prediction, and context-conditioned fusion

Temperature scaling, softmax confidence baselines, Monte Carlo (MC) dropout, deep ensembles, Mahalanobis distance, ODIN-style scores, and modern calibration studies quantify uncertainty from model outputs or learned representations [4]–[9], [36], [37]. Split conformal prediction converts fixed nonconformity scores into finite-sample marginal coverage under exchangeability [10], [38]. RAPS regularizes adaptive sets, simplified adaptive prediction sets (SAPS) reduce dependence on probability magnitudes, learned-feature methods adapt the representation, and recent work studies set efficiency, robustness, risk control, and shift adaptation [11], [39]–[44]. Weighted conformal prediction addresses a specified covariate-shift model [45].

Selective classifiers, learned confidence models, and conformal reject options connect uncertainty to downstream actions [46]–[52]. Conformal assurance monitoring has also been applied directly to traffic-sign recognition through learned embeddings [53]. Mixture-of-experts systems provide the natural modern extension when the correct combination depends on input context: a gate routes or weights specialized experts rather than assigning one fixed coefficient globally [54], [55]. ASCR provides a diagnostic for whether context may be useful by measuring whether the evidence–error association remains stable across environments.

C. Hierarchical calibration units

GTSRB stores multiple frames of each physical sign in tracks [17]. Random image-level development partitions can therefore share physical identities across model fitting and calibration. We quantify this dependence and partition whole tracks into training, temperature tuning, and conformal calibration. Hierarchical and non-exchangeable conformal methods formalize the corresponding distinction between crop-level and physical-object-level inference [56]–[58]. Calibration parameters and score choices remain frozen before final conformal calibration, avoiding tuning-induced dependence [59], [60].

III. Two-Channel Method

A. Semantic channel

Let a semantic classifier–convolutional, transformer-based, or frozen-encoder linear head–produce logits $\ell(X) \in \mathbb{R}^K$ [1]–[3], [15]. Following post-hoc temperature scaling [6], an independently tuned temperature gives

$$p_T(y \mid X) = \mathrm{softmax}(\ell(X)/T)_y \,. \tag{1}$$

For probabilities sorted as $p_{(1)} \geq \cdots \geq p_{(K)}$ and true-label rank r_y, the deterministic RAPS score [11] is

$$s_{\lambda,k}(X,y) = \sum_{j=1}^{r_y} p_{(j)}(X) + \lambda(r_y - k)_+. \tag{2}$$

The split-conformal quantile $\widehat{q}_{1-\alpha}$ is estimated on the independent conformal partition [10], [38]. At test time,

$$m^*(X) = \min\{m : C_m(X) \geq \widehat{q}_{1-\alpha}\}, \tag{3}$$

$$\Gamma_\alpha(X) = \{\pi_j(X) : 1 \leq j \leq m^*(X)\}, \tag{4}$$

where C_m is the cumulative regularized score in sorted order.

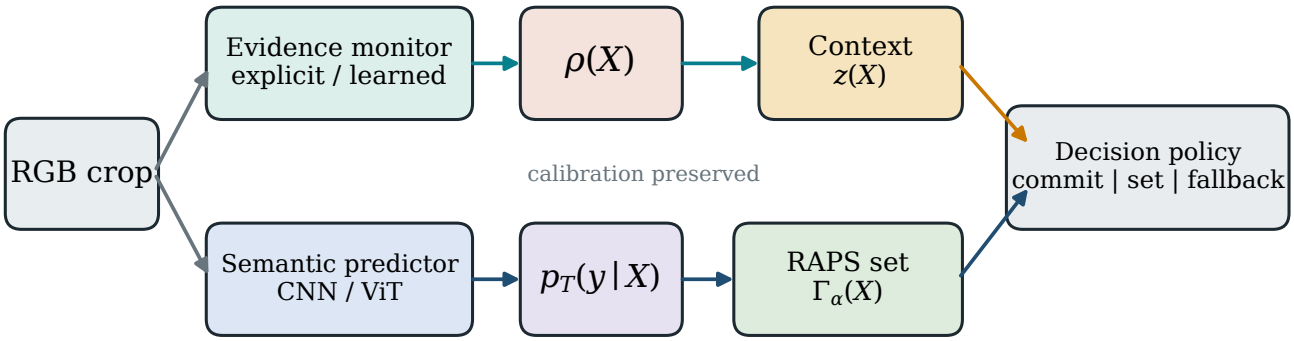
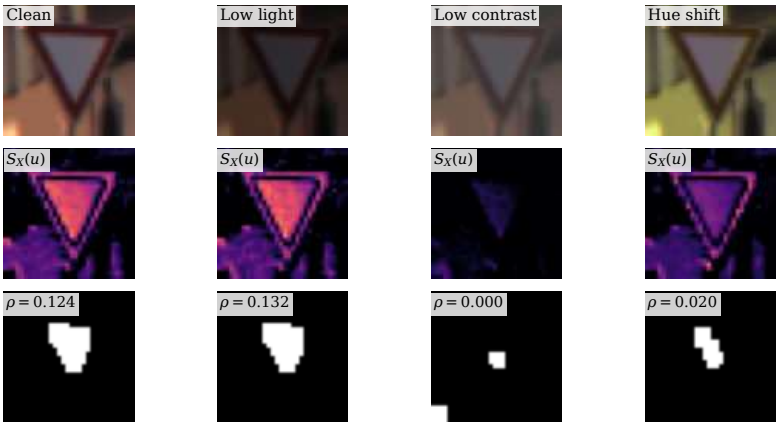
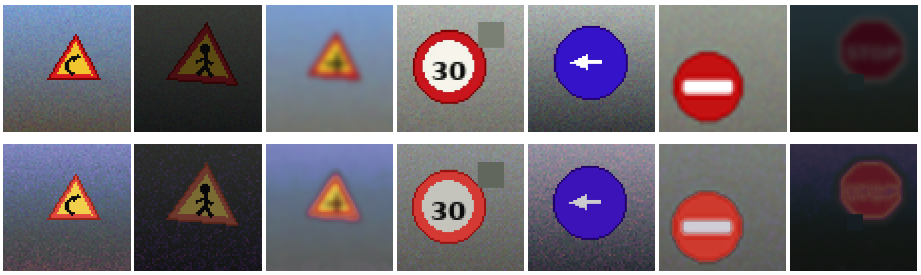


Fig. 1. ASCR separates semantic inference from visual-evidence testing. The semantic branch produces $p_T(y \mid X)$ and a RAPS set $\Gamma_\alpha(X)$. The image-processing branch produces an evidence score $\rho(X)$ from label-free witnesses. Late fusion retains the semantic set or invokes fallback while preserving the calibrated semantic probabilities.

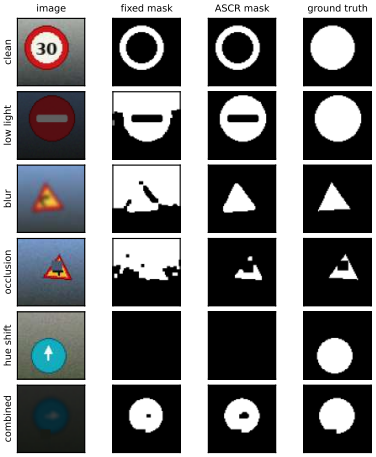
(a)



(a) Matched corruptions on one official crop.



(b) Class-diverse procedural crops and retained masks.



(c) Fixed and adaptive masks against exact masks.

Fig. 2. Visual-evidence inventory. The matched official crop shows how the score field and retained component respond to brightness, contrast, and hue changes; class-diverse procedural examples expose color and contour behavior; exact masks isolate the segmentation mechanism before semantic prediction or conformal calibration.

B. Visual-evidence channel

A *witness* is a normalized, label-free measurement of one expected image property. ASCR accepts a witness vector $\boldsymbol{r}_\phi(X) \in [0,1]^J$. Explicit color and component measurements, frozen-encoder distances, masked-reconstruction fidelity, and segmentation stability share this input–output form [14]–[16]. The reported experiments use four directly inspectable witnesses,

$$\boldsymbol{r}(X) = \big(r_{\text{color}}, r_{\text{area}}, r_{\text{shape}}, r_{\text{frag}}\big), \tag{5}$$

constructed from color constancy, robust dispersion, circular hue distance, thresholding, and connected components [22], [23], [61]–[64]. A weighted geometric product combines the witnesses,

$$\rho_{\boldsymbol{a}}(X) = \text{clip}\left[\prod_{j=1}^{4} r_j(X)^{a_j}, 0, 1\right], \qquad a_j \geq 0, \tag{6}$$

following monotone product aggregation [65]. The appendix specifies the prototypes and thresholds used in the experi-

ments. It also specifies the morphology and exponents.

Figure 2 makes the witness construction concrete: matched real crops show the response to illumination, contrast, and hue, while exact masks separate evidence extraction from semantic recognition.

The score has a clear operational meaning. It is bounded, label-free, monotone in every witness, and equal to zero when a required witness vanishes. Away from clipping,

$$\frac{\partial \log \rho}{\partial \log r_j} = a_j, \tag{7}$$

so a_j gives the relative sensitivity to witness j. The normalization also makes the score approximately invariant to global intensity scaling. The absolute value of ρ depends on the chosen witnesses; ASCR uses its rank and a tuning-defined threshold. We define the retained evidence domain as $G_\tau = \{X : \rho(X) \geq \tau\}$.

C. Environment-conditioned association

Let $Z = \mathbf{1}\{\widehat{Y} \neq Y\}$ denote classification error, let $S(X)$ denote a semantic risk score such as entropy, and let C denote the operating environment (clean, low light, low contrast, blur, pixelation, or hue rotation). We first compare how quickly the two channels change from clean operation:

$$L_\rho(c) = 1 - \frac{\mathbb{E}[\rho \mid C = c]}{\mathbb{E}[\rho \mid C = \text{clean}]}, \tag{8}$$

$$L_F(c) = 1 - \frac{F_1(c)}{F_1(\text{clean})}, \qquad \Lambda(c) = L_\rho(c) - L_F(c). \tag{9}$$

A positive Λ identifies an early evidence response: the evidence monitor moves farther from its clean reference than macro-F_1. This quantity compares relative sensitivities rather than error probabilities. It is useful because an intervention can become informative before semantic performance changes by the same proportion.

We also define the high-confidence/low-evidence region

$$\mathcal{H}_{\tau,\gamma} = \left\{ X : \rho(X) < \tau,\ \max_y p_T(y \mid X) \geq \gamma \right\}, \tag{10}$$

which identifies crops that receive concentrated semantic predictions while lying outside the retained evidence domain.

To measure the local relation between evidence and error, define the environment-conditioned slope

$$\beta_\rho(c) = \frac{\partial}{\partial \rho} \operatorname{logit} \Pr(Z = 1 \mid S, \rho, C = c)\,. \tag{11}$$

The law of total covariance

$$\operatorname{Cov}(\rho, Z) = \mathbb{E}_C[\operatorname{Cov}(\rho, Z \mid C)] + \operatorname{Cov}_C(\mathbb{E}[\rho \mid C], \mathbb{E}[Z \mid C]) \tag{12}$$

separates the average within-environment relation from changes in environment prevalence and difficulty [18]. This decomposition explains a Simpson-type aggregation effect: a pooled coefficient can look stable even when the conditional meaning of ρ changes across environments.

Context requirement. Suppose two environments share the same range of (S, ρ) but assign opposite evidence–error slopes. A global score that always increases, or always decreases, with ρ must then misorder risk in at least one environment. Context resolves this ambiguity: a gate $g_\theta(S, \rho, z)$ can use sensor state, scene information, or learned visual tokens as z [54], [55]. The appendix gives the one-line ordering argument.

Context also has a direct decision value. Let $\mathcal{R}^*_{\text{global}}$ be the minimum expected loss when an action uses only (S, ρ), and let $\mathcal{R}^*_{\text{ctx}}$ use (S, ρ, C). Since every global policy is also available to the context-aware policy class,

$$\begin{aligned} \mathcal{R}^*_{\text{ctx}} &= \mathbb{E}\Big[\min_a \mathbb{E}(L \mid S, \rho, C)\Big] \\ &\leq \mathbb{E}\Big[\min_a \mathbb{E}(L \mid S, \rho)\Big] = \mathcal{R}^*_{\text{global}}. \end{aligned} \tag{13}$$

The inequality is strict whenever the same (S, ρ) region requires different optimal actions in environments of nonzero probability. The observed slope change supports a role for context in cost-sensitive routing and connects the statistical result to context-gated and mixture-of-experts decisions [46], [54], [66]. A Bayes-optimal action further depends on the losses and action costs specified by the application.

D. Decision layer and retained-domain accounting

ASCR represents each crop by a three-coordinate decision state,

$$\boldsymbol{u}(X) = \big(S(X),\, 1 - \rho(X),\, |\Gamma_\alpha(X)|\big), \tag{14}$$

where the coordinates measure semantic risk, visual-evidence departure, and label-set ambiguity. The available actions are: *commit* to a singleton label, *return a set* of plausible labels, or *intervene* by routing the crop to another sensor, model, or operator. For context z, the Bayes action is

$$a^*(X, z) = \arg\min_{a \in \mathcal{A}} \mathbb{E}[L(a, Y, \Gamma_\alpha, \rho, z) \mid X]\,, \tag{15}$$

which links the two-channel state to cost-sensitive and selective decisions [43], [46], [66].

The reported operating point uses a transparent evidence threshold τ, fixed on the tuning partition before conformal calibration and testing:

$$\delta_\tau(X) = \begin{cases} \text{intervene}, & \rho(X) < \tau, \\ \text{commit to } \pi_1(X), & \rho(X) \geq \tau,\ |\Gamma_\alpha(X)| = 1, \\ \text{return } \Gamma_\alpha(X), & \rho(X) \geq \tau,\ |\Gamma_\alpha(X)| > 1. \end{cases} \tag{16}$$

This rule gives each channel one role. RAPS changes the number of plausible labels when semantic ambiguity grows. The evidence score changes the action when the crop leaves the declared visual domain. Because the intervention is applied after probability calibration, it does not alter the meaning of p_T or the ordering of the semantic labels.

Let $A = \{Y \notin \Gamma_\alpha(X)\}$ be semantic set error. On G_τ, suppose a target distribution Q satisfies $dQ/dP \leq \kappa$, let $\eta_Q = Q(G_\tau^c)$, and let $P(A) \leq \alpha$. Then

$$Q(A) \leq \kappa\alpha + \eta_Q. \tag{17}$$

The first term accounts for semantic set error inside the retained evidence domain; the second is the intervention

mass outside it [45], [57], [58]. A system can choose its operating point by minimizing

$$\mathcal{R}(\delta) = \mathbb{E}\big[c_e \mathbf{1}\{\delta \neq \text{intervene}, Y \notin \Gamma\} + c_i \mathbf{1}\{\delta = \text{intervene}\} + c_s |\Gamma| \mathbf{1}\{\delta \neq \text{intervene}\}\big], \tag{18}$$

where c_e, c_i, and c_s encode semantic error, intervention, and set-size costs.

IV. Experimental Protocol

A. Controlled mechanism study

SignStress-12 provides class-diverse procedural crops with exact masks and controlled changes in illumination, contrast, blur, occlusion, hue, scale, and combined appearance. Ten independently generated seeds use separate model-fitting, tuning, conformal-calibration, and test partitions. Exact masks make segmentation and witness behavior directly observable, while seed-level replication supports paired comparisons.

B. Physical-sign-track-disjoint GTSRB

The official GTSRB training data organize 30 frames of each physical sign into tracks [17]. We parse class and track identities and assign whole tracks to training, temperature tuning, or conformal calibration; the official test archive remains untouched. The corrected partition contains 27,509 training crops from 917 tracks and 5,850 crops from 195 tracks in each development holdout. All pairwise track intersections are zero.

The image-level comparator shares 1,305 physical-sign tracks across model fitting and calibration. Every comparator calibration crop belongs to a represented training track. Whole-track assignment gives each development stage distinct physical identities before temperature or conformal estimation. Repeated frames form a hierarchical sample rather than independent calibration units. Conformal methods for grouped and non-exchangeable data formalize the corresponding inferential structure [56]–[58].

C. Models, corruptions, and metrics

The public-data semantic branch uses a residual backbone trained on the track-disjoint partition with AdamW, label smoothing, dropout, and photometric augmentation. The selected model reaches 0.983 tuning macro-F_1, and independent temperature fitting gives $T = 0.564$. The mathematical interface requires calibrated logits and a prediction set; its equations therefore remain unchanged when the semantic backbone changes. A balanced histogram-of-oriented-gradients (HOG) linear support-vector machine (SVM) supplies a head-independent evidence control. The controlled study also evaluates distance-weighted k-nearest neighbors (kNN) and a random forest on the explicit ASCR descriptor [1], [24], [25], [67], [68].

Official test crops are evaluated clean and under brightness reduction, contrast reduction, Gaussian blur, pixelation, and hue rotation. We report macro-F_1, accuracy, expected calibration error (ECE), RAPS coverage, mean set size, singleton rate, evidence-intervention rate, retained-subset coverage, mask intersection-over-union (IoU) and Dice scores, and CPU latency. We use *selective risk* for the classification error among samples retained after sorting by a trust score [47], [69]. Held-out error-ranking comparisons report the area under the receiver-operating-characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC) [70], [71]. "Mean corruption" denotes the arithmetic mean over the stated non-clean conditions.

The conformal calibration uses crops from physical signs absent from model fitting and temperature tuning. The official-test values are crop-level operating characteristics on distinct physical identities. Hierarchical and non-exchangeable conformal methods provide the corresponding route to group-level coverage when the physical sign, rather than the crop, is the inferential unit [56]–[58].

V. Results

A. Controlled recognition, masking, and structural calibration

The ten-seed mechanism study separates representation quality from uncertainty behavior. The adaptive descriptor with a random-forest head reaches 0.828 ± 0.021 clean and 0.642 ± 0.024 mean-corruption macro-F_1, compared with 0.504 ± 0.025 and 0.281 ± 0.023 for fixed-mask kNN. Figure 3(a) shows that the recognition gain persists across seeds. Figure 3(b) then separates coverage from efficiency: temperature-calibrated ASCR gives the compact operating point, while direct probability contraction expands the sets and changes ECE. We therefore preserve the semantic distribution and use visual evidence at the decision stage.

The exact-mask analysis in the appendix locates the mechanism behind this gain. Adaptive masks improve mean IoU and Dice over fixed masks, gain normalization provides the largest combined-corruption benefit, and threshold variants remain interchangeable components rather than part of the central claim. Direct hue rotation activates a distinct prototype regime because a fixed prototype family encodes task-relevant color support rather than arbitrary chromatic rotations.

B. Track-disjoint GTSRB-43 performance

On official GTSRB-43 crops, the calibrated residual backbone reaches 0.954 clean macro-F_1 with ECE 0.010. Model fitting, temperature tuning, and conformal calibration use distinct physical identities. The two-channel equations act on calibrated logits, a conformal set, and a label-free evidence score, which defines a common interface for other semantic and evidence models.

RAPS gives 0.998 empirical clean coverage with mean set size 5.07. Across brightness, contrast, blur, pixelation, and hue rotation, coverage remains between 0.980 and 0.998 and mean set size between 5.07 and 5.82. The RAPS quantile is 0.999950; it is a cumulative-score threshold, whereas mean set size reports the fraction of labels retained operationally. Reporting set size and singleton

TABLE I

PHYSICAL-SIGN-TRACK ANALYSIS AND TRACK-DISJOINT GTSRB PARTITION. THE IMAGE-LEVEL COMPARATOR PLACES 1,305 PHYSICAL-SIGN TRACKS ACROSS FITTING AND CALIBRATION. WHOLE-TRACK ASSIGNMENT GIVES EACH DEVELOPMENT STAGE DISTINCT PHYSICAL IDENTITIES.

Protocol or stage	Crops	Tracks	Crossing tracks	Statistical role
Image-level comparator calibration	6,666	–	1,305	Identity-overlap analysis
Track-disjoint training	27,509	917	0	Model fitting
Track-disjoint tuning	5,850	195	0	Temperature and hyperparameters
Track-disjoint conformal calibration	5,850	195	0	RAPS quantile
Official test	12,630	–	–	Final evaluation

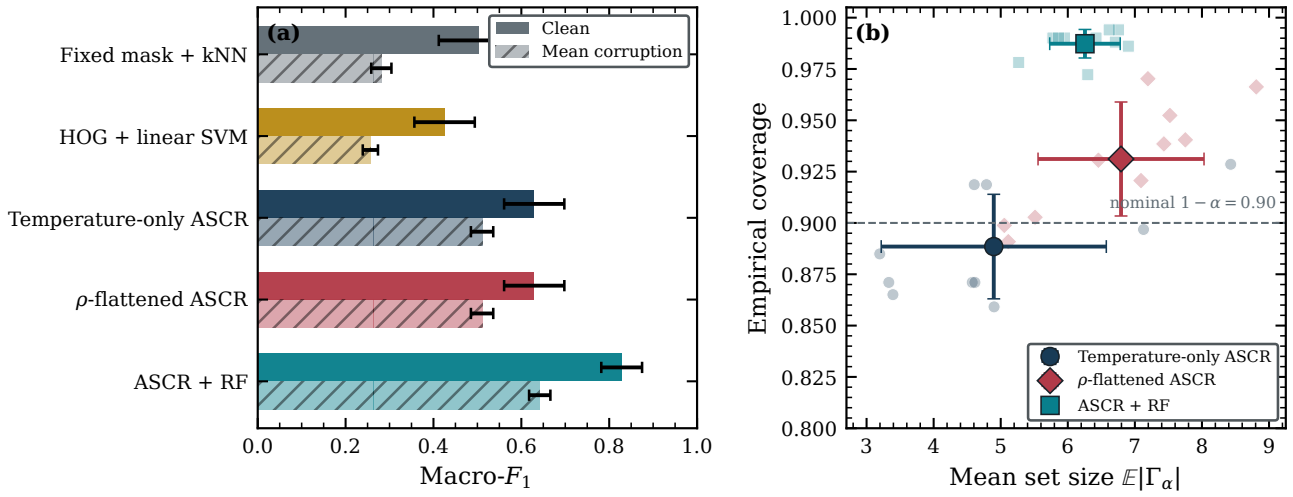


Fig. 3. Ten-seed controlled results with equal-sized panels. (a) Clean and mean-corruption macro-F_1; error bars show one standard deviation. (b) Coverage–efficiency operating points. Small markers show individual seeds, large markers show means, and error bars show one standard deviation.

rate alongside coverage follows the efficiency emphasis of modern conformal evaluation [39], [41]. The condition-level values are summarized in Table II.

C. Visual evidence can lead semantic change

Figure 4(a) compares the channel losses defined in (9). Brightness reduction leaves mean evidence nearly unchanged because the normalization removes a uniform intensity gain. Contrast reduction compresses color separation and weakens the retained components, so mean ρ changes from 0.155 to 0.057 ($L_\rho = 0.629$). The residual classifier still uses coarse shape and learned texture, and macro-F_1 changes more slowly ($L_F = 0.044$). The resulting lead, $\Lambda = +0.585$, therefore has a concrete mechanism: explicit image evidence departs from its clean specification before semantic performance changes by the same relative amount. Pixelation perturbs component geometry while preserving coarse class structure, which gives a smaller positive lead. Blur preserves broad color regions but erases discriminative edges, so semantic loss leads the explicit evidence score.

The condition summary in Table II connects this geometry to action. Clean and brightness-reduced crops have nearly identical mean evidence and intervention rates. Contrast reduction moves a much larger fraction of crops outside G_τ, while RAPS coverage remains high. ASCR therefore separates two useful outputs: the semantic set stays calibrated, and the intervention rate reports evidence-domain departure.

D. Environment-conditioned association

Figure 4(b) reports the slopes in (11). Clean, low-light, low-contrast, blur, and pixelation environments yield $\beta_\rho(c) < 0$ with 95% intervals below zero: within these environments, stronger visual evidence accompanies lower error after conditioning on semantic risk. Hue rotation yields $\beta_\rho(c) = +0.283$ with interval $[0.222, 0.345]$. The reason is structural rather than numerical: a rotated hue can move toward another declared prototype, so a larger score no longer has the same risk meaning. The pooled slope is -0.482 with interval $[-0.518, -0.445]$. Equation (12) shows why one pooled coefficient is incomplete: it combines the within-environment relation with differences in each environment's mean evidence and error rate. The polarity span $\Delta_\beta = 1.374$ summarizes this change of interpretation across environments.

An independent held-out-family analysis shows the same qualitative pattern. Confidence and entropy preserve semantic ranking across both held-out splits, while the fitted coefficient of ρ changes from -3.63 to $+2.39$. This coefficient change agrees with the condition-wise slopes: the operational interpretation of evidence depends on the evaluated environment, whereas semantic uncertainty retains its calibrated semantic role. Figure 5 shows the consequence

TABLE II
TRACK-DISJOINT GTSRB-43 OPERATING CHARACTERISTICS. SEMANTIC METRICS USE THE TEMPERATURE-SCALED RESIDUAL NETWORK AND INDEPENDENTLY CALIBRATED RAPS. THE FINAL COLUMN REPORTS THE SEPARATE EVIDENCE-INTERVENTION RATE AT THE FROZEN TUNING THRESHOLD.

Condition	Macro-F_1	ECE	RAPS coverage	Mean set size	Mean ρ	Intervention rate
Clean	0.954	0.010	0.998	5.07	0.155	0.049
Low light	0.934	0.014	0.995	5.27	0.155	0.050
Low contrast	0.912	0.025	0.987	5.13	0.057	0.310
Blur	0.856	0.051	0.980	5.42	0.142	0.046
Pixelation	0.879	0.032	0.994	5.52	0.128	0.051
Hue shift	0.876	0.043	0.991	5.82	0.136	0.059

across retained fractions. Confidence and entropy rank generic semantic risk, while ρ identifies departure from the declared image specification and therefore enters the action rule rather than the class-probability vector.

E. High confidence with low visual evidence

The region $\mathcal{H}_{\tau,\gamma}$ in (10) gives the main operational insight. Under contrast reduction, the lowest-evidence group retains mean semantic confidence 0.913 and has error 0.162. Under blur, mean confidence remains 0.850 while error reaches 0.267. Figure 6 shows why the second channel matters: confidence stays concentrated because it compares labels within the classifier's representation, whereas the evidence decile records whether the input still satisfies the observable image specification. The rising error in low-evidence groups therefore supplies action information that a concentrated class distribution alone does not express.

Hue rotation illustrates why context matters. A rotated crop can match another declared color prototype, so the evidence score receives a different error meaning in that environment. Class-conditioned prototype banks, frozen representation distance, segmentation stability, or a learned context gate can supply the corresponding context while preserving the same two-channel decision state.

VI. DISCUSSION

A. A modern two-channel interface

ASCR defines a simple input–output contract. The semantic channel supplies calibrated logits and a conformal label set. The evidence channel supplies a label-free score constructed independently of the class probabilities. The reported color and component witnesses make every response traceable to visible pixels, support exact-mask tests, and run in 13–15 ms on a CPU. A learned monitor can occupy the same interface by replacing the witness vector with reconstruction residuals, frozen-feature distances, token consistency, or segmentation stability [14]–[16]. The separation is therefore architectural: each channel can improve without changing the calibration target or the evidence definition of the other.

The absolute scale of ρ follows from its witnesses and exponents. Its operational content comes from ordering crops relative to the frozen tuning threshold and from comparing condition-wise distributions. This rank-based interpretation applies equally to explicit and learned evidence scores.

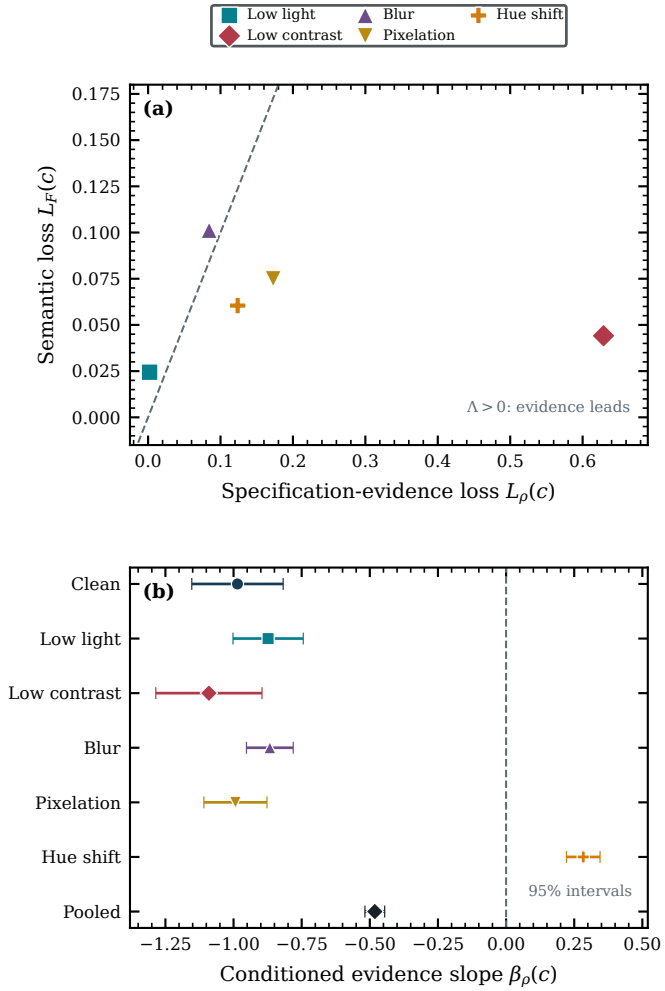


Fig. 4. Environment-conditioned evidence analysis. (a) Normalized visual-evidence loss L_ρ versus semantic loss L_F; positive $\Lambda = L_\rho - L_F$ identifies an evidence-leading regime. (b) Standardized condition-wise logistic slopes between ρ and error with 95% intervals. Hue rotation changes the sign observed in the other environments, while the pooled estimate combines the conditional relations–a Simpson-type aggregation effect.

B. The Simpson-type aggregation insight

The covariance decomposition in (12) separates two sources of association. The first term measures how evidence and error move together inside each environment. The second term measures how their environment means move together after pooling. Figure 4(b) shows that

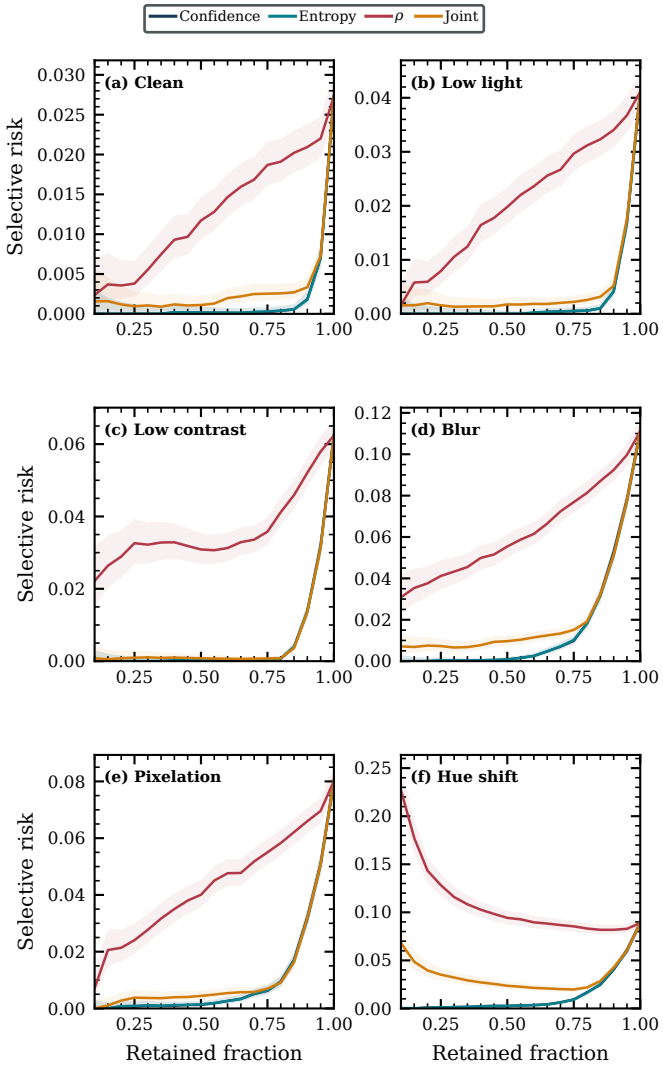


Fig. 5. Selective-risk curves for all public-data conditions. Each selector orders samples from most to least trusted; risk is the classification error among the retained samples [47], [69]. Shading gives 95% Wilson intervals [72]. Confidence and entropy rank semantic risk consistently; the joint score changes with the environment-dependent meaning of ρ.

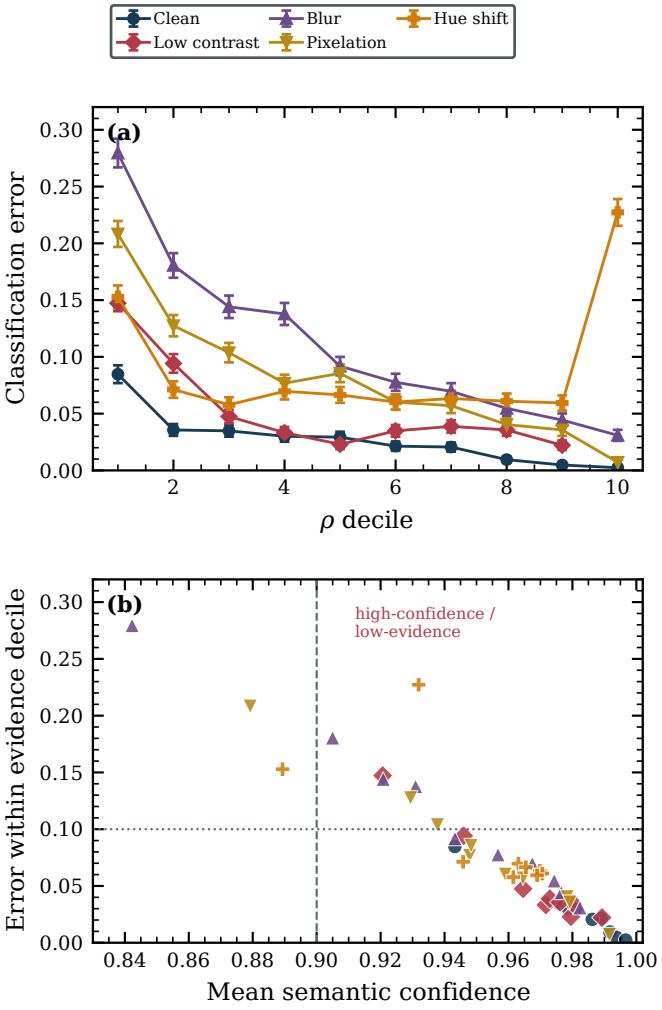


Fig. 6. Evidence-conditioned error and confidence. Samples are grouped by evidence decile; the lowest-evidence groups retain concentrated semantic confidence while the corresponding error rate rises.

the within-environment slope can change sign. A pooled model then averages relations with different meanings. This is the relevant Simpson-type aggregation effect. The decomposition identifies the condition under which one global evidence coefficient becomes insufficient across environments.

The context requirement turns that statistical observation into a design consideration. Global confidence or entropy can rank semantic risk, while visual evidence remains a separate coordinate. When an application learns a joint policy and the evidence–error association varies by operating condition, its gate can condition on environment information such as sensor state, scene metadata, or visual tokens [54], [55]. This arrangement permits nonlinear fusion while preserving the interpretation of the calibrated semantic output.

C. Decision geometry and analytical redundancy

The state $\boldsymbol{u}(X)$ in (14) separates label ambiguity, evidence departure, and set size. Equation (16) maps these quantities to three concrete actions: commit, return a set, or intervene. Figure 7 gives representative examples from these operating regions. The high-confidence/low-evidence region is especially important because an evidence intervention changes the action while leaving the semantic probabilities untouched.

This arrangement follows analytical redundancy in fault monitoring: independently constructed measurements add value when they respond to different mechanisms [73]. Equation (13) states the corresponding value-of-information result. A context-aware policy class contains the global rule and can achieve lower expected loss when the optimal action changes with the environment. The experiments establish environment-dependent evidence–error associations. Equation (18) then expresses the operational trade through application-defined costs for semantic error, intervention, and set size.

D. Physical identity as the calibration unit

The physical-track analysis shows that the partition defines the prediction task. Image-level splitting evaluates another frame of a represented sign. Whole-track splitting evaluates a new physical identity. Frames from one track share shape, background, viewpoint, and acquisition conditions. Whole-track assignment therefore aligns the statistical unit with the physical object. Grouped or non-exchangeable conformal methods provide the corresponding theory when the sign is the inferential unit [56]–[58].

The controlled and public datasets play complementary roles. Exact masks identify how each witness responds to a known image change. Real tracks test calibration and recognition across physical identities. Full-scene datasets add detector confidence and non-sign candidates, and adverse-condition sequences add naturally occurring shifts [33]–[35]. The two-channel state provides a natural route for future extension to full-scene data by adding detector outputs to the semantic channel and task-appropriate evidence measurements to the visual channel.

E. Scope and limitations

The empirical evaluation focuses on cropped traffic-sign recognition with GTSRB and the procedural SignStress-12 study. It covers the five stated test corruptions. The reported robustness results characterize this evaluation domain. Broader road-scene shifts such as weather variation and camera faults form separate evaluation settings. Adversarial perturbations and detector errors also form separate settings. Conformal coverage is reported at the crop level for physical identities disjoint from model fitting and calibration. Repeated observations of one physical sign call for hierarchical or non-exchangeable conformal constructions when group-level coverage is required. The evidence score follows the declared color-and-geometry specification. Its scale and relation to error can vary across sign standards and sensors. They can also vary across operating environments. The decision rule is evaluated through its operating characteristics. Deployment evaluation can incorporate target-environment calibration data. It can also incorporate context information when the evidence–error relation varies and application-defined intervention costs.

VII. Conclusion

ASCR separates two quantities that can diverge under image shift: semantic uncertainty and observable visual evidence. The semantic channel returns calibrated probabilities and RAPS prediction sets, while the evidence channel measures retained color, area, shape, and component structure without modifying the class probabilities.

In the evaluated GTSRB and controlled-corruption experiments this separation exposes cases that confidence alone does not summarize. Under contrast reduction the mean visual-evidence score changes substantially before macro-F_1 changes by a comparable proportion. The lowest-evidence group retains mean confidence 0.913 while reaching 0.162 error. The conditional evidence–error slope also changes sign across the tested environments. These observations support treating visual evidence as a context-dependent decision variable instead of assigning one global monotone coefficient the same interpretation under every evaluated shift.

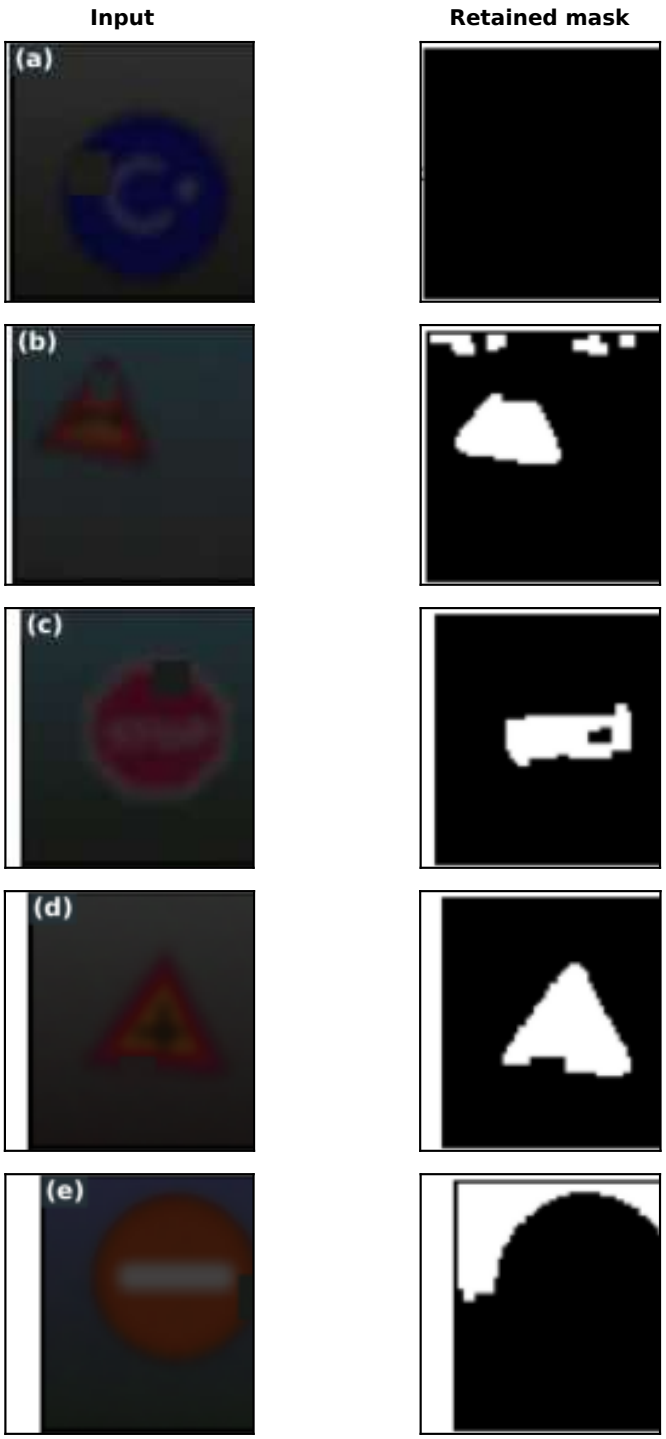


Fig. 7. Representative ASCR decision states. Each row pairs an input with its retained evidence mask and illustrates evidence departure, fragmentation, semantic ambiguity, or context-dependent prototype response.

The physical-track analysis further shows why the calibration unit matters for repeated traffic-sign crops. Assign-

ing complete tracks keeps physical identities separate across model fitting and temperature tuning. The same separation is maintained for conformal calibration. Together with the controlled study this provides a transparent evaluation protocol for the proposed two-channel interface.

The present results establish the behavior of ASCR for cropped GTSRB recognition and the stated controlled corruptions. Full-scene detection and additional sensors provide direct settings for further evaluation. Naturally occurring weather and camera shifts provide another setting. Group-level conformal guarantees extend the evaluation to repeated observations of physical signs.

Appendix
Implementation and Supporting Analyses

A. Exact visual-evidence implementation

The main text defines the evidence score through its witnesses and required properties. This appendix specifies the visual-evidence instance used in the reported experiments.

Following gray-world color constancy [61], for RGB channel means μ_c and $\bar{\mu} = \frac{1}{3}\sum_c \mu_c$,

$$X_c^{\mathrm{gw}}(u) = X_c(u)\frac{\bar{\mu}}{\mu_c + \epsilon}, \tag{19}$$

$$\widehat{X}(u) = \mathrm{clip}\left[X^{\mathrm{gw}}(u)\frac{0.45}{\mathrm{mean}(X^{\mathrm{gw}}) + \epsilon}, 0, 1\right]. \tag{20}$$

After HSV conversion, circular hue distance $d_{\mathbb{S}^1}$ follows standard directional-statistics treatment of angular data [63]; prototypes $\mathcal{H} = \{0, 0.60, 0.13\}$ define

$$S_X(u) = \max_{h_0 \in \mathcal{H}} \exp\left[-\frac{d_{\mathbb{S}^1}(h(u), h_0)^2}{2\sigma_{h_0}^2}\right] \times g_s(s(u))g_v(v(u)). \tag{21}$$

where g_s and g_v are clipped linear gates. Using a quantile and median absolute deviation as robust location/dispersion summaries [23], [62], the adaptive threshold is

$$\tau_X = \mathrm{clip}[Q_{0.86}(S_X) + 0.25\,\mathrm{MAD}(S_X), 0.10, 0.65]. \tag{22}$$

Opening, closing, and a ten-pixel component rule remove isolated responses, following standard binary morphology and connected-component practice [22], [23], [64]. For largest retained component $\widehat{C}$, area fraction A_f, perimeter estimate P, and plausible-component count N_c,

$$r_{\mathrm{color}} = |\widehat{C}|^{-1}\sum_{u \in \widehat{C}} S_X(u), \tag{23}$$

$$r_{\mathrm{area}} = \exp\left[-\frac{(\log A_f - \log 0.15)^2}{2(0.85)^2}\right], \tag{24}$$

$$r_{\mathrm{shape}} = \mathrm{clip}\left(\frac{4\pi|\widehat{C}|}{P^2 + \epsilon}, 0, 1\right), \tag{25}$$

$$r_{\mathrm{frag}} = \exp[-0.35(N_c - 1)_+], \tag{26}$$

$$\rho(X) = \mathrm{clip}\left[r_{\mathrm{color}} r_{\mathrm{area}}^{0.6} r_{\mathrm{shape}}^{0.5} r_{\mathrm{frag}}^{0.75}, 0, 1\right]. \tag{27}$$

We fix all parameters before test evaluation. The exponent perturbation analysis varies all exponents by ±25% and tests component-heavy and fragmentation-heavy configurations.

The ASCR descriptor concatenates edge summaries, downsampled score and mask fields, component statistics, and a score-weighted 12-bin hue histogram, for 827 features. Training-partition standardization precedes distance-weighted five-nearest-neighbor or random-forest classification [67], [68]. The kNN probability uses the five-neighbor set $\mathcal{N}_5(z)$,

$$p_0(y \mid X) = \frac{\sum_{i \in \mathcal{N}_5(z)} \mathbf{1}(y_i = y)(\|z - z_i\|_2 + \epsilon)^{-1}}{\sum_{i \in \mathcal{N}_5(z)}(\|z - z_i\|_2 + \epsilon)^{-1}}. \tag{28}$$

TABLE III
Track-disjoint class-domain diagnostic. The two class slices separate color/contour structure from fine printed-symbol discrimination and clarify the respective roles of the explicit descriptor and semantic backbone.

Model	Digit-heavy IDs 0–11 F_1	Cov.	Size	Color/contour Diverse-12 F_1	Cov.	Size
ASCR + kNN	0.200	0.978	11.47	0.333	0.967	10.87
ASCR + RF	0.202	0.959	9.39	0.329	0.953	6.06

B. Structural probability control

To quantify the calibration–efficiency geometry of direct score contraction, the structural control evaluates

$$\widetilde{p}_\rho(y \mid X) = \rho(X)p_T(y \mid X) + \frac{1 - \rho(X)}{K}. \tag{29}$$

It satisfies

$$\|\widetilde{p}_\rho - u\|_1 = \rho\|p_T - u\|_1, \tag{30}$$

$$\widetilde{p}_{\rho,i} - \widetilde{p}_{\rho,j} = \rho(p_{T,i} - p_{T,j}). \tag{31}$$

Thus ranking is unchanged for $\rho > 0$. Decreasing ρ moves the vector toward uniformity and increases entropy by majorization. At a fixed conformal threshold, sets weakly expand; when the transformed score is recalibrated, its own quantile must be used. All mixture results in this paper use independent recalibration. Paired effects over ten seeds quantify the ranking-preserving calibration–efficiency trade-off and support the separate action-layer construction used in the main text.

Across ten independent repetitions the contracted score changes mean corrupted coverage by +0.043 and mean set size by +1.90. ECE changes by +0.078. Macro-F_1 remains unchanged because the class ranking is preserved.

C. Mechanism, calibration, and efficiency analyses

Exact masks isolate the image-processing mechanism. Figure 8 shows higher adaptive-mask IoU on clean, hue-shifted, and combined crops and identifies gain normalization as the largest combined-corruption contributor. The severity analysis separates gain-invariant low-light behavior from prototype-sensitive hue rotation, while the reliability analysis agrees with the ECE trends in Table II.

Exponent perturbations change combined-corruption coverage by at most 0.028 and mean set size by at most 0.31 classes. CPU latency is 13.49 ms/image for ASCR+kNN and 14.53 ms/image for ASCR+RF. The set-size trigger has precision 0.408 and recall 0.807 for incorrect argmax predictions; the stricter $\rho \le 0.10$ trigger has precision 0.897 and recall 0.212. Set size emphasizes broad recall, while explicit evidence departure produces a selective, high-precision intervention.

D. Class-domain and classical controls

The class-domain results are reported in Table III. Their head-independent pattern locates the handcrafted descriptor in color and contour variation, while the semantic model carries fine printed-symbol discrimination.

The HOG–linear-SVM control reaches 0.832 clean macro-F_1. Its evidence-intervention response is close to the residual model because ρ is deliberately independent of the semantic head. This supports interpreting the visual channel as an image-domain monitor, while semantic performance remains dependent on the recognition model.

Across all six public-data conditions, the HOG control exhibits the same qualitative evidence response as the residual model. The low-contrast intervention therefore follows the image specification itself and remains stable across distinct semantic heads.

E. Proof details and numerical checks

For the context requirement, fix a semantic coordinate $S = s$ and choose $\rho_1 < \rho_2$ in the common support. A fusion monotone increasing in ρ orders $f(s, \rho_1) \le f(s, \rho_2)$ in every environment; a monotone decreasing score imposes the reverse order everywhere. Opposite signs of $\beta_\rho(c_1)$ and $\beta_\rho(c_2)$ require opposite Bayes risk orderings for the

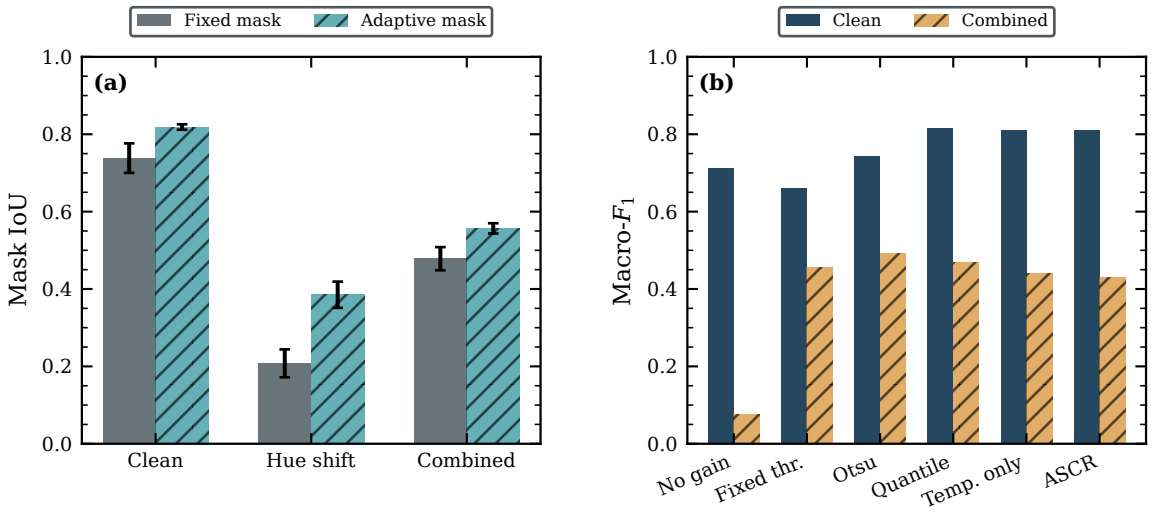


Fig. 8. Supporting mechanism analyses. (a) Fixed and adaptive mask IoU. (b) Component ablation on clean and combined crops. Error bars report seed variation where available.

same pair, so a global monotone direction conflicts with one of the two environment-conditioned orderings.

For (17), decompose $Q(A) = Q(A \cap G_\tau) + Q(A \cap G_\tau^c)$. The density-ratio condition gives $Q(A \cap G_\tau) \le \kappa P(A \cap G_\tau) \le \kappa\alpha$, and $Q(A \cap G_\tau^c) \le Q(G_\tau^c) = \eta_Q$. Summing the terms yields the stated bound.

Data Availability

The German Traffic Sign Recognition Benchmark (GTSRB) images and annotations used in this study are publicly available from the German Traffic Sign Benchmarks website at https://benchmark.ini.rub.de/gtsrb_dataset.html. An archival copy of the official GTSRB training and test files is publicly available at https://sid.erda.dk/public/archives/daaeac0d7ce1152aea9b61d9f1e19370/published-archive.html. The GTSRB analyses can be independently verified from these public data by following the experimental protocol described in the manuscript. SignStress-12 is a procedurally generated controlled dataset. Its analyses can be regenerated from the construction and transformation procedures described in the Methods. No private data were used.

Conflict of Interest

The author declares no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgment

The author acknowledges publication support from the Deanship of Graduate Studies and Scientific Research at Qassim University (www.qu.edu.sa) under QU-APC-2026.

Author Contributions

A. S. Alturki: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Resources, Writing – original draft, Writing – review & editing, Project administration, and Funding acquisition.

Funding

The research reported in this article received no dedicated research funding. Publication support was provided by the Deanship of Graduate Studies and Scientific Research at Qassim University under QU-APC-2026.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[3] Z. Liu, H. Mao, C.-Y. Wu *et al.*, "A convnet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[4] Y. Ovadia, E. Fertig, J. Ren *et al.*, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[5] M. Minderer, J. Djolonga, R. Romijnders *et al.*, "Revisiting the calibration of modern neural networks," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 15 682–15 694.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 2017, pp. 1321–1330.

[7] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48. PMLR, 2016, pp. 1050–1059.

[8] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6402–6413.

[9] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[10] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3581–3591.

[11] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," in *International Conference on Learning Representations*, 2021.

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[13] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[14] K. He, X. Chen, S. Xie *et al.*, "Masked autoencoders are scalable vision learners," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[15] M. Oquab, T. Darcet, T. Moutakanni *et al.*, "DINOv2: Learning robust visual features without supervision," 2023.

[16] A. Kirillov, E. Mintun, N. Ravi *et al.*, "Segment anything," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[17] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, vol. 32, pp. 323–332, 2012.

[18] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society: Series B*, vol. 13, no. 2, pp. 238–241, 1951.

[19] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497, 2012.

[20] S. B. Wali, M. A. Abdullah, M. A. Hannan *et al.*, "Vision-based traffic sign detection and recognition systems: Current trends and challenges," *Sensors*, vol. 19, no. 9, p. 2093, 2019.

[21] A. S. Alturki, "Traffic sign detection and recognition using adaptive threshold segmentation with fuzzy neural network classification," in *2018 International Symposium on Networks, Computers and Communications (ISNCC)*, 2018, pp. 1–7.

[22] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[23] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[26] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *International Joint Conference on Neural Networks*, 2011, pp. 2809–2813.

[27] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333–338, 2012.

[28] D. Tabernik and D. Skocaj, "Deep learning for large-scale traffic-sign detection and recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 4, pp. 1427–1440, 2020.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, vol. 28, 2015, pp. 91–99.

[30] W. Liu, D. Anguelov, D. Erhan *et al.*, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.

[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[32] T.-Y. Lin, P. Goyal, R. Girshick *et al.*, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*, 2017, pp. 2999–3007.

[33] S. Houben, J. Stallkamp, J. Salmen *et al.*, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in *International Joint Conference on Neural Networks*, 2013, pp. 1–8.

[34] Z. Zhu, D. Liang, S. Zhang *et al.*, "Traffic-sign detection and classification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2110–2118.

[35] D. Temel, G. Kwon, M. Prabhushankar, and G. AlRegib, "CURE-TSR: Challenging unreal and real environments for traffic sign recognition," in *NeurIPS Workshop on Machine Learning for Intelligent Transportation Systems*, 2017.

[36] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations*, 2017.

[37] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 948–10 957.

[38] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.

[39] J. Huang, H. Xi, L. Zhang *et al.*, "Conformal prediction for deep classifier via label ranking," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 20 331–20 347.

[40] S. Kiyani, G. J. Pappas, and H. Hassani, "Conformal prediction with learned features," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 24 749–24 769.

[41] G. S. Dhillon, G. Deligiannidis, and T. Rainforth, "On the expected size of conformal prediction sets," in *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 238. PMLR, 2024, pp. 1549–1557.

[42] S. H. Zargarbashi, M. S. Akhondzadeh, and A. Bojchevski, "Robust yet efficient conformal prediction sets," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 17 123–17 147.

[43] A. N. Angelopoulos, S. Bates, A. Fisch *et al.*, "Conformal risk control," in *International Conference on Learning Representations*, 2024.

[44] K. Kasa, Z. Zhang, H. Yang, and G. W. Taylor, "Adapting prediction sets to distribution shifts without labels," in *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 286. PMLR, 2025, pp. 1990–2010.

[45] R. J. Tibshirani, R. F. Barber, E. J. Candès, and A. Ramdas, "Conformal prediction under covariate shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 2526–2536.

[46] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.

[47] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4878–4887.

[48] ——, "Selectivenet: A deep neural network with an integrated reject option," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 2151–2159.

[49] C. Corbière, N. Thome, A. Bar-Hen *et al.*, "Addressing failure prediction by learning model confidence," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[50] L. F. Cattelan and D. Silva, "How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks," in *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, vol. 244. PMLR, 2024, pp. 547–584.

[51] Y.-C. Wu, S.-H. Lyu, H. Shang *et al.*, "Confidence-aware contrastive learning for selective classification," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 53 706–53 729.

[52] A. García-Galindo, M. López-De-Castro, and R. Armañanzas, "Multi-class classification with reject option and performance guarantees using conformal prediction," in *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, ser. Proceedings of Machine Learning Research, vol. 230. PMLR, 2024, pp. 295–314.

[53] D. Boursinos and X. Koutsoukos, "Assurance monitoring of learning enabled cyber-physical systems using inductive conformal prediction based on distance learning," 2021.

[54] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[55] S. Pavlitska, B. Keskin, A. Faßbender *et al.*, "Extracting uncertainty estimates from mixtures of experts for semantic

segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2025, pp. 311–320.
[56] R. Dunn, L. Wasserman, and A. Ramdas, "Distribution-free prediction sets for two-layer hierarchical models," *Journal of the American Statistical Association*, vol. 118, no. 544, pp. 2491–2502, 2023.
[57] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023.
[58] R. I. Oliveira, P. Orenstein, T. Ramos, and J. V. Romano, "Split conformal prediction and non-exchangeable data," *Journal of Machine Learning Research*, vol. 25, no. 225, pp. 1–38, 2024.
[59] H. Zeng, K. Liu, B. Jing, and H. Wei, "Parametric scaling law of tuning bias in conformal prediction," in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 74 133–74 156.
[60] F. Wu, L. Niu, S. Kaski, and M. Caprio, "Decoupled conformal optimisation: Efficient prediction sets via independent tuning and calibration," 2026, arXiv:2605.18354.
[61] G. Buchsbaum, "A spatial processor model for object colour perception," *Journal of the Franklin Institute*, vol. 310, no. 1, pp. 1–26, 1980.
[62] W. Niblack, *An Introduction to Digital Image Processing*. Prentice Hall, 1986.
[63] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Wiley, 1999.
[64] A. Rosenfeld and J. L. Pfaltz, "Sequential operations in digital picture processing," *Journal of the ACM*, vol. 13, no. 4, pp. 471–494, 1966.
[65] E. P. Klement, R. Mesiar, and E. Pap, *Triangular Norms*. Dordrecht: Kluwer Academic Publishers, 2000.
[66] S. Zhao, M. P. Kim, R. Sahoo *et al.*, "Calibrating predictions to decisions: A novel approach to multi-class calibration," in *Advances in Neural Information Processing Systems*, vol. 34, 2021.
[67] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[68] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[69] J. Traub, T. J. Bungert, C. T. Lüth *et al.*, "Overcoming common flaws in the evaluation of selective classification systems," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 2323–2347.
[70] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[71] T. Saito and M. Rehmsmeier, "The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015.
[72] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927.
[73] R. Isermann, "Model-based fault-detection and diagnosis—status and applications," *Annual Reviews in Control*, vol. 29, no. 1, pp. 71–85, 2005.