Assessing Calibration and Conformal Coverage of SAR Perception Models under Sim-to-Real Shift

Diego Argüello Ron[1,*], Christyan Cruz Ulloa[2], Naresh Ravichandran[3], Anders Lansner[3], Pawel Herman[3], Jaime del Cerro[2] and Oscar García Perales[1]

[1]*Data Analytics for Industries 4.0 (i4RI), Xàtiva, Spain*
[2]*Centro de Automática y Robótica (UPM–CSIC), Universidad Politécnica de Madrid, Madrid, Spain*
[3]*Computational Science and Technology, EECS and Digital Futures, KTH Royal Institute of Technology, Stockholm, Sweden*

Abstract

Search-and-rescue (SAR) robots are often trained or validated in simulation before deployment in real environments, a setting relevant to high-risk AI transparency and robustness obligations under the EU AI Act. In this regime, accuracy alone is an incomplete trust signal: a model may be correct often enough but still report misleading confidence or silently avoid difficult parts of the scene. We study this risk on an IsaacSim-to-real SAR benchmark using two deployed perception models: a brain-inspired Bayesian Confidence Propagation Neural Network (BCPNN) terrain classifier and a YOLOv8m-seg segmenter. We evaluate three forms of uncertainty evidence: raw calibration, measured by expected calibration error (ECE, the average gap between confidence and observed accuracy); post-hoc calibration, which tests whether confidence can be corrected after training; and conformal prediction, which tests whether prediction sets achieve controlled coverage. The two models reveal different hidden failures. BCPNN is overconfident in both domains and degrades modestly under sim-to-real shift, with ECE increasing from 0.385 to 0.403. YOLO appears to improve, with ECE dropping from 0.448 to 0.035, but this is an abstention artefact: on real-domain images the model labels only $1.0\,\%$ of pixels (down from $6.0\,\%$ in simulation), so the low ECE reflects a small high-confidence subset rather than the full scene. Post-hoc calibration also behaves differently across architectures: isotonic regression reduces BCPNN's real-domain ECE by $82\,\%$, while temperature scaling inflates YOLO's real-domain ECE by $18\times$. For BCPNN, standard split and weighted conformal prediction always return the full set of classes, making them uninformative; adaptive conformal inference, by contrast, still produces useful prediction sets at every tested level. Read against EU AI Act Art. 13/15 and NIST AI 800-3, these results support joint reporting of accuracy, calibration, emission coverage, and conformal set behaviour, with model-specific validation of uncertainty methods. Operationally, this helps prevent acceptance based on task accuracy alone when confidence is misleading or predictions cover only a small part of the scene.

Keywords

calibration, conformal prediction, sim-to-real, search and rescue robotics, uncertainty quantification, EU AI Act

1. Introduction

Operational motivation. Search-and-rescue (SAR) robots increasingly rely on perception models trained or validated in simulation before deployment in physical environments [1, 2, 3]. In this setting, model confidence is not merely a diagnostic quantity: it affects how operators, safety supervisors, and downstream autonomy modules interpret uncertain terrain or obstacle predictions, so an often-correct but systematically overconfident model can still create operational risk or warrant additional validation. This concern aligns with the EU AI Act's emphasis on transparency for interpreting system outputs (Art. 13) and on appropriate accuracy and robustness throughout the system lifecycle (Art. 15) [4, 5], and with recent NIST guidance on uncertainty quantification for AI benchmarks [6, 7]. SAR perception on a quadruped operating in a NIST-Orange-Zone-style disaster arena [8] is therefore a high-risk-relevant safety-critical AI setting. Depending on intended use, integration, and deployment context, such a system may fall within high-risk obligations under Article 6 of the same Regulation, either as a safety component of a product covered by EU harmonisation legislation or through an Annex III use case [4].

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.
*Corresponding author.
diego.arguello@i4ri.com (D. A. Ron); christyan.cruz.ulloa@upm.es (C. C. Ulloa); nbrav@kth.se (N. Ravichandran); ala@kth.se (A. Lansner); paherman@kth.se (P. Herman); j.cerro@upm.es (J. d. Cerro); oscar.garcia@i4ri.com (O. G. Perales)

We therefore treat Articles 13 and 15 as design-relevant disclosure targets rather than asserting a final legal classification of the research prototype.

Conceptual gap. Most sim-to-real perception evaluations still focus on task metrics such as accuracy, intersection-over-union, or mAP. These metrics answer whether a prediction is correct on the evaluated subset, but they do not answer three trustworthiness questions that matter in deployment. First, do confidence scores correspond to empirical correctness? If the model reports 0.9 confidence, is it right about $90\,\%$ of the time? Second, does the model silently abstain from difficult parts of the scene, so that headline metrics are conditional on a shrunk subset rather than the full input? Third, can uncertainty estimates be converted into prediction sets whose coverage of the true label can be controlled? We therefore evaluate calibration, emission coverage (the fraction of the input on which the model actually labels), and conformal coverage as complementary evidence rather than as replacements for standard performance metrics.

Study design. We study two existing SAR perception components on a shared IsaacSim-to-real benchmark (1,240 images, four terrain classes): a brain-inspired Bayesian Confidence Propagation Neural Network (BCPNN) terrain classifier and a YOLOv8m-seg segmentation model, both deployed in the SAR pipeline of Cruz Ulloa et al. [3]. We do not treat the two models as a direct leaderboard comparison, because their output spaces and scoring pools differ. Instead, they are used as diagnostic case studies: the central question is whether conventional performance metrics hide different trustworthiness failures under sim-to-real shift, and whether commonly-recommended uncertainty-correction methods (post-hoc calibration, split/weighted/adaptive conformal prediction) recover trustworthy behaviour for both. All headline numbers carry $10{,}000$-resample bootstrap $95\,\%$ confidence intervals.

Contributions. The paper makes four evidence-backed contributions:

- **C1. First (to our knowledge) calibration characterisation of BCPNN under sim-to-real shift.** BCPNN remains moderately accurate but becomes systematically more overconfident on real data; the patch-level bootstrap CI for ΔECE excludes zero, while the image-clustered sensitivity interval includes zero (Table 1, Section 4.1).
- **C2. Post-hoc calibration is not architecture-agnostic.** Isotonic regression transfers cleanly from simulation to real for BCPNN ($-82\,\%$ real-domain ECE), while temperature scaling severely inflates YOLO's real-domain ECE ($18\times$) because the deployment confidence distribution changes shape (Table 2, Section 4.4).
- **C3. Conformal behaviour of atomic-posterior classifiers.** Because BCPNN's posterior scores are highly concentrated, standard split and weighted conformal prediction return the full class set for every test example: coverage is valid, but the sets are uninformative. ACI is the only tested method that remains informative at all tested levels $\alpha \in \{0.05, 0.10, 0.20\}$; ConfTS becomes informative at $\alpha \geq 0.10$ (Table 3, Section 4.5).
- **C4. Calibration must be interpreted together with emission coverage.** YOLO's apparent ECE improvement on real imagery is caused by a $6\times$ drop in emission coverage: the model labels only $1\,\%$ of real-domain pixels. An accuracy-, mAP-, or ECE-only evaluation could therefore accept a detector whose reported performance is measured on a small emitted subset rather than on the full scene (Fig. 2, Section 4.2).

Read against EU AI Act Art. 13/15 [4] and NIST benchmark-evaluation guidance [9, 6], the operational consequence is a deployment-oriented reporting framework: SAR perception evaluations should report accuracy, calibration, emission coverage, and conformal set behaviour together, and no single uncertainty-quantification method should be assumed reliable under sim-to-real shift without model-specific validation. The rest of the paper develops this assessment. Section 2 positions the methods used; Section 3 describes the models, data, and statistical protocol; Section 4 presents the raw-calibration, post-hoc calibration, and conformal prediction results; and Section 5 maps the findings to the regulatory context.

2. Background and Related Work

This section connects four operational trustworthiness concepts (calibration, post-hoc correction, emission coverage, and conformal coverage) to sim-to-real SAR perception. Formal metric definitions are deferred to Section 3.2.

2.1. Calibration, correction, and emission coverage

Standard task metrics (accuracy, IoU, mAP) measure whether predictions are correct but not whether confidence scores are meaningful. Calibration addresses the second question: a well-calibrated classifier assigning, e.g., $80\,\%$ confidence is correct about $80\,\%$ of the time. Accuracy says how often the model is right; calibration says whether the model represents its own uncertainty. Expected calibration error (ECE) [10] summarises the gap between predicted confidence and empirical correctness across bins; reliability diagrams provide the visual diagnostic (curve below diagonal = overconfident). For detection and segmentation, Küppers et al. [11, 12] proposed D-ECE and Kuzucu et al. [13] proposed LaECE_0; continuous-IoU extensions remain active [14].

Once miscalibration is measured, a natural question is whether it can be corrected without retraining the perception model. Post-hoc calibration adjusts confidence after training. Temperature scaling [10] fits a single scalar to logits; Platt scaling [15] fits a per-class sigmoid; isotonic regression [16] fits a per-class non-parametric monotone map. Temperature preserves argmax; Platt and isotonic can re-rank. Such scalers can substantially improve calibration when calibration and deployment distributions are compatible. Under domain shift this assumption may fail: Ovadia et al. [17] showed temperature scaling degrades under shift, while Kuzucu et al. [13] reported that class-wise isotonic remained robust across domains.

For detection and segmentation, calibration must also be read together with the fraction of the scene on which the model emits a prediction. Otherwise a low ECE on a tightly selected high-confidence subset can mask operational incompleteness. We call this quantity *emission coverage* to distinguish it from the conformal coverage discussed next. Reject-option and selective-prediction methods [18, 19, 20] make the coverage / conditional-risk trade-off explicit; in the YOLO setting studied here it is made implicitly by the detector's confidence threshold and is therefore easy to miss in a single-metric report.

2.2. Conformal prediction and set informativeness

Conformal prediction [21, 22] provides a complementary form of uncertainty quantification. Instead of trying to make confidence values themselves reliable, conformal methods construct prediction sets intended to contain the true label with probability at least $1 - \alpha$ under suitable exchangeability assumptions. For classification, Romano et al. [23] introduced adaptive prediction sets (APS), which build instance-dependent sets from the model's sorted class probabilities. Tibshirani et al. [24] extended the marginal-coverage guarantee under covariate shift via density-ratio weighting; Gibbs and Candès [25] introduced adaptive conformal inference (ACI), which treats α_t as an online control variable that adapts to the observed empirical miscoverage rate; Xi et al. [26] combined temperature scaling with APS (ConfTS) to minimise mean set size; Luo et al. [27] applied conformal methods to robot safety assurances.

Coverage alone, however, is not sufficient. A method can achieve valid coverage by returning very large sets; in the extreme, returning all K classes always gives perfect coverage but no useful information. Conformal prediction must therefore be evaluated through coverage *and* mean set size. This distinction becomes important when a classifier's score distribution is highly concentrated, because valid conformal sets may then become uninformatively large. Section 4.5 returns to this point on the BCPNN posterior.

2.3. SAR sim-to-real perception and prior work

BCPNN [28, 29, 30] is a modular Bayesian-Hebbian recurrent network with softmax winner-take-all (WTA) competition within each hypercolumn. In this architecture, hypercolumns represent discrete variables and minicolumn activities provide probability-like scores over competing states. These outputs are therefore useful for uncertainty analysis, but probability-like scores are not automatically calibrated empirical probabilities, especially after sim-to-real shift. We are not aware of prior published evaluations of BCPNN calibration error in a sim-to-real SAR setting.

Sim-to-real quadruped perception evaluations [2, 1] are still dominated by mAP and top-1 accuracy. The SAR pipeline assessed here was released by Cruz Ulloa et al. [3], who reported task accuracy and identified uncertainty quantification as future work; this leaves open whether the pipeline's confidence scores can support thresholding, abstention analysis, or uncertainty communication during deployment. An earlier TRUST-AI workshop paper by some of the present authors applied XAI and D-ECE [11] to an earlier YOLOv8m-seg checkpoint. The present paper extends these studies by evaluating calibration and conformal behaviour for BCPNN and YOLOv8m-seg, and by interpreting the results in relation to transparency and uncertainty-reporting needs under EU AI Act Art. 13/15 [4] and NIST benchmark-evaluation guidance [9, 6].

3. Experimental Setup

3.1. Models and benchmark

The first model is a pretrained **BCPNN terrain classifier** released with the sim-to-real pipeline of Cruz Ulloa et al. [3]. Its architecture is a three-layer Bayesian Confidence Propagation Neural Network [28, 30] with 256×20 input hypercolumns (GMM-whitened 64×64 RGB patches), 100×100 hidden hypercolumns, and a 1×5 output layer (4 terrain classes: *ground*, *grass*, *gravel*, *wood*; plus one background readout). Inference uses 3 recurrent loops and softmax winner-take-all within each hypercolumn.

The second model is a **YOLOv8m-seg segmenter** trained on a 9-class NIST-Orange-Zone-adjacent SAR dataset [8] (classes: NS, S100, S50, S75, *concrete*, *grass*, *gravel*, victim, *wood*). We deliberately evaluate the native 9-class model rather than a filtered 4-class variant: a serial probe of three candidate single-task checkpoints confirmed that none matches the BCPNN label set exactly, and the most abundant YOLO emission on our imagery is the SAR-specific *S100* class. We do not include MC Dropout, Deep Ensembles, or Laplace baselines [31, 32, 17]; scope is bounded to models already deployed in our SAR pipeline, and a broader Bayesian baseline sweep is future work (Section 6).

The evaluation dataset [33] (1,240 RGB images at 1280×720, Pascal-VOC polygon labels) spans two domains. $\mathrm{D}_{\mathrm{SIM}}$ (`d1`; IsaacSim photoreal renders of a disaster-arena USD scene) provides $349/104/52$ train/val/test images. $\mathrm{D}_{\mathrm{REAL}}$ (`d2`; photographs of the same arena instantiated physically) provides $519/144/72$. For the patch-level BCPNN pipeline we tile each image into 64×64 patches and retain those with $\geq 80\%$ single-class polygon coverage on one of the four terrain classes, yielding $8{,}058$ (d1_test) and $11{,}413$ (d2_test) classifiable patches.

3.2. Scoring units, metrics, and statistics

The two models are evaluated on their native output units. BCPNN is scored on retained 64×64 terrain patches (one class-probability vector each). YOLO is scored at pixel level only where the predicted mask exceeds the threshold and the ground-truth class is in the name-matchable terrain subset (*grass*, *gravel*, *wood*); pixels below threshold are excluded from accuracy and ECE but counted in emission coverage, so YOLO's accuracy and ECE are *conditional-on-emission* metrics. SAR-specific classes without a terrain-name match (e.g., *S100*) are retained for the emission-distribution analysis (Appendix C) but not scored as correct terrain predictions.

We compute accuracy, ECE (15 equal-width bins [10]; for BCPNN, ECE and the sim-to-real ΔECE are insensitive to the bin count over $[5, 30]$, Appendix I), Brier, and NLL. Main tables report accuracy/E-

Table 1
Raw-model calibration under sim-to-real shift. Primary CIs use paired-sample patch-level bootstrap (10k); ‡ rows use image-clustered bootstrap (10k) as a sensitivity check, resampling $52/72$ source images for d1/d2 (point estimates unchanged; CIs widen by $\sim 3\times$). YOLO uses image-level bootstrap (10k). Δ columns are d2 $-$ d1 with bootstrap 95 % CIs. All ECE and Brier values are probabilities in $[0, 1]$; lower is better. YOLO entries are conditional-on-emission: pred-mask coverage is $6.0\,\%$ on d1_test and $1.0\,\%$ on d2_test (see Section 4.2).

Model	Split	Accuracy	ECE	Brier
BCPNN	d1_test	0.604 [0.594, 0.615]	0.385 [0.374, 0.396]	0.775 [0.754, 0.796]
	d2_test	0.581 [0.573, 0.590]	0.403 [0.394, 0.412]	0.816 [0.798, 0.833]
	Δ (d2 $-$ d1)	-0.023 [-0.037, -0.009]	$+0.018$ [$+0.005$, $+0.032$]	$+0.041$ [$+0.014$, $+0.068$]
	Δ (d2 $-$ d1)‡	-0.023 [-0.064, $+0.018$]	$+0.018$ [-0.022, $+0.059$]	$+0.041$ [-0.041, $+0.121$]
YOLOv8m-seg	d1_test	0.540 [0.425, 0.733]	0.448 [0.266, 0.572]	0.899
	d2_test	0.985 [0.969, 0.998]	0.035 [0.007, 0.058]	0.041
	Δ (d2 $-$ d1)	$+0.445$ [$+0.255$, $+0.561$]	-0.414 [-0.541, -0.233]	—

‡ Image-clustered bootstrap ($n_{\text{img}} = 52/72$): point estimates unchanged; Δ-ECE CI now *includes zero*, so the sim-to-real ECE shift is not significance-robust to within-image dependence at the present arena-image counts.

CE/Brier; NLL is quoted inline where relevant. For YOLO we additionally report *emission coverage* (the fraction of pixels for which the predicted mask exceeds the threshold) and reserve *conformal coverage* for the fraction of examples whose true label is contained in a conformal prediction set. LaECE$_0$ [13] reduces algebraically to ECE under our discrete-pixel proxy $\mathrm{loc}_i = \mathbb{1}[\arg\max_i = y_i]$, so we report both labels for continuity with detection-calibration literature and leave continuous-IoU LaECE$_0$ to future work (Section 6). The BCPNN head emits five probabilities but the published `testlbls.bin` has zero class-0 samples; calibration is reported on the four-class filtered pool as the primary statistic, with the unfiltered comparison in Appendix A.

All headline numbers carry $95\,\%$ bootstrap CIs (10,000 resamples). BCPNN primary CIs use paired-sample patch-level resampling (a first-order i.i.d. approximation after GMM whitening); image-clustered CIs ($52/72$ source images for d1/d2) are reported in Table 1 (‡ rows) as a within-image-dependence sensitivity check, widening by $\sim 3\times$. YOLO uses image-level resampling, which avoids treating spatially correlated pixels as independent. A float-precision artefact in the APS cumulative sum was fixed mid-study (Appendix G).

3.3. Conformal protocol

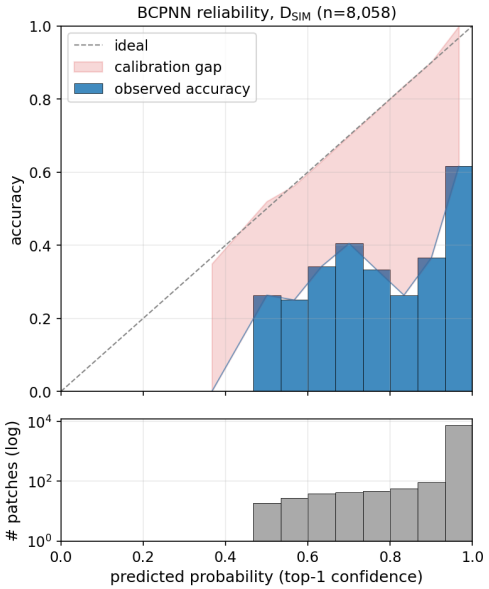
We calibrate on $\mathrm{D}_{\mathrm{SIM}}$ validation (d1_val) and evaluate on d1_test and d2_test at coverage levels $\alpha \in \{0.05, 0.10, 0.20\}$. BCPNN calibration-fit / test pool sizes are 16,673 / 8,058 / 11,413 for d1_val / d1_test / d2_test; YOLO pixel-level pool sizes are 651,621 / 544,015 / 217,528.

Four methods are compared: *split conformal* with adaptive prediction sets [23]; *weighted conformal* [24] with density-ratio weights from an L_2-regularised logistic domain classifier ($C = 1$) on per-image $(\bar{R}, \bar{G}, \bar{B}, \sigma_R, \sigma_G, \sigma_B)$ features broadcast to all 220 patches per image, with weights clipped to $[0.01, 100]$; *ConfTS* [26], which tempers the softmax by T before scoring; and *adaptive conformal inference* (ACI) [25] with $\gamma = 0.01$ and warmup of 1000 steps on the d2_test stream.[1] The clip bounds and ACI step size are non-critical over the tested ranges (Appendix I).

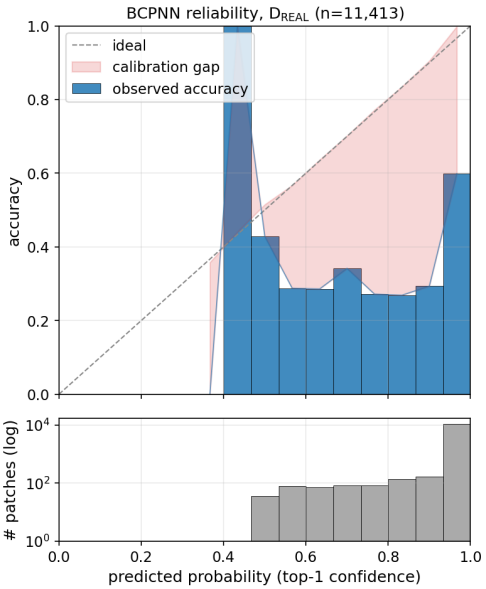
4. Results: Calibration Failures and Correction under Shift

We first report raw-model calibration, before any correction, and then evaluate whether post-hoc calibration or conformal prediction recovers reliable uncertainty behaviour. Raw calibration numbers are summarised jointly in Table 1; reliability diagrams are shown in Figures 1–2.

[1] The domain classifier achieves cross-validated AUROC 0.999 (image-level shift is identifiable); the fitted weights have a narrow range (median 0.066, max 0.403).



(a) D_{SIM} (d1_test): acc 0.604, ECE 0.385, Brier 0.775 ($n = 8{,}058$).



(b) D_{REAL} (d2_test): acc 0.581, ECE 0.403, Brier 0.816 ($n = 11{,}413$).

Figure 1: BCPNN reliability diagrams. The accuracy curve sits consistently below the identity diagonal across both domains, reflecting systematic overconfidence; D_{REAL} is modestly worse.

4.1. BCPNN: uniform degradation

BCPNN is already miscalibrated in-domain: on D_{SIM} the expected calibration error is 0.385 with bootstrap 95 % CI $[0.374, 0.396]$, indicating substantial overconfidence at accuracy 0.604. Domain shift to D_{REAL} shifts ECE to 0.403 $[0.394, 0.412]$. Patch-level paired bootstrap gives $\Delta\text{ECE} = +0.0183$ with CI $[+0.0046, +0.0324]$; the image-clustered bootstrap (Table 1, ‡ row) widens this CI to $[-0.0223, +0.0592]$ and *includes zero*, so the shift is robust under the i.i.d. patch-pool assumption but not significance-resolvable once within-image dependence is preserved at the present arena-image counts. Brier and NLL degrade commensurately ($+0.041$ and $+0.42$, patch-level); accuracy drops -2.3 percentage points.

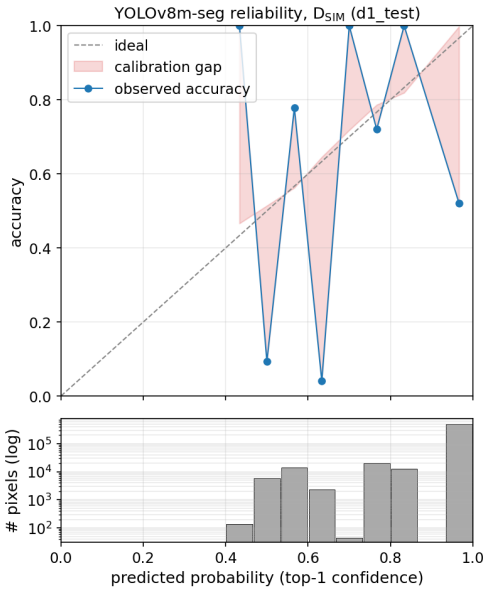
Per-class one-vs-rest reliability is shown in Appendix F; *ground* carries the largest sim-to-real ECE shift. Figure 1 plots the reliability diagrams on both splits.

4.2. YOLO: abstention-masked "improvement"

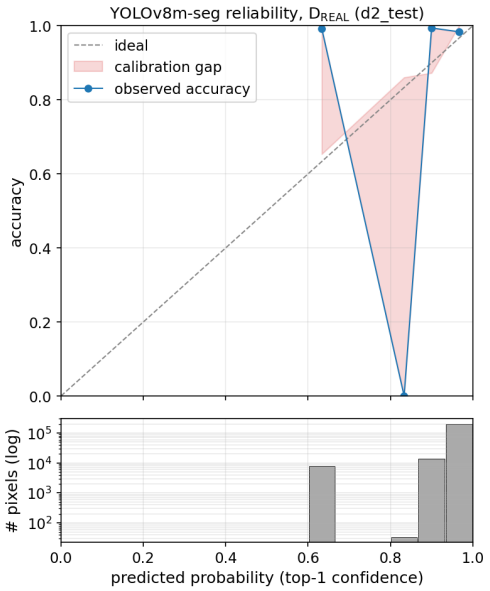
YOLO's d2_test ECE is 0.035 $[0.007, 0.058]$ versus d1_test ECE 0.448 $[0.266, 0.572]$ ($\Delta\text{ECE} = -0.414$ $[-0.541, -0.233]$). *This apparent improvement is an abstention artefact, not evidence that YOLO is a calibrated classifier on real imagery.* On D_{SIM} YOLO's predicted mask exceeds the confidence threshold on 6.0 % of pixels; on D_{REAL} only 1.0 % (a 6× drop). On the retained D_{REAL} pixels accuracy is 98.5 % and confidence is concentrated near 1.0, so ECE stays low because the retained predictions are concentrated in a high-confidence subset while the rest of the scene is unlabelled. Figure 2 shows the bimodal D_{SIM} curve against the nearly degenerate D_{REAL} curve; per-image overlays are in Appendix B.

4.3. Synthesis: two silent failure modes

Both models could appear acceptable under a conventional task-specific acceptance gate (BCPNN near 0.58 patch accuracy, YOLO 0.98 on the pixels it labels), but only the joint reporting of ECE *and* emission coverage separates a uniformly-degraded Bayesian classifier from an abstention-biased detector. This is the paper's first main finding: under sim-to-real shift, calibration fails in architecturally distinct ways that are *silent* to headline accuracy.



(a) D_{SIM} (d1_test): acc 0.540, ECE 0.448 (n = 544,015 pixels). Bimodal confidence.



(b) D_{REAL} (d2_test): acc 0.985, ECE 0.035 (n = 217,528 pixels). Peaked-at-1 confidence on 1 % pixel coverage.

Figure 2: YOLO pixel-level reliability diagrams. The "improvement" in ECE from D_{SIM} to D_{REAL} is accompanied by a 6× drop in prediction coverage: most REAL pixels fall below the detector's confidence threshold and do not enter the scoring pool.

Table 2
Post-hoc calibration fit on d1_val, transferred to both test splits. BCPNN values carry paired-bootstrap 95 % CIs; YOLO entries are point estimates at large pixel-level n. Bold marks the lowest ECE per model-split cell; the YOLO d2_test uncalibrated entry is bolded with the reminder that this low ECE is the abstention regime of Fig. 2b.

Model	Scaler	d1_test acc	d1_test ECE	d2_test acc	d2_test ECE
BCPNN	uncal.	0.604	0.385 [0.374, 0.395]	0.581	0.403 [0.394, 0.412]
	temperature	0.604	0.055 [0.045, 0.065]	0.581	0.067 [0.062, 0.076]
	Platt	0.685	0.046 [0.037, 0.055]	0.650	0.073 [0.065, 0.081]
	isotonic	**0.689**	**0.044** [0.039, 0.055]	**0.667**	**0.071** [0.064, 0.080]
YOLO	uncal.	0.540	0.448	0.985	**0.035**†
	temperature	0.540	0.201	0.985	0.642
	Platt	0.566	**0.301**	0.985	0.245
	isotonic	0.570	0.301	0.985	0.248

† Abstention-regime ECE (pred-mask coverage 1.0 %); not a calibrated classifier on real imagery (see Section 4.2).

On commensurability. The two models sit on no shared metric scale—BCPNN per 64 × 64 patch (a forced five-class choice), YOLO per pixel behind an implicit-threshold abstention (conditional-on-emission)—so the "unified" claim is methodological, not a head-to-head ranking (Appendix J).

4.4. Post-hoc calibration: model-dependent transfer under shift

We fit temperature scaling [10], Platt scaling [15], and class-wise isotonic regression [16] on the D_{SIM} validation pool and transfer each scaler unmodified to the two test splits. Results appear in Table 2.

BCPNN: isotonic transfers cleanly sim-to-real. On D_{SIM}, isotonic reduces ECE from 0.385 to 0.044 (88 % reduction, disjoint CIs); transferred unmodified to D_{REAL} it drops ECE from 0.403 to 0.071 (−82 %, CIs still disjoint). Platt performs comparably; fitted temperature $\hat{T} = 12.01$ is large because it scales log-probabilities, but the univariate form still captures most of the miscalibration. Platt

Table 3
Conformal prediction on BCPNN posteriors with APS scores. Target coverage is $1-\alpha$; $K=5$ classes. Split conformal and weighted conformal both collapse to the trivial full-set predictor because $87\,\%$ of calibration scores lie within 10^{-7} of the atom $s=1$, so the quantile τ snaps to 1 for every α tested. ConfTS partially breaks the atom at $\alpha \geq 0.10$; ACI attains target empirical online coverage across the full range.

Method	Split	$\alpha = 0.05$ Cov / Size	$\alpha = 0.10$ Cov / Size	$\alpha = 0.20$ Cov / Size
Split CP (APS)	d1_test	1.00 / 5.00	1.00 / 5.00	1.00 / 5.00
Split CP (APS)	d2_test	1.00 / 5.00	1.00 / 5.00	1.00 / 5.00
Weighted CP	d2_test	1.00 / 5.00	1.00 / 5.00	1.00 / 5.00
ConfTS ($\hat{T}=12.01$)	d1_test	1.00 / 5.00	0.862 / 4.02	0.788 / 2.59
ConfTS ($\hat{T}=12.01$)	d2_test	1.00 / 5.00	0.930 / 4.03	0.794 / 2.53
ACI ($\gamma=0.01$)	d2_test	**0.957 / 4.22**	**0.907 / 3.52**	**0.806 / 2.53**

and isotonic also improve accuracy by $+8.5/+8.6$ pp (CIs disjoint from uncalibrated), since per-class rescaling preserves within-class ordering but can shift the inter-class argmax; temperature preserves argmax by construction.

YOLO: the same recipe degrades substantially. On $\mathrm{D}_{\mathrm{REAL}}$, temperature scaling *inflates* ECE from 0.035 to 0.642 ($18\times$). The fitted $\hat{T}=18.96$ adapts to d1_test's broad bimodal distribution; on d2_test, where retained pixels peak near confidence 1.0 and are $98.5\,\%$ correct, the temperature-softened confidences fall well below accuracy and open a large underconfidence gap. Platt and isotonic also degrade d2_test ECE (to 0.25) but less severely. The lesson: a scaler fit on one confidence distribution cannot be applied to a structurally different one without held-out real-domain validation. Reliability diagrams are in Appendix E.

4.5. Conformal prediction under residual uncertainty

Given that even isotonic BCPNN retains ECE 0.07 on $\mathrm{D}_{\mathrm{REAL}}$ and that the three post-hoc scalers tested above do not provide a reliable real-domain correction for YOLO under the current emission regime, we ask what distribution-free coverage guarantees remain available. We apply four conformal prediction methods to the BCPNN posterior with adaptive prediction sets (APS) [23]. Results are summarised in Table 3.

Split and weighted conformal prediction collapse on BCPNN. Of the $16{,}673$ d1_val APS scores, $87\,\%$ lie within 10^{-7} of the atom $s=1$ and $65\,\%$ are bit-exact. The finite-sample quantile $\hat{\tau}$ therefore lands inside the atom for every $\alpha \in [0.05, 0.20]$, every test point's cumulative vector reaches mass ≤ 1 at all five ranks, and the set is $C(x)=\{0,1,2,3,4\}$ uniformly: coverage 1, size $K=5$. Weighted conformal inherits the pathology because the permuted quantile lands in the same atom. *This is a score-distribution failure, not a domain-shift failure*, intrinsic to BCPNN's near-deterministic softmax winner-take-all [30] and independent of d1/d2.

ConfTS partially breaks the atom. Following Xi et al. [26], we fit $\hat{T}=12.01$ to minimise mean set size on calibration; the at-1 fraction drops from $83\,\%$ to $\sim 8\,\%$. At $\alpha=0.10$ ConfTS produces informative sets but coverage is asymmetric: undercovering on d1_test (0.862, ~ 11 standard errors below target at $n=8{,}058$) and overcovering on d2_test (0.930), mean set size ~ 4. ConfTS is therefore best read as an offline set-size-reduction heuristic, not the most reliable coverage-control method in this setting. At $\alpha=0.20$ coverage is 0.79 on both splits with mean size 2.5; at $\alpha=0.05$ the residual $8\,\%$ atom still captures the quantile and the set remains trivial.

ACI attains target empirical online coverage at every α tested. With $\gamma=0.01$ and warmup-gated α_t updates on the d2_test stream [25], observed coverage is $0.957/0.907/0.806$ at $\alpha=0.05/0.10/0.20$ (within 0.8 pp of target at every level), with mean set sizes $4.22/3.52/2.53$. ACI is the only method whose prediction sets remain informative at $\alpha=0.05$. Set-size distributions across all α levels and both test splits are in Appendix D.

Together, these correction experiments support the second main finding: *the right correction depends on both model architecture and score distribution.* Post-hoc calibration improves BCPNN but not YOLO; standard CP leaves BCPNN with trivial sets while ACI recovers target coverage across all α. No single UQ method should be assumed reliable without model-specific validation.

5. Discussion

Silent failure is a central operational concern. A conventional task-specific acceptance gate passes both models on D_{REAL} (BCPNN $58\,\%$ patch accuracy, YOLO $98.5\,\%$ pixel accuracy on the $1\,\%$ of pixels it labels), masking that the YOLO number is conditional on an abstention a field operator would not see on the overlay. EU AI Act Art. 13(3)(b)(ii) and Art. 15(1) [4, 5] are actionable only when calibration and emission coverage are disclosed alongside accuracy.

Post-hoc prescriptions are architecture-specific. Bayesian-Hebbian posteriors with WTA within hypercolumns [30] admit isotonic recalibration that transfers cleanly sim-to-real; abstention-biased detectors do not, as a scaler fit on D_{SIM}'s broad distribution cannot be applied unmodified to D_{REAL}'s peaked-at-1 distribution ($18\times$ ECE inflation in our data). This indicates that post-hoc calibration should be validated on held-out real-domain data before deployment, rather than assumed to transfer from simulation.

Conformal prediction as a coverage diagnostic with model-specific cost. Split and weighted CP collapse to trivial full-set predictors on near-deterministic Bayesian classifiers: the atom at $s = 1$ is a *score-distribution* failure intrinsic to any classifier placing $\sim 90\,\%$ of its posterior on one label. ConfTS is the most effective offline method among those tested at moderate α; ACI is the only method tracking observed shift online and remains informative at $\alpha = 0.05$. Aligning with NIST AI 800-3 [6] and NIST AI 800-2 Practice 3.1 [9], we recommend an offline ConfTS fit plus an online ACI update.

From uncertainty signals to deployment decisions. High BCPNN ECE means confidence is not a probability: recalibrate offline and gate field terrain decisions on ACI sets, escalating non-singletons. YOLO's low ECE with collapsed D_{REAL} emission is a reject-and-fallback trigger; for both models, go/no-go should depend on uncertainty evidence, not accuracy alone.

Limitations. The study uses one SAR arena (52 simulated, 72 real images); additional arenas, broader hidden-layer sweeps, and classifier baselines such as MC Dropout [31] and Deep Ensembles [32] remain future work. The image-clustered ‡ rows in Table 1 check within-image dependence.

6. Conclusion and Future Work

Calibration failures under sim-to-real shift are not captured by headline metrics: BCPNN degrades uniformly, while YOLO's apparent calibration improvement is abstention-masked. Correction is model-dependent: post-hoc calibration transfers for BCPNN but fails for YOLO; split/weighted CP collapse on BCPNN's APS atom, while ACI remains informative. Together this supports joint reporting of accuracy, calibration, emission coverage, and conformal set behaviour under EU AI Act Art. 13/15 and NIST AI 800-3. Future work extends weighted CP, $LaECE_0$, classifier baselines, and arenas [8].

Acknowledgments

Supported by EXTRA-BRAIN (Horizon Europe GA No. 101135809) and THEMIS 5.0 (Horizon Europe GA No. 101121042).

Declaration on Generative AI

The authors used Anthropic Claude for language editing, grammar, and spelling checks; they reviewed and edited all content and take full responsibility.

References

[1] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.

[2] E. Bonetto, C. Xu, A. Ahmad, GRADE: Generating realistic and dynamic environments for robotics research with isaac sim, 2023. arXiv:2303.04466.

[3] C. Cruz Ulloa, N. B. Ravichandran, D. Argüello Ron, A. Prieto, A. Lansner, P. Herman, J. del Cerro, Sim-to-real neural perception for terrain classification in search-and-rescue robotics with brain-inspired BCPNN, in: Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), 2025.

[4] European Parliament and Council, Regulation (EU) 2024/1689 of the European parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (AI Act), Official Journal of the European Union, OJ L 2024/1689, 2024.

[5] D. Schnurr, Effective implementation of requirements for high-risk AI systems under the AI act: Transparency and appropriate accuracy, CERRE Issue Paper, 2025.

[6] National Institute of Standards and Technology, Expanding the AI Evaluation Toolbox with Statistical Models, Technical Report NIST AI 800-3, NIST, 2026.

[7] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Technical Report NIST AI 100-1, NIST, 2023.

[8] A. Jacoff, E. Messina, H.-M. Huang, A. Virts, A. Downs, R. Norcross, R. Sheh, Standard Test Methods for Response Robots: ASTM International Standards Committee on Homeland Security Applications E54.09, Technical Report, National Institute of Standards and Technology (NIST), 2020. Orange-Zone arena specifications for urban search and rescue robotics.

[9] National Institute of Standards and Technology, Practices for Automated Benchmark Evaluations, Technical Report NIST AI 800-2 (Initial Public Draft), NIST, 2026. Practice 3.1: Conduct statistical analysis and uncertainty quantification.

[10] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 1321–1330. arXiv:1706.04599.

[11] F. Küppers, J. Kronenberger, A. Shantia, A. Haselhoff, Multivariate confidence calibration for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020. arXiv:2004.13546.

[12] F. Küppers, A. Haselhoff, J. Kronenberger, J. Schneider, Confidence calibration for object detection and segmentation, in: Deep Neural Networks and Data for Automated Driving, Springer, 2022. arXiv:2202.12785.

[13] S. Kuzucu, K. Oksuz, J. Sadeghi, P. K. Dokania, On calibration of object detectors: Pitfalls, evaluation and baselines, in: Proceedings of the European Conference on Computer Vision (ECCV), 2024. arXiv:2405.20459, oral presentation.

[14] T. Popordanoska, A. Tiulpin, M. B. Blaschko, Beyond classification: Definition and density-based estimation of calibration in object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024. arXiv:2312.06645.

[15] J. C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: A. J. Smola, P. Bartlett, B. Schölkopf, D. Schuurmans (Eds.), Advances in Large-Margin Classifiers, MIT Press, 1999, pp. 61–74.

[16] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 694–699.

[17] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[18] Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks, in: Advances in Neural

Information Processing Systems (NeurIPS), 2017. arXiv:1705.08500.
[19] Y. Geifman, R. El-Yaniv, SelectiveNet: A deep neural network with an integrated reject option, in: Proceedings of the 36th International Conference on Machine Learning (ICML), 2019, pp. 2151–2159. arXiv:1901.09192.
[20] V. Franc, D. Průša, V. Voracek, Optimal strategies for reject option classifiers, Journal of Machine Learning Research 24 (2023) 1–49.
[21] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, 2005.
[22] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. arXiv:2107.07511.
[23] Y. Romano, M. Sesia, E. J. Candès, Classification with valid and adaptive coverage, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.
[24] R. J. Tibshirani, R. F. Barber, E. J. Candès, A. Ramdas, Conformal prediction under covariate shift, in: Advances in Neural Information Processing Systems (NeurIPS), 2019. arXiv:1904.06019.
[25] I. Gibbs, E. J. Candès, Adaptive conformal inference under distribution shift, in: Advances in Neural Information Processing Systems (NeurIPS), 2021. arXiv:2106.00170.
[26] H. Xi, J. Huang, L. Yang, N. Guan, C. Sun, H. Dai, Does confidence calibration improve conformal prediction?, Transactions on Machine Learning Research (TMLR) (2024). arXiv:2402.04344.
[27] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, M. Pavone, Sample-efficient safety assurances using conformal prediction, The International Journal of Robotics Research (2024). arXiv:2109.14082.
[28] A. Lansner, Ö. Ekeberg, A one-layer feedback artificial neural network with a Bayesian learning rule, International Journal of Neural Systems 1 (1989) 77–87.
[29] N. B. Ravichandran, A. Lansner, P. Herman, Learning representations in bayesian confidence propagation neural networks, 2020. arXiv:2003.12415.
[30] N. B. Ravichandran, A. Lansner, P. Herman, Unsupervised representation learning with Hebbian synaptic and structural plasticity in brain-like feedforward neural networks, Neurocomputing 626 (2025) 129440.
[31] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2016.
[32] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[33] C. Cruz Ulloa, Synthetic RGB-D dataset for terrain classification in search and rescue robotics, https://doi.org/10.5281/zenodo.17233719, 2025. doi:10.5281/zenodo.17233719.
[34] K. Achig, YOLOv8m-seg trained model for SAR terrain segmentation, https://doi.org/10.5281/zenodo.15674867, 2025. doi:10.5281/zenodo.15674867.

A. Robustness check: unfiltered BCPNN calibration

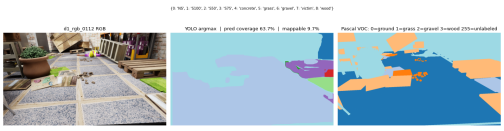
The main-body BCPNN calibration is reported on the class-0-filtered pool because the published `testlbls.bin` contains zero class-0 samples, so class-0 emissions from the model are unscorable against the reference labelling. For transparency we also report the unfiltered numbers that include $\sim 33\,\%$ class-0 "no polygon coverage" patches (Table 4). The unfiltered comparison loses the sim-to-real significance signal (the class-0 noise swamps the signal), but the magnitude of in-domain miscalibration remains consistent.

B. YOLO prediction-mask overlays

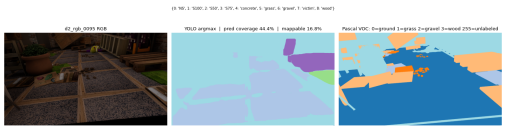
Figure 3 shows per-image prediction coverage for representative dense examples in each domain. Aggregate coverage is $6.0\,\%$ on $\mathrm{D}_{\mathrm{SIM}}$ and $1.0\,\%$ on $\mathrm{D}_{\mathrm{REAL}}$: any pixel-level calibration metric reported without this coverage figure is silent on $\geq 94\,\%$ of the input surface.

Table 4
BCPNN calibration, unfiltered (class-0 patches retained). Compare to the filtered numbers in Table 1. The Δ row is a robustness check only: the unfiltered Δ-ECE CI spans zero because class-0 noise dominates.

Split	n	Accuracy	ECE	Brier
d1_test	11,440	0.427 [0.418, 0.436]	0.563 [0.554, 0.572]	1.131 [1.112, 1.149]
d2_test	15,840	0.419 [0.411, 0.427]	0.567 [0.560, 0.575]	1.142 [1.127, 1.158]
Δ	—	-0.008	$+0.004$ [-0.008, $+0.017$]	$+0.012$



(a) D_{SIM} example (63.7 % pred coverage).



(b) D_{REAL} example (44.1 % pred coverage).

Figure 3: YOLOv8m-seg prediction-mask overlays: dense examples per domain.

C. YOLOv8m-seg class distribution under shift

Table 5 reports the per-split distribution of YOLO's argmax output across the native 9-class label set. The `S100` SAR-specific class dominates on both domains; the two name-matchable terrain classes *grass* and *wood* together capture 12 % and 41 % of emitted pixels on D_{SIM} and D_{REAL} respectively. The `no_pred` row makes the abstention regime of Section 4.2 explicit: 94 % of D_{SIM} pixels and 99 % of D_{REAL} pixels receive no YOLO label above threshold.

Table 5
YOLOv8m-seg argmax distribution, fraction of total pixels per split. "no_pred" is the fraction below the detector's confidence threshold (0.001).

Class	d1_test	d1_val	d2_test	d2_val
S100	0.0522	0.0532	0.0052	0.0039
grass	0.00156	0.00064	0.00055	0.00062
wood	0.00564	0.00228	0.00313	0.00567
gravel	0.00008	0	0	0.00000
concrete	0.00002	0.00002	0	0
other	0.00069	0.00069	0.00012	0.00035
no_pred	0.9398	0.9433	0.9911	0.9893

D. Conformal set-size distributions across α and domains

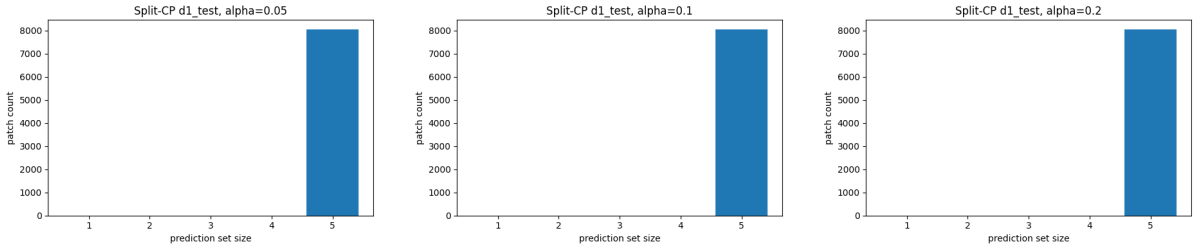
Conformal set-size distributions across all $\alpha \in \{0.05, 0.10, 0.20\}$ and both test splits are shown in Fig. 4.

E. Post-hoc reliability diagrams (D_{REAL})

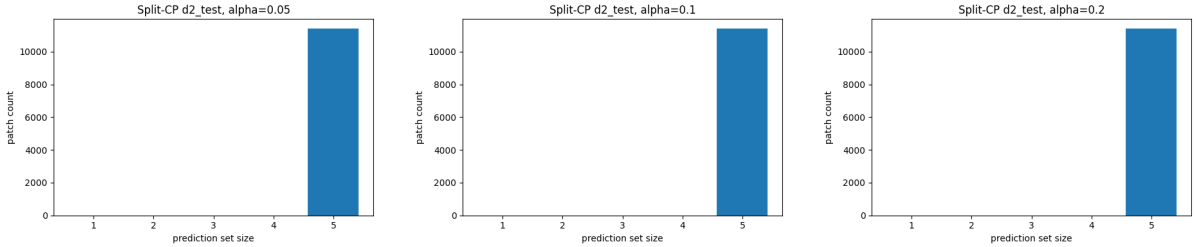
Figure 5 contrasts the post-hoc calibration outcomes on D_{REAL} for both models.

F. Per-class BCPNN reliability

Figure 6 shows BCPNN one-vs-rest reliability per class. The *ground* class bears most of the sim-to-real degradation; *wood* is the highest-absolute miscalibrated class in-domain but barely shifts across domains.

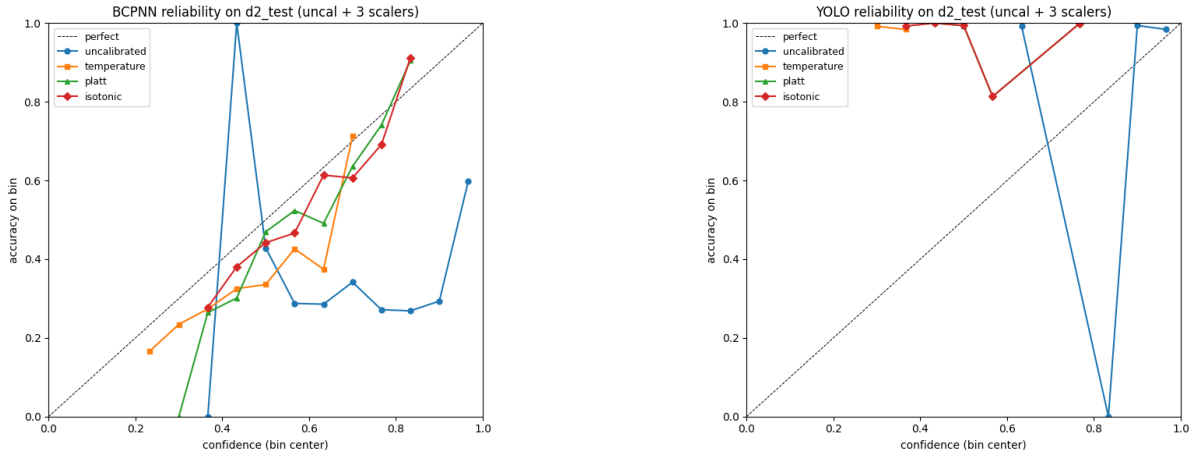


(a) d1_test, $\alpha = 0.05$. (b) d1_test, $\alpha = 0.10$. (c) d1_test, $\alpha = 0.20$.



(d) d2_test, $\alpha = 0.05$. (e) d2_test, $\alpha = 0.10$. (f) d2_test, $\alpha = 0.20$.

Figure 4: Conformal set-size distributions across all $\alpha \in \{0.05, 0.10, 0.20\}$ and both test splits. Split and weighted conformal (not shown separately: all mass at size 5) collapse to the trivial full-set predictor at every α due to the APS atom at $s = 1$.



(a) BCPNN recovery: isotonic collapses the reliability curve onto the diagonal.

(b) YOLO failure: temperature drags the curve sharply off the diagonal ($18\times$ ECE inflation).

Figure 5: Post-hoc calibration outcomes on D_{REAL}. BCPNN admits sim-to-real recalibration; YOLO's detection-regime shift is not amenable to the same treatment.

G. Reproducibility: float-precision fix in APS scoring

During review of the main conformal-prediction implementation we identified and fixed a single float-precision artefact in the APS cumulative-sum implementation. On float32 posteriors re-accumulated in float64, the per-row cumulative sum can exceed 1 by approximately 10^{-7}; under a strict $\mathrm{cum} > \tau$ comparison at $\tau = 1$, this caused occasional exclusion of the final-ranked class from test sets even though the calibration quantile had landed exactly at the atom. A clamp $\mathrm{cum} \leftarrow \min(\mathrm{cum}, 1)$ restores the expected plateau-inclusion behaviour; a regression test now guards the invariant. The fix changed the reported split / weighted CP numbers by 2–9 percentage points in coverage before collapsing to the qualitative trivial-set result of Table 3. Independently, an off-by-one in the ACI α_t update during the warmup window (the update fired on the pre-warmup error $\mathrm{err}_t = 0$ and drifted α towards its clip

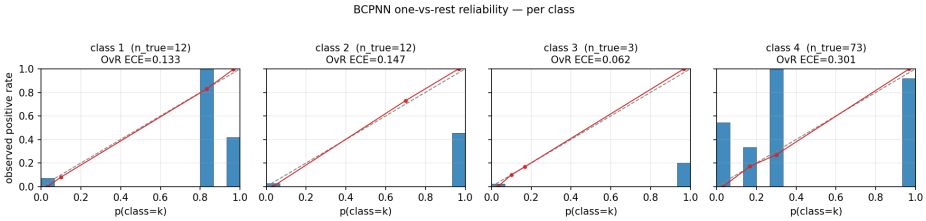


Figure 6: BCPNN per-class one-vs-rest reliability on D_{SIM} and D_{REAL}. *ground* degrades most sharply ($0.253 \rightarrow 0.324$, $+7.1$ pp) and carries the largest absolute miscalibration on D_{REAL}; *wood* has the highest SIM-ECE (0.295) but barely shifts (-1.8 pp); *grass* and *gravel* both remain below 0.15 on either domain.

ceiling) was corrected by gating the update on the post-warmup branch; a unit test asserts $\alpha_t = \alpha_0$ throughout warmup.

H. Hyperparameters

BCPNN: $H_{\text{in}} \times M_{\text{in}} = 256 \times 20$; $H_{\text{hid}} \times M_{\text{hid}} = 100 \times 100$; $H_{\text{out}} \times M_{\text{out}} = 1 \times 5$; $n_{\text{actHi}} = 25$; $n_{\text{silHi}} = 231$; activation function SOFTMAX; learning rule BCPNN; $n_{\text{enpat}} = 70$; $n_{\text{loop}} = 3$; noise amplitude $n_{\text{ampl}} = 0.01$; 64×64 patches.

YOLOv8m-seg: checkpoint [34]; image size 640; inference confidence threshold 0.001; default `retina_masks=True`.

Post-hoc calibration: temperature solved by L-BFGS on calibration NLL; Platt logistic regression with `L-BFGS`; isotonic regression with `scikit-learn` default parameters and marginal-fallback for zero-positive classes; all fit on d1_val, evaluated on d1_test and d2_test.

Conformal: $\alpha \in \{0.05, 0.10, 0.20\}$; APS score convention $C(x) = \{y : s(x, y) \leq \tau\}$; ConfTS temperature found by minimising mean set size over $\alpha = 0.10$ on d1_val; ACI $\gamma = 0.01$, warmup $1{,}000$ on d2_test stream.

I. Parameter sensitivity

We report sensitivity analyses for the three fixed hyperparameters most likely to influence the headline conclusions: the ECE bin count, the weighted-CP weight-clip bounds, and the ACI step size γ. All numbers are recomputed with the same loaders and estimators used for the main tables.

ECE bin count. Table 6 recomputes BCPNN ECE on both test splits for equal-width bin counts $B \in \{5, 10, 15, 20, 25, 30\}$. Point ECE on each split and the sim-to-real ΔECE are stable to within 10^{-4} across the whole range: the reported $\Delta\text{ECE} = +0.018$ is not an artefact of the $B = 15$ choice.

Table 6
BCPNN ECE and sim-to-real ΔECE as a function of the number of equal-width ECE bins. Point estimates on the filtered d1_test ($n = 8{,}058$) and d2_test ($n = 11{,}413$) pools; compare the $B = 15$ row to Table 1.

Bins B	ECE (d1_test)	ECE (d2_test)	ΔECE (d2−d1)
5	0.3850	0.4032	+0.0182
10	0.3850	0.4033	+0.0183
15	0.3850	0.4033	+0.0183
20	0.3850	0.4033	+0.0183
25	0.3850	0.4033	+0.0183
30	0.3850	0.4033	+0.0183

Weighted-CP clip bounds. The density-ratio weights are clipped to $[0.01, 100]$, but at the fitted setting this clip is *non-binding*: the empirical weights span $[0.037, 0.403]$ (median 0.066), strictly inside the clip, so no weight is actually clipped. Moreover (Section 4.5), $\sim 87\,\%$ of the 16,673 d1_val APS calibration scores lie within 10^{-7} of the atom $s = 1$, leaving only $\sim 12\,\%$ of calibration mass below the atom. Even under an adversarial reweighting at the extreme in-range ratio ($0.403/0.037 \approx 11$), the weighted fraction of mass below the atom is at most ≈ 0.61, which is below the $1 - \alpha$ target for $\alpha \leq 0.10$; the weighted $(1 - \alpha)$-quantile therefore remains pinned at $s = 1$ and split/weighted CP collapse to the full set (Table 3) for any clip in, e.g., $\{[0.001, 1000], [0.01, 100], [0.05, 20]\}$. The collapse is a property of the APS atom, not of the clip choice.

ACI step size. Table 7 reports ACI empirical online coverage and mean set size for $\gamma \in \{0.005, 0.01\}$ at each α (the full rolling-coverage streams are the source of Table 3). Coverage moves by at most 0.7 percentage points between the two step sizes and stays within ~ 1.5 pp of the target $1 - \alpha$ at every level, so the reported $\gamma = 0.01$ operating point is not finely tuned.

Table 7
ACI empirical online coverage / mean set size on the d2_test stream ($n_{\text{post-warmup}} = 10{,}413$; warmup $1{,}000$) for two step sizes γ. Target coverage is $1 - \alpha$.

α	$\gamma = 0.005$ Cov / Size	$\gamma = 0.01$ Cov / Size	Target $1 - \alpha$
0.05	0.965 / 4.38	0.957 / 4.22	0.95
0.10	0.914 / 3.65	0.907 / 3.52	0.90
0.20	0.812 / 2.62	0.806 / 2.53	0.80

J. Model commensurability

The unified assessment protocol is applied to two operationally different predictors, which we deliberately do not place on a common numeric scale. BCPNN emits one class-probability vector per retained 64×64 terrain patch and is scored as a forced choice over five classes. YOLOv8m-seg emits per-pixel class scores and abstains wherever the predicted mask falls below its confidence threshold, so its accuracy and ECE are conditional-on-emission (Section 3.2), and "coverage" carries two distinct meanings for it: *emission coverage* (fraction of pixels labelled) and *conformal coverage* (fraction of examples whose label lies in the prediction set). Because the two models differ in prediction unit (patch vs pixel), scoring rule (forced choice vs threshold-plus-abstention), and even in what a low ECE *means*, we draw no head-to-head ranking between them. The single unified claim is methodological: applying one disclosure protocol—report calibration *and* prediction/emission coverage alongside accuracy—surfaces a distinct, architecture-specific trust failure in each model (patch-level overconfidence for BCPNN; abstention-masked ECE for YOLO). The two cases are complementary instances of the same disclosure gap, not competing measurements of one quantity.