CERTOD: CLASS-CONDITIONAL PAIR CERTIFICATES FOR MATCHED OBJECT DETECTIONS IN AUTONOMOUS DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detection combines a discrete class and a continuous box. Certifying both under class-specific error budgets requires a joint guarantee conditional on the true class, which is unknown at inference, so selecting a geometric margin by the detector's possibly incorrect top-1 label does not establish that guarantee. CertOD instead emits per-class pair certificates: one calibrated box for each surviving class hypothesis. Under within-class exchangeability it provides finite-sample coverage of the event that the true class is retained and its associated box contains the object, at each class's prescribed level for matched detections of certifiable classes. Against a matched single-box control, pairs reduce retained true-class pedestrian area by 6.8% (median) and 22.2% (p90) on BDD100K at a 0.47% exclusion rate; CODA2022 reproduces the tail benefit (6.6%–35.6% across four common classes). The gain requires a consumer that retains class–box associations: the smallest axis-aligned envelope of all pairs is exactly the conservative single box. Supremum scoring lowers the finite-quantile sample threshold roughly fourfold against coordinate Bonferroni without shrinking pedestrian boxes, and unified joint scoring is larger in median area wherever the class set is held fixed. We also report three boundaries: 33.3% of BDD100K pedestrians match no proposal, zero-shot transfer of BDD thresholds to CODA2022 misses all eight tested score–class targets, and the repair weakens as the detector sharpens—on the misclassified slices its joint coverage falls from 67.9% to 12.7% across four backbones, because a better classifier yields smaller class sets.

1 INTRODUCTION

A detection combines a discrete class and a continuous box. A useful certificate must cover their conjunction: the true class must remain a candidate, and the box associated with that class must contain the object. Containing a pedestrian's geometry while excluding `person` from the candidate set does not satisfy this target. Driving further motivates class-specific error tolerances, since the consequences of a miss differ across road users. Conformal prediction supplies finite-sample coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2021); the question here is how to apply it to this joint, class-conditional target when the class itself is uncertain.

The uncertainty creates a calibration–inference mismatch. To control error for objects whose true class is c, calibration must establish a bound conditional on that true class. At inference, however, the detector provides a predicted label and class scores. Selecting the box margin by its argmax can apply another class's calibrated margin to a misclassified object (Figure 1(a)). Calibration conditional on the predicted class also changes the event being controlled: a bound for objects predicted as c does not imply a bound for objects that are actually c. This distinction matters even when aggregate coverage is close to target. On BDD100K, the frozen detector misclassifies 1.27% of matched pedestrians. That rate motivates examining margin selection; it does not itself establish failure of the joint certificate, because a candidate set can still contain the true class after a top-1 mistake.

A conformal class set offers a conservative repair. Test every candidate against its own true-class threshold, then enlarge one box by the coordinatewise largest margin any retained candidate demands. When the true class is retained that box includes its true-class-calibrated box, so class exclusion and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

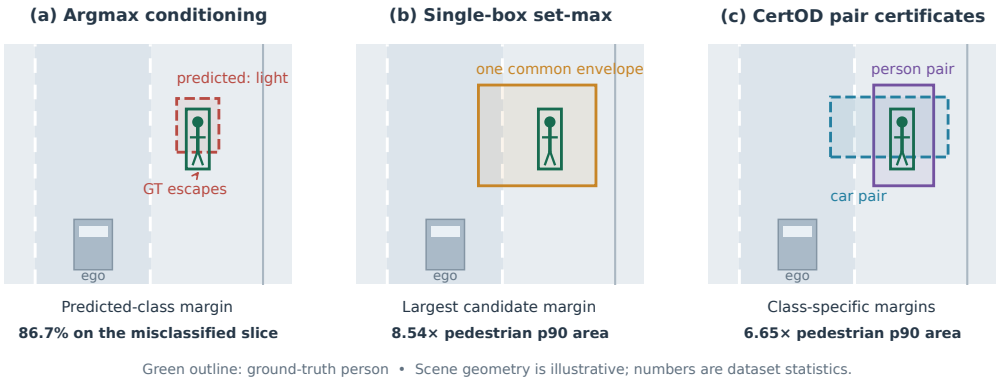


Figure 1: Three certification outputs (schematic, not detector output). **(a)** Argmax conditioning applies the predicted class's margin; aggregate coverage can conceal low coverage on a misclassified slice. **(b)** A single set-max box applies the largest candidate margin. **(c)** Pair certificates retain a box for each class hypothesis and reduce retained true-pair pedestrian p90 area by 22.2% for a class-aware consumer. Their smallest common axis-aligned envelope equals (b) (Appendix B.5).

containment can be bounded together. The cost is in the output: one rectangle must reach the most outward edge any candidate demands, even for a consumer interested in a single class. The matched single-box control reaches $\times 8.53$ p90 retained true-pair pedestrian area on BDD100K (Figure 1(b)).

CertOD retains the association that the single-box output discards. It emits $\{(c, \hat{B}_c(x)) : c \in S^+(x)\}$, one box calibrated for each surviving class hypothesis (Figure 1(c)). All pairs share the detector's proposal, but each carries its own class-specific margin. No test-time true label is required: every candidate is evaluated against fixed thresholds, and the consumer receives all retained pairs. Splitting the error budget between class exclusion and box noncontainment gives the joint guarantee for the true-class pair, for certifiable classes of matched detections under within-class exchangeability.

This representation benefits a consumer that retains class-specific hypotheses. A consumer that instead requires one axis-aligned envelope around all hypotheses obtains exactly the conservative single box; CertOD does not reduce that envelope. The evaluation therefore compares the two outputs using identical quantiles, class sets, budgets, and scores, so that any difference in per-class certified area can be attributed to representation. It also audits the limits that this change cannot address, including missed proposals and distribution shift.

The paper makes three contributions:

- **Pair output and its boundary.** On BDD100K, pairs reduce retained true-pair pedestrian area by 22.2% at p90 and 6.8% at the median. The common axis-aligned envelope is exactly the single-box control, so the gain requires retaining class–box associations.
- **Design choices and guarantee boundaries.** The study separates finite-sample feasibility from geometric tightness. Supremum scoring lowers the finite-quantile sample threshold fourfold and makes motorcycle certifiable on all splits; it does not shrink pedestrian boxes where both methods are feasible. Unified scoring, misclassified slices, object sizes and detector backbones establish further empirical limits, and 33.3% of pedestrians match no proposal at all.
- **Replication and transfer.** CODA2022 reproduces the conditional p90 reduction (6.6%–35.6%), while frozen BDD thresholds miss all eight transfer targets. Pairs also beat the single envelope of a same-detector RAPS–Bonferroni baseline (Angelopoulos et al., 2021; Ries et al., 2026) (Appendix D.6).

2 Related Work

Conformal prediction and conditional coverage. Standard constructions cover classification (Sadinle et al., 2019; Angelopoulos et al., 2021; Romano et al., 2020), quantile regression (Romano

et al., 2019; Sesia & Candès, 2020), and monotone risk control (Bates et al., 2021; Angelopoulos et al., 2025; 2024); see Fontana et al. (2023) and Zhou et al. (2025) for surveys. Our class candidate set is the thresholded true-class score of Sadinle et al. (2019), calibrated per class rather than pooled. Group-conditional methods partition calibration data by class (Ding et al., 2023), target a tractable family of conditioning events (Gibbs et al., 2023), or localize quantiles around the test point (Guan, 2023; Fuchs et al., 2026). CertOD uses true-class Mondrian calibration (Vovk et al., 2003), whose finite-quantile feasibility depends on class sample size. At inference the true class is unknown, so selecting a margin by the predicted class changes the conditioning event (Section 3.2).

Exchangeability and shift. Weighted conformal prediction addresses covariate shift by reweighting calibration data (Tibshirani et al., 2019); other approaches provide doubly robust corrections (Yang et al., 2024) or bound coverage gaps under departures from exchangeability (Barber et al., 2023). Our construction applies none of these corrections and does not guarantee validity under arbitrary shift. We audit frozen thresholds across datasets (Section 4.2) and scene attributes (Appendix A.9).

Detection uncertainty. MC-dropout, ensembles, loss attenuation, and evidential approaches estimate uncertainty (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Barbarit–Gaboriau et al., 2026) but do not supply the finite-sample joint guarantee studied here. Conformal detection includes sequential box risk control (Andéol et al., 2025) and coordinate-wise split CP scaled by detector variance (Ries et al., 2026); neighbouring work addresses instance segmentation and post-hoc instance uncertainty (Lu et al., 2026; Zhang et al., 2026).

Detector interfaces. CertOD requires a full class-score vector per proposal, which query-based detectors supply: DETR (Carion et al., 2020), Deformable DETR (Zhu et al., 2021), ViT-based detection (Fang et al., 2021), and later refinements (Zhang et al., 2023; Zhao et al., 2024). An output retaining one label per box is insufficient. Appendix A.5 compares four backbones.

Confidence calibration. Post-hoc calibration of classifiers (Guo et al., 2017) and detectors (Lee et al., 2026) controls confidence, not the true-class joint event of label inclusion and box containment, and therefore complements the certificate; the cross-backbone audit (Appendix A.5) keeps the two apart.

3 The CertOD Framework

CertOD wraps a frozen detector by constructing a conformal class set, retrieving a true-class-calibrated geometric margin for each candidate, and emitting one enlarged box per candidate as a class–box pair. We first state the target and conditioning problem, then define the pair certificate, its geometric score, and the inference procedure.

3.1 Problem Formulation

An input image I contains ground-truth entities $\mathcal{G} = \{(B^*_m, c^*_m)\}_{m=1}^{M}$, where $B^*_m = [x^*_{1,m}, y^*_{1,m}, x^*_{2,m}, y^*_{2,m}] \in \mathbb{R}^4$ is a box and $c^*_m \in \mathcal{K} = \{1, \dots, K\}$ a category. A frozen detector such as DETR (Carion et al., 2020) returns proposals $\mathcal{D}_{\text{det}} = \{(\hat{B}_j, p_j)\}_{j=1}^{J}$, each carrying a box $\hat{B}_j$ and a class distribution $p_j \in \Delta^K$. Class-agnostic greedy matching at IoU ≥ 0.5 pairs proposals with ground truth and yields matched rows $Z_i = (X_i, Y_i)$, with input $X_i = (\hat{B}_i, p_i)$ and label $Y_i = (B^*_i, c^*_i)$. The detector remains frozen; CertOD wraps its matched outputs.

The target. A certificate turns one matched row into two objects: a set of candidate categories $S(x) \subseteq \mathcal{K}$, and for each candidate c an enlarged box $\hat{B}_c(x) \supseteq \hat{B}(x)$. It is correct on a row when the true category is among the candidates *and* the box attached to that category contains the true box. We require that to happen with a prescribed probability for the category an object *is*, not the one it was predicted to be. Write $\mathcal{K}_{\text{cert}}$ for the classes with finite class and box quantiles (the mask in Section 3.3); the guarantee is stated only for those classes:

$$P\Big(c^* \in S(x) \ \wedge \ B^* \subseteq \hat{B}_{c^*}(x) \ \Big| \ c^* = c\Big) \geq 1 - \alpha_c, \qquad \forall c \in \mathcal{K}_{\text{cert}}. \tag{1}$$

CertOD instantiates this target with the candidate set $S^+(x)$ and class-specific boxes (Section 3.3).

Heterogeneous safety budgets. The assignment $\alpha : \mathcal{K} \to (0, 1)$ is fixed before calibration and can reflect class-dependent consequences of a miss. The construction accepts any such assignment for classes with finite quantiles. Different budgets make the conditioning choice consequential: a guarantee under the predicted class does not establish the required guarantee under the true class. Appendix A.1 gives the experimental assignment.

What the guarantee is over. The finite-sample statements require within-class exchangeability of matched calibration rows and a fresh matched test row. Image-disjointness alone does not establish this assumption, so the primary pooled-row results are empirical. Under exchangeable images, a separate one-row-per-image-class audit implements a sufficient design (Appendix B.1). The guarantee is marginal over matched detections in $\mathcal{K}_{\text{cert}}$, not over images or all ground-truth objects.

3.2 Why Conditioning on the Argmax Fails

Split conformal prediction calibrates score thresholds on matched rows and transfers their coverage to a fresh exchangeable row, without retraining.

One target, two thresholds. Equation (1) is a conjunction, and its two conjuncts are different kinds of object. Membership $c^* \in S(x)$ is a discrete event decided by the classifier scores; containment $B^* \subseteq \hat{B}_{c^*}(x)$ is a geometric event decided by four box coordinates. No single score orders both without choosing a relative scale, so the default construction certifies them separately and pays for both out of the one budget α_c. For a row whose true class is c, write the two failure events

$$\mathcal{E}_c^{\text{cls}} = \big\{c \notin S(x)\big\}, \qquad \mathcal{E}_c^{\text{box}} = \big\{B^* \not\subseteq \hat{B}_c(x)\big\}, \tag{2}$$

so that the target fails exactly on $\mathcal{E}_c^{\text{cls}} \cup \mathcal{E}_c^{\text{box}}$. Bounding that union by the sum of its parts gives the decomposition every construction below is built on.

Lemma 1 (Budget decomposition)**.** *Fix a class c and a split $\alpha_c = \alpha_c^{\text{cls}} + \alpha_c^{\text{box}}$ into non-negative parts. If a certificate satisfies the two conditions*

$$P\big(c \in S(x) \,\big|\, c^* = c\big) \geq 1 - \alpha_c^{\text{cls}} \qquad \textit{and} \qquad P\big(B^* \subseteq \hat{B}_c(x) \,\big|\, c^* = c\big) \geq 1 - \alpha_c^{\text{box}}, \tag{3}$$

then it satisfies Equation (1) *at level α_c.*

The result follows from the conditional union bound; Appendix B.4 gives the proof.

The decomposition applies to both single-box and pair outputs. Unless stated otherwise, $\alpha_c^{\text{cls}} = \alpha_c^{\text{box}} = \alpha_c/2$; Section 3.5 considers a unified score. Both calibrated bounds must condition on the true class. Inference must apply these bounds without knowing the true class.

The geometric certificate enlarges the detector box outward to contain the true box.

Definition 1 (Outward inflation)**.** For a box $\hat{B} = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2]$ and per-edge margins $\mathbf{q} = (q_0, q_1, q_2, q_3) \in \mathbb{R}^4_{\geq 0}$, write

$$\hat{B} \oplus \mathbf{q} = [\hat{x}_1 - q_0,\ \hat{y}_1 - q_1,\ \hat{x}_2 + q_2,\ \hat{y}_2 + q_3].$$

Section 3.4 maps scalar quantiles to these per-edge margins.

The conditioning problem. Pooled calibration controls an average error rate, whereas Equation (3) requires true-class bounds. Mondrian calibration (Vovk et al., 2003) therefore partitions rows by class and uses a separate quantile for each bucket.

Calibrating the box margin on the true-class bucket $\{i : c_i^* = c\}$ yields the second condition verbatim,

$$P\big(B^* \subseteq \hat{B}(x) \oplus \hat{\mathbf{q}}_c^{\text{box}} \,\big|\, c^* = c\big) \ \geq\ 1 - \alpha_c^{\text{box}}, \tag{4}$$

but applying it at test time means selecting the margin of a class nobody has observed. The deployable construction buckets by the prediction $\hat{c} = \arg\max_k p_k(x)$ and applies $\hat{\mathbf{q}}_{\hat{c}}^{\text{box}}$ instead, which certifies

$$P\big(B^* \subseteq \hat{B}(x) \oplus \hat{\mathbf{q}}_{\hat{c}}^{\text{box}} \,\big|\, \hat{c} = c\big) \ \geq\ 1 - \alpha_c^{\text{box}}. \tag{5}$$

Equations (4) and (5) condition on different events, and neither implies the other. They coincide on the rows the detector classifies correctly and diverge on the rest, where the margin applied is the one calibrated for another class.

The semantic half avoids selecting a class in advance by testing every candidate against its own true-class threshold. Geometry can likewise retain every candidate's margin, either in one conservative box or as separate class–box pairs.

On the diagnostic slice $\mathcal{E}_{\text{leak}} = \{i : c_i^* = \text{person},\ \hat{c}_i \neq \text{person}\}$, argmax selection applies another class's margin. The two budgets involved can differ by a factor of twenty under our assignment ($\alpha_{\text{traffic light}}/\alpha_{\text{person}} = 0.20/0.01$), but that ratio does not quantify joint-certificate error; Appendix A.4 reports containment and joint coverage separately.

The Set-Max repair, and its price. The leak is caused by conditioning the margin on a label that can be wrong, so the repair is to stop conditioning on it. Replace $\hat{c}$ by a conformal class set $S(x)$ that contains c^* with high probability, and inflate one box by the coordinatewise largest margin any candidate in that set would demand:

$$\hat{B}_{\text{Set-Max}}(x) = \hat{B}(x) \oplus \bigvee_{c \in S(x)} \mathbf{q}_c^{\text{box}}(x), \tag{6}$$

with $\oplus$ as in Definition 1 and $\vee$ the coordinatewise maximum. We call this construction **Set-Max** and carry it through the paper as the single-box comparator, calibrated on the *same* true-class buckets, the same class sets and the same budget split as our own output, so that when the two are compared in Section 4.1 the only difference left is the representation. Concurrent work (Ries et al., 2026) reaches a similar safe-box step from a different starting point; we implement their two-step method on the same detections in Appendix D.6 rather than argue from the difference.

Set-Max avoids argmax margin selection, but every candidate shares the largest margin in its set. The pair construction retains the class attached to each margin, allowing a class-aware consumer to use the corresponding box.

3.3 The Pair Certificate

CertOD emits, instead of one box, one box *per surviving hypothesis*. The certificate has two halves, calibrated separately on true-class subsets and combined by a union bound: a class candidate set, and one geometric margin per candidate.

The semantic half: a class candidate set. Score a category by its classifier shortfall, $s^{\text{cls}}(x, c) = 1 - p_c(x)$ (Sadinle et al., 2019), and let $\hat{q}_c^{\text{cls}}$ be the $(1 - \alpha_c^{\text{cls}})$ conformal quantile of that score over the calibration instances whose *true* class is c. Then

$$\begin{aligned} S(x) &= \big\{k \in \mathcal{K} : p_k(x) \geq 1 - \hat{q}_k^{\text{cls}} \ \wedge\ M_k = 1\big\}, \\ S^+(x) &= S(x) \cup \big\{k^*(x) : k^*(x) = \arg\max_{k \in \mathcal{K}} p_k(x),\ M_{k^*(x)} = 1\big\}. \end{aligned} \tag{7}$$

The set CertOD emits is $S^+(x)$. The *certifiability mask* $M_k = \mathbb{I}\big(n_k \geq \lceil 1/\alpha_k^{\text{cls}} \rceil - 1 \ \wedge\ n_k \geq \lceil 1/\alpha_k^{\text{box}} \rceil - 1\big)$ records whether class k has enough calibration instances to support its own budget at all; a class failing it has an infinite quantile and is reported as uncertifiable rather than silently padded.

The geometric half uses a margin satisfying the true-class containment bound in Equation (4); Section 3.4 constructs it.

The certificate. Putting the halves together,

$$\mathcal{C}_{\text{pair}}(x) = \Big\{\big(c,\ \hat{B}_c(x)\big) : c \in S^+(x)\Big\}, \qquad \hat{B}_c(x) = \hat{B}(x) \oplus \mathbf{q}_c^{\text{box}}(x), \tag{8}$$

where the scalar $\hat{q}_c^{\text{box}}$ inducing that margin vector is calibrated on instances whose *true* class is c. All pairs share one detector proposal and differ only in their calibrated margins; the representation requires neither extra detector passes nor additional proposals.

Corollary 2 (Validity of the pair certificate)**.** *For every class c with $M_c = 1$, the pair certificate of Equation* (8) *discharges both conditions of Lemma 1 and therefore satisfies Equation* (1) *at level α_c.*

Proof. True-class split conformal calibration bounds class exclusion and box noncontainment by α_c^{cls} and α_c^{box}, respectively. Augmentation from S to S^+ preserves membership, and Lemma 1 combines the bounds. Appendix B.4 gives the full proof. □

The margin is indexed by candidate c, so Equation (4) is the relevant calibrated bound. True labels are used only offline; inference evaluates fixed thresholds and margins for every candidate in $S^+(x)$ without knowing which one is true.

The mask keeps scarcity local: a class without finite quantiles is reported as uncertifiable and excluded, leaving other classes usable. Admitting its infinite margin into Set-Max would instead make that set's box vacuous. Adding a certifiable argmax label preserves the membership bound and ensures semantic coverage no worse than top-1 prediction within the certified label space; it does not guarantee containment (Appendix B.4).

3.4 Constructing the Geometric Margin

It remains to produce $\hat{q}_c^{\mathrm{box}}$, the scalar that Equation (8) turns into an envelope. Three choices separate it from a textbook regression interval: the residuals are one-sided, the four coordinates are collapsed by a supremum rather than a union bound, and the scale is normalised.

3.4.1 One-Sided Outward Containment Residuals

Define the four one-sided outward residuals

$$\mathbf{s}(x, B^*) = (\hat{x}_1 - x_1^*,\ \hat{y}_1 - y_1^*,\ x_2^* - \hat{x}_2,\ y_2^* - \hat{y}_2), \tag{9}$$

so that $B^* \subseteq \hat{B} \oplus \mathbf{q}$ holds exactly when $s_j \le q_j$ on all four edges. A two-sided score additionally penalizes inward errors and cannot yield a smaller margin at the same level (Appendix B.2).

3.4.2 Supremum Scalarization in Place of Bonferroni Splitting

Coordinate Bonferroni assigns $\alpha_{\mathrm{box}}/4$ to each edge; supremum scoring calibrates their maximum at α_{box}. A finite quantile needs at least $\lceil 1/\alpha_{\mathrm{box}} \rceil - 1$ rows for the maximum, versus $\lceil 4/\alpha_{\mathrm{box}} \rceil - 1$ for Bonferroni. Under the equal split, the pedestrian thresholds are 199 and 799 rows. Below its threshold the quantile is infinite (Theorem 5). This is a feasibility advantage; area comparisons remain empirical (Appendix A.6).

At the absolute pixel scale the supremum nonconformity score is

$$S_{\max}(x, B^*) = \max_{j \in \{0,1,2,3\}} s_j(x, B^*), \tag{10}$$

calibrated by one scalar quantile $\hat{q}_{1-\alpha_{\mathrm{box}}}$ over $\{S_{\max}(x_i, B_i^*)\}_{i=1}^n$. Section 3.4.3 applies the same maximum after scale normalization to obtain CertOD's default score. The residual equivalence calibrates joint containment directly, without a coordinatewise union bound (Appendix B.2).

3.4.3 Affine Scale Normalization and Size-Stratified Validity

Fixed pixel margins disproportionately inflate small objects. We therefore normalize residuals by predicted dimensions with a class-specific offset:

$$S_{\mathrm{affine}}^{(c)}(x, B^*) = \max \left\{ \frac{\hat{x}_1 - x_1^*}{\hat{w} + \gamma_c},\ \frac{\hat{y}_1 - y_1^*}{\hat{h} + \gamma_c},\ \frac{x_2^* - \hat{x}_2}{\hat{w} + \gamma_c},\ \frac{y_2^* - \hat{y}_2}{\hat{h} + \gamma_c} \right\}, \tag{11}$$

where $\hat{w} = \hat{x}_2 - \hat{x}_1$, $\hat{h} = \hat{y}_2 - \hat{y}_1$, and one nonnegative offset γ_c is selected for each class on a disjoint tuning split $\mathcal{D}_{\mathrm{tune}}$. The same offset is added to width and height. It is a class-specific hyperparameter fixed before calibration, rather than a uniformly beneficial regularizer. Its tuning results are in Appendix A.6. Calibrating Equation (11) yields the scalar margin in Equation (8).

Size-stratified audit. Class-conditional validity does not imply validity within size strata. The offset minimizes tuning-set median area, not a stratum constraint. Coverage on post-hoc, within-true-class ground-truth-area quintiles—fixed on the full matched audit population and unused by calibration—is therefore an audit, reported in Appendix A.7.

3.5 Whether the Budget Must Be Split

The union bound is valid without independence but can be conservative when class and box failures overlap. An alternative takes the maximum of their scores on a common scale and calibrates once at α_c, directly certifying the conjunction. Appendix B.6 defines this unified score and proves its validity.

Coverage does not imply smaller boxes: weakening the class term can shrink boxes by admitting more categories. We therefore tune its relative scale on a disjoint split and compare area alongside class-set size. We retain the split-budget default and report the comparison in Section 4.3 and Appendix A.6.

3.6 The Procedure

Algorithm 1 assembles the pieces. Offline calibration, which produces the thresholds it consumes, is Algorithm 2 in Appendix C.

With thresholds, offsets, and the mask fixed offline, each proposal needs K comparisons and $O(|S^+(x)|)$ inflation operations after one detector pass, without additional parameters or proposals.

Algorithm 1 Inference in the CertOD framework

Require: Image I, base detector f, calibrated thresholds $\{(\hat{q}_c^{\text{cls}}, \hat{q}_c^{\text{box}})\}_{c=1}^K$, offsets $\{\gamma_c\}_{c=1}^K$, and certifiability mask M.
Ensure: Set of certified pairs $\mathcal{P}$.

```
Run the base detector: {(B̂_j, p_j)}_{j=1}^J ← f(I).
P ← ∅.
for each proposal (B̂, p) do
    S ← {c ∈ K : p_c ≥ 1 − q̂_c^cls and M_c = 1}.
    k* ← arg max_k p_k; if M_{k*} = 1, set S ← S ∪ {k*}.
    for each category c ∈ S do
        Δw ← q̂_c^box(ŵ + γ_c),  Δh ← q̂_c^box(ĥ + γ_c).
        B̂_c ← [x̂_1 − Δw, ŷ_1 − Δh, x̂_2 + Δw, ŷ_2 + Δh].
        P ← P ∪ {(c, B̂_c)}.
    end for
end for
return P.
```

4 Experiments

Protocol. A frozen COCO-pretrained DETR-ResNet-50 is evaluated on BDD100K and the exact person/car/bus/truck label intersection of CODA2022 (Carion et al., 2020; Yu et al., 2020; Li et al., 2022). Class-agnostic matching uses IoU ≥ 0.5. Results average 100 image-disjoint splits: $80/20$ calibration/evaluation for fixed scores and $40/40/20$ tuning/calibration/evaluation for fitted offsets or joint temperatures. Class targets are 99% for person, 95% for car/bicycle/motorcycle, 90% for bus/truck/train, and 80% for traffic light, with equal class/box budget splits. Coverage is the joint event in Equation (1) on matched detections. These tables pool rows within images and therefore give empirical coverage; a separate one-row-per-image-class audit instantiates a sufficient rank design under exchangeable images, for a changed selected-row population (Appendix B.1). Appendix A gives the complete protocol and results.

4.1 BDD100K: Isolating Representation

Both true-class arms meet the joint target on all seven certifiable classes (Table 1); train has only two matched rows and is uncertifiable. Holding class sets, quantiles, scores, and budgets fixed, pairs

reduce retained true-pair pedestrian area by 6.8% at the median and 22.2% at p90 (exclusion rate 0.47%), with joint coverage 99.04% versus 99.11% for the single box. Both arms use the affine-supremum default; the coordinate-Bonferroni representation control is retained in Appendix D.7. The benefit belongs to a consumer retaining class–box associations: collapsing all pairs to one axis-aligned envelope recovers the single box exactly (Proposition 7).

Table 1: BDD100K representation comparison over 100 image-disjoint $40/40/20$ tune/calibration/evaluation splits, the protocol used whenever the affine offset is fitted. The outputs differ only in representation. Areas condition on true-class retention (the emitted basis of Appendix A.1) and are normalized by ground-truth area; Excl. is its complement.

		Joint coverage (%)			**Area (median)**		**Area (p90)**	
Category	**Nominal**	Single box	**CertOD**	**Excl.**	Single box	**CertOD**	Single box	**CertOD**
person	99.0%	99.11	**99.04**	0.47%	×4.78	×**4.45**	×8.53	×**6.64**
car	95.0%	96.09	**95.01**	2.50%	×3.61	×**3.07**	×8.26	×**5.08**
bicycle	95.0%	96.28	**95.98**	2.33%	×4.35	×**4.21**	×7.69	×**7.34**
motorcycle	95.0%	96.54	**95.38**	2.65%	×4.58	×**4.02**	×7.43	×**6.17**
bus	90.0%	91.88	**90.54**	5.03%	×2.76	×**2.38**	×5.29	×**4.36**
truck	90.0%	91.13	**90.39**	4.87%	×2.86	×**2.67**	×5.38	×**4.89**
traffic light	80.0%	91.10	**89.72**	0.20%	×2.90	×**2.72**	×5.83	×**4.05**
train	90.0%	*Uncertifiable: fewer than* 19 *calibration rows at* $\alpha_{\text{box}} = 0.05$						

4.2 CODA2022: Replication and Transfer

In-domain calibration reproduces aggregate target coverage on the four common classes and reduces conditional p90 area by 6.6%–35.6% (Table 2). Frozen BDD certificates instead miss all eight score–class transfer targets. A 5% target-domain prefix makes all four affine-supremum certificates finite and above target in mean, but larger prefixes leave a 0.45-point bus shortfall. The representation effect replicates; target-domain calibration remains necessary (Appendix A.3).

Table 2: CODA2022 replication over 100 image-disjoint $40/40/20$ tune/calibration/evaluation splits. The outputs differ only in representation. Areas condition on true-class retention (the emitted basis of Appendix A.1) and are normalized by ground-truth area; Excl. is its complement.

				Median area		**p90 area**		
Class	**Nominal**	**Pair joint**	**Excl.**	Single	**Pair**	Single	**Pair**	**p90 reduction**
person	99%	99.05%	0.51%	×6.33	×6.25	×8.73	×8.15	6.6%
car	95%	95.12%	2.50%	×4.18	×3.96	×11.18	×7.20	35.6%
bus	90%	90.33%	4.98%	×4.28	×4.20	×9.82	×8.39	14.5%
truck	90%	90.50%	5.06%	×3.65	×3.37	×8.59	×7.91	7.9%

4.3 Design Choices and Size-Stratified Coverage

Table 3 separates the sample-efficiency and geometric contributions of the score. Supremum scoring makes motorcycle certifiable on $100/100$ splits versus $0/100$ for coordinate Bonferroni, but does not shrink pedestrian boxes when both methods are feasible. Scale-relative scoring provides the area reduction: pedestrian median area falls from $\times 11.32$ to $\times 4.44$, and p90 from $\times 30.41$ to $\times 6.36$.

The affine offset is a per-class choice: on pedestrians pure relative scoring slightly improves on the affine arm, while positive offsets benefit large vehicles (19.4% median reduction for bus, 15.0% for truck). Where the class set is held fixed, unified joint scoring has 8.4% larger median area on the six categories satisfying that constraint; it is tighter only on traffic light, the one category whose tuned temperature also enlarges the class set ($+7.7\%$ median across all seven). Appendix A.6 has both comparisons in full.

Aggregate class coverage conceals size-dependent behavior (Figure 2, quintiles computed within the pedestrian class). Absolute scoring inflates the smallest pedestrians by $\times 30.47$ yet covers only

Table 3: Box score formulations on the pedestrian category ($\alpha = 0.01$), with the number of categories certifiable on all 100 splits and the number of true-class area-quintile strata with mean evaluation count at least 20 that miss their target by more than one point. No construction clears every stratum. Areas are on the same emitted basis as Table 1, with which the `affine_max` row agrees.

Score construction	Joint cov.	Area (med.)	Area (p90)	Cert. cls	Bad strata
Coordinate Bonferroni (`abs_bonf`)	99.10%	×10.48	×28.46	6/7	6
Supremum max (`abs_max`)	99.04%	×11.32	×30.41	**7/7**	6
Normalized max (`norm_max`)	99.04%	×10.52	×28.27	**7/7**	6
Scale-relative max (`rel_max`)	99.06%	×**4.44**	×**6.36**	**7/7**	**3**
Affine scale-adaptive (`affine_max`)	99.04%	×4.45	×6.64	**7/7**	6

96.91% of the largest quintile; affine scoring caps the smallest quintile at ×5.76 and lifts largest-quintile coverage to 97.95%, still short of the 99% target, and relative scoring gives 98.24% there. No tested score clears every class–size stratum: normalization improves geometric balance without establishing size-conditional validity (Appendix A.7).

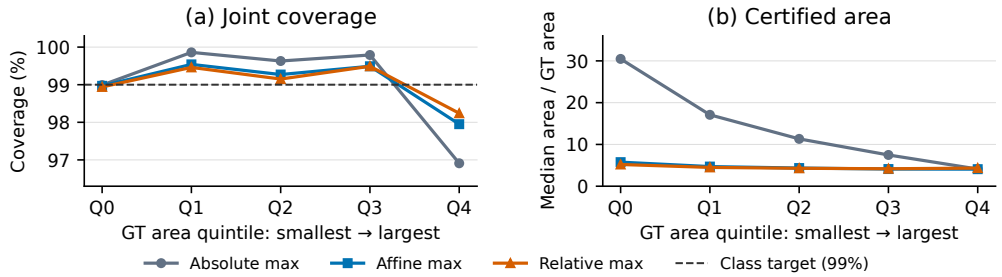


Figure 2: Pedestrian size audit across within-person ground-truth area quintiles. **(a)** Joint coverage; the dashed line is the class-level target, not a guarantee for each quintile. **(b)** Median certified area normalized by ground-truth area. The two panels expose the size imbalance of absolute scoring and the remaining large-object deficit after normalization. Values reproduce Table 10; the full numerical table is retained in the appendix.

4.4 Consumer Geometry and Guarantee Boundaries

A class-aware area reduction need not change whole-frame occupancy much. In the downstream geometric audit the smallest common envelope of all pairs equals the true-quantile single box at every reported digit, and a consumer using only its class's pair recovers at most 1.31 percentage points of frame area (0.11 for pedestrians, Appendix D.4). These are geometric measurements without a planner or braking evaluation, and they bound the practical reading of the per-class area gains.

Misclassified slices stay limited by true-class membership, and validity under shift does not follow from the aggregate guarantee (Appendices A.4 and A.9). Finally, 33.3% of pedestrians match no proposal: lowering the threshold raises the end-to-end coverage proxy from 59.4% to 66.1% and then saturates, so conditional box calibration cannot repair that ceiling (Appendix A.8).

5 Conclusion

CertOD retains class-specific boxes rather than collapsing hypotheses into one conservative envelope, cutting retained true-pair pedestrian p90 area by 22.2% on BDD100K with a replicated tail gain on CODA2022. The gain needs a class-aware consumer: the common axis-aligned envelope is exactly the single-box control. Coverage is conditional on a matched detection, a certifiable class and within-class exchangeability, and implies nothing about missed objects, background false positives, or scene-level, size-stratified or shifted-domain validity. The recall ceiling, the transfer failures, and the narrowing of the repair as the classifier sharpens bound post-hoc certification.

Ethics Statement

This work uses the publicly released BDD100K (Yu et al., 2020) and CODA2022 (Li et al., 2022) driving datasets under their published licences and introduces no new data collection involving human subjects. The images contain pedestrians and vehicles recorded in public road scenes; we use the released annotations as distributed and perform no identification of individuals. The intended application is safety-critical perception, so we note two risks explicitly. First, the coverage guarantee holds conditionally on detection, and a deployment that reads it as an unconditional safety certificate would overstate the assurance provided; Appendix A.8 and Section 4.2 quantify this bound. Second, the guarantee is distribution-free only under exchangeability, and both within-dataset shifts and BDD-to-CODA transfer show coverage degradation. Class-specific risk budgets encode a normative ordering over road users, which is a policy decision rather than a statistical one and should be set by the deploying authority.

Reproducibility Statement

Every theoretical claim is stated with its assumptions in Section 3 and proved in full in Appendix B. Calibration and inference are given as explicit pseudocode in Algorithm 2 and Algorithm 1, including the certifiability mask, the budget split, and the offset and temperature grids. Appendix A.1 and Appendix D fix everything the numbers depend on: the public detector checkpoint (`facebook/detr-resnet-50`), its operating point and proposal cap, the BDD100K-to-COCO label mapping, the CODA2022 label intersection, the class-agnostic $\mathrm{IoU} \geq 0.5$ matching rule, the per-class budgets, the $80/20$ and $40/40/20$ image-disjoint split protocols, and the 100-split averaging. The five box scores are defined in closed form in Section 3.4, the unified joint score in Appendix B.6, and the conformal quantile in Proposition 4. Both datasets and the detector are publicly available under their own licences.

AI Use Statement

In this work, we used generative AI tools for experiment planning, code assistance, result auditing, language editing, and manuscript formatting. The tools did not execute unreviewed analyses or supply unverified citations: all AI-assisted code was read and tested by the authors, every reported number was regenerated by the evaluation pipeline, and every reference was checked against its original source. The authors selected the research questions, approved all experiments, interpreted the evidence, and take full responsibility for the final text, claims, code, and artifacts.

References

Léo Andéol, Luca Mossina, Adrien Mazoyer, and Sébastien Gerchinovitz. Conformal object detection by sequential risk control. *arXiv preprint arXiv:2505.24038*, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Simon Barbarit–Gaboriau, Hind Laghmara, Rémi Boutteau, and Samia Ainouz. Ev-YOLO: Uncertainty-aware object detection via a unified evidential formulation. *arXiv preprint arXiv:2609.24668*, 2026.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 2020.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

Clément Fuchs, Tim Bary, and Benoît Macq. Localized conformal prediction for image classification with vision-language models. *arXiv preprint arXiv:2606.31577*, 2026.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Juyong Lee, Seungjin Jung, Jungmin Lee, Sunju Lee, and Jongwon Choi. Rethinking detection calibration: A coordinate and direction perspective. *arXiv preprint arXiv:2607.29040*, 2026.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, and Hang Xu. CODA: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

Kerri Lu, Dan M. Kluger, Stephen Bates, and Sherrie Wang. Conformal prediction sets for instance segmentation. *arXiv preprint arXiv:2602.10045*, 2026.

Christopher Ries, Moussa Kassem Sbeyti, Nicola Bianco, and Nadja Klein. Probabilistic object detection with conformal prediction. *arXiv preprint arXiv:2605.07549*, 2026.

Yaniv Romano, Evan Patterson, and Emmanuel Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Matteo Sesia and Emmanuel J. Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1), 2020.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. Technical report, On-line Compression Modelling project, Royal Holloway, University of London, 2003.

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

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B*, 86(4), 2024.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Chongzhe Zhang, Zifan Zeng, Qunli Zhang, Feng Liu, and Zheng Hu. Instance-level post hoc uncertainty quantification in object detection. *arXiv preprint arXiv:2606.04656*, 2026.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. DETRs beat YOLOs on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *ACM Computing Surveys*, 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.

APPENDIX

This appendix provides the complete proofs, sample-complexity derivations, algorithmic specifications, and extended empirical tables.

A EXTENDED EXPERIMENTS

A.1 EXPERIMENTAL SETUP

Data and detector. The primary benchmark is the BDD100K (Yu et al., 2020) validation split: $10{,}000$ images and $151{,}221$ ground-truth boxes mapped to 8 COCO classes (Lin et al., 2014). The detector is a frozen COCO-pretrained DETR-ResNet-50 (Carion et al., 2020; He et al., 2016), evaluated zero-shot and never fine-tuned; Appendix A.5 repeats the whole protocol on three further backbones. The same checkpoint and matching code are then applied to CODA2022 (Li et al., 2022), whose $9{,}768$ images carry $80{,}180$ annotated corner-case objects. Only $47{,}107$ of those fall in the exact label intersection person/car/bus/truck, and we keep it exact: we do not coerce cyclist to bicycle, tram to train, or tricycle to any COCO class, because a certificate for a category the detector was never calibrated on is not a certificate. Appendix D gives the BDD label mapping and the architecture.

Protocol. Every number is a mean over 100 random splits, and every split is image-disjoint. Comparisons that fit nothing use $80/20$ calibration/evaluation; comparisons that fit the affine offset or the joint temperature reserve a third part and use $40/40/20$ tune/calibration/evaluation, so a class receives 40% of its matched rows for calibration in those runs. The same two protocols are used on both datasets, and each table states which applies. Across-split standard deviations are reported in Appendix D.7.

Result provenance. The headline representation comparison (Tables 1 and 2) is produced by one evaluation path that fits `affine_max` on $40/40/20$ image splits and reports emitted true-pair area and all-output load. The score ablation (Table 3) and the size-stratified audit (Table 10) come from a second path, and the image-cluster bootstrap (Table 17) from a third, which uses the absolute coordinate-Bonferroni score and says so in its caption. No headline representation number is imported from the ablation or the bootstrap path, and no table mixes them. All three report certified area on the emitted basis defined above, so one construction carries one set of digits wherever it appears: the `affine_max` pedestrian row is $\times 4.45$ median and $\times 6.64$ at p90 in Tables 1, 3 and 12 alike. Two audits stand outside this and say so in their own captions, because they deliberately evaluate a different construction: the across-split variance table (Table 16) and the shift breakdown (Table 18) use the earlier absolute coordinate-Bonferroni certificate on all matched rows, and neither is compared against a headline number.

The split unit is the image. Objects within an image share a scene, pose, exposure and occlusion pattern and need not be exchangeable with each other. Every split is therefore taken at image granularity, with sizes targeted by row count so per-class calibration sets stay comparable. The effect is small but not nothing—moving from row-level to image-level splits shifts no certifiable class by more than 0.56 points—and a row-level split calibrates on near-duplicates of what it evaluates, with optimism not boundable from the numbers. Image-disjointness alone still does not establish exchangeability of the pooled object rows, so these primary results are empirical. The separate sampling-unit audit retains one deterministic row per image and class; under exchangeable images it supports the usual rank argument for that selected-row population. Neither protocol covers missed objects or background false positives.

Safety budgets. We instantiate the assignment of Section 3.1 by criticality: person 0.99 ($\alpha = 0.01$); car, bicycle and motorcycle 0.95 ($\alpha = 0.05$); bus, truck and train 0.90 ($\alpha = 0.10$); traffic light 0.80 ($\alpha = 0.20$). Each class budget is divided evenly between the two halves, $\alpha_c^{\mathrm{cls}} = \alpha_c^{\mathrm{box}} = \alpha_c/2$, so the pedestrian box is calibrated at the 99.5th percentile of its own true-class residuals.

What we report. *Joint coverage* is the measured rate of the event in Equation (1): the true class is in the emitted set *and* the box attached to that class contains the ground-truth box. It is the guarantee, and we report it in place of box containment wherever the two differ. *True-pair area* is the area of the true-class box divided by ground-truth area, summarized only on rows where that pair is actually emitted; class exclusion is reported separately. Every certified-area number in this paper is on this emitted basis, including the score ablation (Table 3) and the joint-score comparison (Table 12), so the same construction carries the same digits wherever it appears. Where two arms do not share a class set, as in Table 12, their emitted populations differ and the exclusion rate is printed beside the areas. Thus $\times 1$ would be a retained certificate the size of the object itself. For an oracle-free accounting of the whole output, *output load* sums the normalized areas of every emitted pair for a proposal, counting overlaps, and compares it with the one Set-Max box. It measures hypothesis burden, not occupied image area; frame-level union occupancy is reported separately in Appendix D.4. A class whose calibration set fails the mask M_c is reported as uncertifiable rather than silently dropped, and the fraction of splits on which it is certifiable travels with every number we quote for it.

A.2 Main Benchmark Comparison

Table 1 reports the comparison, and the first thing to check is validity. Both per-class arms clear their nominal target on every certifiable category. The values are means over 100 image-disjoint $40/40/20$ tune/calibration/evaluation splits, because the class-specific affine offsets are fitted on the tuning partition.

Two entries in the table are worth reading closely because they are the method's own machinery showing through. Traffic light over-covers by 9.72 points against its permissive 80% target, which follows from forcing the argmax into $S^+(x)$ (Section 3.3): at $\alpha = 0.20$ the conformal set alone would return empty sets, while the detector's top-1 label is right on 99.81% of these rows (Table 15), so semantic coverage cannot sit anywhere near 80%. The overshoot is what non-inferiority costs, and it is paid in the safe direction. Train, meanwhile, is the certifiability mask firing: every calibration fold contains fewer than the 19 rows its budget requires, so it is reported as uncertifiable rather than handed a margin it has not earned. This is the locality claimed in Section 3.3—the scarcity stays inside train's own pairs and leaves the other seven categories usable, where an infinite margin admitted into Equation (6) would empty every Set-Max box in the dataset.

The comparison we care about is the last four columns, because the single-box arm draws its quantiles from the same true-class buckets, the same class sets, the same budget and the same fitted affine-supremum score as CertOD. Only the representation differs. On that control, among rows retaining the true class, pair certificates reduce p90 pedestrian area by 22.2% ($\times 8.53 \rightarrow \times 6.64$) and the median by 6.8% ($\times 4.78 \rightarrow \times 4.45$), at a 0.47% exclusion rate and a joint-coverage difference of 0.07 points. The conditional p90 reduction is positive for every certifiable category, ranging from 4.6% on bicycle to 38.5% on car.

Every area in the main table is the cost of an *emitted* pair carrying the row's true class; rows excluding that class are not assigned a hypothetical box. The saving therefore reaches only a consumer that keeps the class label attached to the box. One that collapses the output to a single axis-aligned envelope recovers none of it: the smallest such envelope *is* the conservative single box, which we prove in Appendix B.5 and confirm to every reported digit in Appendix D.4. The 22.2% and the zero are one measurement read by two different consumers; both must be reported.

Counting every emitted hypothesis reverses the efficiency comparison: for pedestrians, the median summed pair-area load is 35.3% above one Set-Max box and its p90 is 150.6% above. This sum counts overlap and is not physical occupancy, but it makes explicit that CertOD reduces geometry per selected hypothesis rather than total output volume (Table 4).

Neither consumer escapes the underlying cost. Certification takes free space from 87.5% to 75.2% of the frame and joins 15.0% of the detection pairs that the raw boxes keep apart, against 12.9% under argmax conditioning (Appendix D.4). What the representation buys is a smaller certificate per hypothesis, not a less crowded scene.

One further caveat on the baseline. The deployable single-box arm draws its quantiles from the *predicted* class, and comparing against that version instead reports a larger median gain (21.1%) by moving the calibration conditioning at the same time; we report the clean control.

A.3 Second-Dataset Replication, Transfer, and Recovery

CODA2022 separates a representation that transfers from a certificate that does not.

The representation replicates. The label audit leaves four exact common classes; all four have finite supremum and coordinate-Bonferroni quantiles on every split. After in-domain calibration, the pair certificate meets each aggregate target (Table 2). The clean single-box control uses the same class sets, risk split, fitted offsets, and true-class affine-supremum quantiles. Conditional on true-class retention, pair output reduces median area for all four classes and p90 area by 6.6%–35.6%. The smaller effect on person median area (1.3%) and larger tail effect (6.6%) show that the representation mainly removes rare worst-class inflation rather than uniformly shrinking every box.

So do the score conclusions, including the unflattering one. Supremum scalarization reduces the finite-quantile threshold fourfold on CODA as well, and again does not uniformly reduce area. All five scores meet the four class-level targets, yet every score misses at least four class–size strata by more than one point. A second dataset therefore reproduces the BDD finding we would least like to be robust: class-conditional validity is not size-conditional validity.

Zero-shot transfer fails. We freeze the BDD100K class thresholds, box quantiles and affine offsets before reading any CODA label. Coordinate Bonferroni then falls below nominal by 0.08, 2.88, 2.27 and 0.76 points on person, car, bus and truck; affine supremum misses by 1.04, 4.50, 7.05 and 1.90

Table 4: Area-accounting audit for the affine-supremum representation comparison. True-pair reductions condition on that class being emitted; Excl. reports the omitted fraction. Output-load change compares the sum of all emitted pair areas (overlap counted) with one Set-Max box on every row. Positive load changes are additional hypothesis volume, not additional frame occupancy.

Dataset	Class	Excl.	$\lvert S\rvert$	True med. red.	True p90 red.	Load med. change	Load p90 change
BDD	person	0.47%	1.95	6.8%	22.2%	+35.3%	+150.6%
	car	2.50%	2.24	14.8%	38.5%	+86.5%	+154.3%
	bicycle	2.33%	2.44	3.2%	4.6%	+104.7%	+151.3%
	motorcycle	2.65%	3.00	12.2%	16.9%	+140.0%	+192.8%
	bus	5.03%	2.34	13.8%	17.7%	+102.5%	+190.6%
	truck	4.87%	2.15	6.5%	9.1%	+80.4%	+170.1%
	traffic light	0.20%	1.19	6.2%	30.6%	+0.3%	+28.7%
CODA	person	0.51%	1.18	1.3%	6.6%	+2.0%	+41.2%
	car	2.50%	1.53	5.4%	35.6%	+23.8%	+73.9%
	bus	4.98%	2.04	1.9%	14.5%	+83.7%	+123.6%
	truck	5.06%	1.83	7.6%	7.9%	+65.9%	+115.7%

points. All eight score–class targets fail, and $21/40$ size strata miss by more than one point. This is an exchangeability stress test, not a cross-domain guarantee.

Recalibration recovers the aggregate, not the strata. Calibrating on the target domain repairs the aggregate result, but aggregate success masks remaining deficits. In each of 100 splits we reserve one fixed 20% CODA holdout and draw nested calibration prefixes from the remaining 80%. With the BDD-tuned affine offset frozen, a 5% prefix (median 325 images and 1,430 matched objects) makes all four classes finite and yields joint coverage $99.24/95.10/92.32/91.07\%$ for person/car/bus/truck. At larger prefixes bus stabilises at 89.55%, only 0.45 points below nominal; we therefore call this recovery to the target's neighbourhood, not stable restoration. Coordinate Bonferroni first makes all four classes finite and meets their mean targets at a 25% prefix (median 1,630 images and 7,122 matched objects). Even then, five of twenty size strata miss their targets by over one point.

Appendix D.6 adds two further CODA results: a same-detector implementation of the Ries et al. (2026) two-step baseline, against which replacing only the safe envelope by per-class pairs cuts p90 area by 13.4%–41.4%; and the proposal-recall boundary, which is 77.9% for person and only 30.7% across the 36 novel CODA classes, so no closed-set certificate can repair those misses.

A.4 Diagnostic: Breakdown of the Argmax Leak

Table 5 breaks true pedestrians down by what the detector called them, which makes $\mathcal{E}_{\text{leak}}$ of Section 3.2 visible slice by slice. It repays being read column by column, because the containment column and the guarantee column tell different stories. Box containment under argmax conditioning does fall on two of the misclassified slices, to 86.7% where a pedestrian is called a traffic light and 85.6% where it is called a motorcycle. Weighted across all five off-diagonal slices it is 97.3%, because the largest of them—pedestrians called cars—has full containment. Pair certificates do restore containment to essentially 100% everywhere there.

But containment is not the guarantee. The guarantee is the conjunction, and on these slices it is capped by class-set membership: if `person` is in the set only 61.4% of the time, no box can lift joint coverage above 61.4%. The measured joint coverage of the pair certificate on the misclassified slices is 38.9% to 65.9%, not 100%. The repair is real for the box and partial for the guarantee, and any account that quotes only the containment column overstates it.

The counts also deserve their own column. The traffic-light slice is 256 row-instances observed across 95 of 100 splits, drawn from 14 distinct underlying objects; the across-split standard deviation of its joint coverage is 37.6 points. These slices motivate the method. They cannot carry a headline number.

Table 5: The argmax leak on true pedestrians, stratified by detector prediction. *Box cov.* columns are containment $P(B^* \subseteq \hat{B})$ only; *joint* is the conjunction with class-set membership, which is what the method guarantees and which $P(\text{person} \in S)$ bounds. n is the mean per-split count in splits where the cell occurs; N is the total row-instances observed and k the number of splits, so a cell with small n and large N is a handful of objects resampled many times.

Predicted as	n	N / k	**Box cov. argmax**	**Box cov. pair**	**Joint cov. pair**	$P(\text{person} \in S)$
person	1641.1	164113 / 100	99.60%	99.56%	**99.56%**	100.0%
car	14.6	1456 / 100	100.0%	100.0%	**65.91%**	65.9%
traffic light	2.7	256 / 95	**86.68%**	100.0%	**61.42%**	61.4%
bicycle	1.9	152 / 81	100.0%	100.0%	**60.47%**	60.5%
motorcycle	1.7	139 / 80	**85.62%**	97.92%	**38.85%**	40.9%
truck	1.3	82 / 63	100.0%	100.0%	**57.94%**	57.9%

Table 6: The same protocol on four DETR-family backbones. *Budget* is the pedestrian top-1 misclassification rate divided by $\alpha_{\text{person}} = 0.01$; it is a detector-error-to-budget ratio, not joint-certificate miscoverage. *Leak joint* is the pair certificate's joint coverage on the misclassified slices. We do not report mAP: measuring it needs a COCO evaluation protocol we do not implement, and quoting a published figure obtained under a different protocol beside numbers measured here is precisely the practice this table replaces.

Backbone	**Person rec.**	**Prec.**	**FP/img**	**Misclass.**	**Budget**	**Leak joint**
YOLOS-small	53.40%	28.31%	20.65	1.73%	**173%**	67.88%
DETR-ResNet-50	59.27%	35.00%	16.68	1.27%	**127%**	62.32%
DETR-ResNet-101	62.00%	41.19%	13.12	1.18%	**118%**	57.89%
Deformable DETR	56.61%	**65.91%**	**4.37**	**0.58%**	**58%**	12.74%

A.5 Does the Leak Generalise Across Backbones?

If a better-calibrated detector misclassifies few enough pedestrians, the premise of this paper weakens on its own. We therefore ran the whole protocol on three further backbones: a deeper DETR, the plain-ViT detector YOLOS (Fang et al., 2021), and Deformable DETR (Zhu et al., 2021). The constraint on which detectors can appear is an interface, not accuracy: the class-set half reads a full class score vector per detection, so the detector must emit set predictions with one row per query. An NMS pipeline that collapses each surviving box to a single label cannot supply that vector, so a YOLO-style detector cannot be evaluated here at all. This limits the supported detector interfaces.

The top-1 error rate exceeds the nominal pedestrian budget on three of four backbones. Deformable DETR is below it: at 0.58% its misclassification rate is 58% of α_{person}, and it is by a wide margin the cleanest detector in the comparison (65.9% precision, 4.4 false positives per image). The claim that residual misclassification in a strong detector is numerically larger than the target error rate is therefore true here for three architectures and false for one. None of these ratios determines the pair certificate's joint error, which also depends on class-set membership and box containment.

A second pattern was not one we anticipated. The pair certificate's joint coverage on the misclassified slices *falls* as the detector sharpens: 67.9%, 62.3%, 57.9%, then 12.7%. A better-calibrated classifier produces smaller conformal class sets—mean $|S|$ drops from 2.12 to 1.29—so on exactly the rows where it is wrong, the true class is less likely to be in the set at all. The repair weakens fastest on the detectors that need it least. That bounds the method's domain of usefulness to detectors whose classifier is imperfectly calibrated, which is worth knowing.

What is stable is the certificate itself: pair joint coverage lies in 99.04% to 99.08% against a 99% target on all four backbones, with median pedestrian area between $\times 8.18$ and $\times 9.36$.

A.6 Ablation of the Geometric Margin

Table 3 isolates each score construction on the pedestrian category ($\alpha = 0.01$). Supremum scalarization does *not* shrink the certified area relative to Bonferroni—it is slightly larger—which is what the

weak inter-edge correlation predicts. Scale-relative scoring is what compresses the envelope, taking the median pedestrian area from $\times 11.32$ to $\times 4.44$ and the p90 from $\times 30.41$ to $\times 6.36$.

Two qualifications follow, and both cut against the default we chose. `rel_max` is better than `affine_max` on every reported efficiency column and leaves three rather than six true-class size strata short. The affine offset earns its place per class rather than globally (Appendix A.7). And no construction clears every size stratum, which is why Table 3 reports a count of failing strata rather than a pass mark: on this axis there is nothing to award.

Where the sample-efficiency argument bites. A finite quantile at level α needs $n \geq \lceil 1/\alpha \rceil - 1$, so under the split budget the supremum score needs 199 pedestrian instances against Bonferroni's 799, and 39 against 159 for the $\alpha = 0.05$ classes; the measured frontier lands on those values for all twelve (class, method) pairs. The argument bites only if a class falls in the window. Calibration receives 40% of a class's rows—person 3,339, car 26,210, bicycle 193, motorcycle 100—so exactly one class does, and it is precisely the class `abs_bonf` certifies on $0/100$ splits. Bicycle clears the Bonferroni threshold by 34 instances, so the boundary sits close to this dataset.

On margin width the usual claim runs the wrong way: where both are feasible, total pedestrian padding at $n = 4000$ is $183.0\,\mathrm{px}$ for the supremum score against $178.4\,\mathrm{px}$ for Bonferroni. The supremum score buys the window, not the width.

The affine offset is a per-class device. Sweeping γ over $[0, 800]\,\mathrm{px}$ as a fixed score (Appendix D.2) gives a monotone trade-off, not an interior optimum: for pedestrians, median area, p90 and tiny-object area all increase with c while large-object coverage falls, so $c = 0$ dominates on every column. There is no tiny-object coverage cliff at $\gamma = 0$—Q_0 coverage is 98.94% there and 98.99% at positive offsets, a spread of 0.05 points. The offset earns its place on *large vehicles*: bus median area falls 19.4% ($\times 2.93 \rightarrow \times 2.36$ at $\gamma = 100$) and truck 15.0%, at unchanged coverage. The tuning split recovers exactly this per-class structure on its own: the offset it selects has median 0 for person, $10\,\mathrm{px}$ for bicycle, car, motorcycle and traffic light, and $100\,\mathrm{px}$ for bus and truck. The selection is made independently on each split and is not constant—the corresponding means are 4.9, 23.3, 14.1, 13.6, 16.0, 108.3 and $78.8\,\mathrm{px}$—which is why the pedestrian `affine_max` row of Table 3 is close to, but not identical with, `rel_max`.

The unified joint score is worse where the class set is held fixed. Both constructions hold the nominal level, as Theorem 8 requires. On the six categories whose tuned τ satisfies the equal-set-size constraint, the unified score is *worse* in median area on every one, by 2.0% to 9.3%, with a median of 8.4% (Table 12); no certifiable category is tighter at no larger set size. Traffic light is the single exception and we report it rather than drop it: there the unified score is 15.8% *tighter*, and it is also the one category whose tuned τ admits a larger class set ($|S|$ $1.19 \rightarrow 1.27$), so the exception is an instance of the same mechanism rather than a counterexample to it. Across all seven certifiable categories the median change is $+7.7\%$. The sweep shows why: as τ grows the pedestrian median area falls from $\times 8.61$ at $\tau = 1$ to a floor of $\times 3.85$ from $\tau = 3$, while mean $|S|$ rises from 1.37 to 7.00—every certifiable category. The gain is realised exactly where the class set stops carrying information. Tuning τ to minimise area without constraining the set reports a 16.3% reduction while mean $|S|$ goes from 2.19 to 6.63 of 8; that is the number a reader would obtain from "τ fitted on $\mathcal{D}_{\mathrm{tune}}$", and it is not a margin saving.

A.7 Size-Stratified Audit

Class-conditional coverage conceals size-conditional failure (Table 10): `abs_max` under-covers large pedestrians by 2.09 points while inflating tiny ones by $\times 30.47$. Scale-adaptive scoring greatly reduces the spread—`affine_max` caps tiny-pedestrian area at $\times 5.76$ and recovers 1.04 points on large objects—without eliminating the deficit, still missing Q_4 pedestrians by 1.05 points. We therefore do not claim size-stratified validity: across the five named scores and eight affine offsets, none meets every certifiable class's target with every stratum within one point, the best leaving three strata short. A stratum the calibration never optimises is not guaranteed by it. The largest absolute-score deficit is instead the largest-car quintile, missed by 11.3–12.4 points.

A.8 The Recall Ceiling

Every guarantee in this paper is conditional on the object having been detected, $P(\text{Covered} \mid \text{Detected}) \geq 1 - \alpha$, so the end-to-end quantity is $P(\text{Covered} \mid \text{Detected}) \cdot \text{Recall}$. We write this $\text{Cov} \times \text{Rec}$ rather than $P(\text{Safe})$: it is a coverage proxy on the evaluated event, not a statement about a vehicle, since it excludes background false positives and says nothing about all objects in an image being covered simultaneously.

Lowering the proposal threshold is the natural remedy, and it trades recall against false positives, so controlling both ends is a two-stage design: Stage 1 picks the threshold whose miss rate over vulnerable road users is provably controlled, and Stage 2 filters the false positives the lower threshold admits. We implement Stage 1—the loss is the per-image miss rate over vulnerable road users, the image is the exchangeable unit, and conformal risk control (Angelopoulos et al., 2024) selects the least permissive λ clearing $(n\hat{R}(\lambda) + B)/(n+1) \leq \beta$, writing β for the miss-rate budget to keep it apart from the per-class certificate budget α_c, with the loss and its monotonicity stated in Appendix B.7—and not Stage 2. Running the detector at 0.01 with 100 proposals per image gives $817{,}326$ proposals, of which 12.5% match any ground truth.

The ceiling—the fraction of each class matched by *any* proposal—is 66.67% for person, 60.06% for bicycle, 70.43% for motorcycle and 47.86% for traffic light. No threshold recovers an object that was never proposed. A 95% pedestrian recall target is therefore not 36 points of tuning away from 59.27%; it is 28 points above the detector's ceiling. Conformal risk control agrees: of five miss-rate targets only $\beta = 0.40$ admits a threshold on any split, landing at $\lambda = 0.200$ with held-out risk 0.4008 ± 0.0145—the guarantee bounds the *expected* risk, so a single held-out estimate sitting 0.0008 above the target is the expected behaviour, not a violation—and costing 28.9 false positives per image at 29.5% precision on the risk-control evaluation fold. That fold is the $3{,}560$ images that contain a vulnerable road user, which is the exchangeable unit the loss is defined on; Table 11 reports 22.2 false positives per image at the same λ over all evaluation images, and the two populations are why the counts differ. The targets $\beta \in \{0.05, 0.10, 0.20, 0.30\}$ are cleared by no threshold in $[0.01, 0.94]$ on any split.

The useful positive finding is that lowering the threshold costs the certificate nothing. Across $\lambda \in [0.01, 0.50]$, pedestrian joint coverage stays within 98.87%–99.14% and median certified area within $\times 4.44$–$\times 4.69$; the entire price of expanded recall is false positives, 71.6 per image at 0.01 against 17.1 at 0.30. $\text{Cov} \times \text{Rec}$ improves from 59.4% to 66.1% and then stops, pinned by the ceiling. Post-hoc conformal prediction cannot repair a detector's recall, and we think that is the most important limitation in this paper.

We do not implement Stage 2. Sizing it is straightforward from the above: at the only feasible target it must remove roughly 70% of surviving detections while retaining recalled vulnerable road users.

A.9 Distribution Shift

Conformal validity relies on exchangeability, which driving environments violate. Section 4.2 broke it once already, by moving the calibration set to a different dataset, and the certificate did not survive the move. Here we break it inside a single dataset, where the images come from one collection and only a scene attribute changes—a far milder stress, and the one a deployed system meets every night. We evaluate daytime-to-night, clear-to-adverse-weather and unoccluded-to-occluded shift; the first two are image-level attributes whose pools share no image, while the third is an *object* attribute whose pools share $8{,}475$ images, making it a stratification rather than a domain shift. Appendix D.8 gives the seven-category breakdown.

The advantage over argmax conditioning is large: at night, bus coverage rises from 37.00% under argmax conditioning to 83.55% under pair certificates, and truck from 50.48% to 80.23%. But the pair certificates are themselves below their 90% nominal target there, so the supported claim is far more graceful degradation, not retained validity. Pedestrians hold up best, 98.65% at night, with certified area falling from $\times 5.54$ to $\times 4.22$. Two categories are uncertifiable under the occlusion stratification because the unoccluded calibration pool leaves them below the sample-size floor; their 0.00% is calibration scarcity, not a coverage measurement, so we print the certifiability fraction beside every number.

B Formal Proofs and Theoretical Derivations

B.1 Exchangeability Unit and Scope of the Guarantee

Fix the detector, matching rule, and all hyperparameters chosen on the disjoint tuning split. For each certifiable class c, the finite-sample arguments in this paper assume that, conditional on the true class being c, the calibration matched rows and a fresh matched test row are exchangeable for the score being calibrated. Under this assumption, the test score has a uniform rank among the $n_c + 1$ scores when ties are absent (with the usual conservative conformal handling of ties). Classification and box scores on the same row may be dependent; Lemma 1 requires no independence.

Image-disjoint assignment ensures that no image contributes to both calibration and evaluation, but this alone does not imply exchangeability of pooled object scores. An independently sampled image may contain several matched objects of the same class whose scores are dependent. As a simple counterexample, suppose each image contributes two class-c objects with an identical continuous score. Two calibration images then supply four scores (U_1, U_1, U_2, U_2), while a new image supplies (U_3, U_3). The nominal 80% conformal threshold from the four calibration scores is $\max(U_1, U_2)$, but a test object is covered with probability $P(U_3 \le \max(U_1, U_2)) = 2/3$ for independent, identically distributed image scores.

One sufficient design for the usual object-level rank argument is to sample images independently, apply the same image-wise matching and selection rule throughout, and let each image contribute at most one matched row of class c. Then, conditional on an image contributing such a row, the class-c scores come from distinct exchangeable images. The primary experiments do not implement this design: their image-disjoint splits are targeted by total row count and retain all matched rows inside an assigned image. Their pooled class-c rows therefore do not inherit the rank condition merely from image-disjointness, and their coverage values are empirical measurements.

Audit implementing the sufficient design. For a separate $40/40/20$ experiment, after assigning whole images we retain at most one matched row for each image and true class in every partition. The rule is fixed before score calibration: choose the row with the smallest matched ground-truth annotation index, using original row order only to break a tie. It does not inspect classifier or box conformity scores. Tuning, calibration, and evaluation all apply the same rule. Consequently, under the assumption that images are exchangeable, the selected rows of class c are exchangeable across the class-contributing images and satisfy the sufficient design above. This changes the estimand: the guarantee concerns one deterministically selected matched detection per contributing image–class, not a uniformly drawn object from the pooled matched detections.

Over 100 splits, the affine-supremum construction remains finite on every split for all non-train BDD classes and all four CODA classes. BDD retains on average 1066 person, 137 bicycle, and 86 motorcycle calibration images; the other certifiable classes range from 398 to 3849, while train has fewer than one and remains uncertifiable. Mean joint coverage is 99.06% for person and at or above nominal for the other certifiable classes except motorcycle, whose 94.83% estimate is 0.17 points short with a 4.32-point split standard deviation. CODA retains 336–2994 calibration images per class and obtains $99.09/95.16/90.57/90.49\%$ for person/car/bus/truck. These fixed-dataset estimates are not the theorem, whose probability is over a fresh exchangeable calibration–test draw; they are means over the 100 splits, and the per-class calibration-image counts above bound how much any one split can move them.

Thus the theorems remain conditional on within-class matched-row exchangeability. The strict audit supplies one implementable route to that condition under exchangeable images; the primary pooled-object tables do not. Neither statement extends to missed objects, uncertifiable classes, or all objects in an image simultaneously.

B.2 Supremum Scalarization: Finite-Sample Multi-Coordinate Containment

Definition 2 (Outward coordinate residuals)**.** Let $B^* = [x_1^*, y_1^*, x_2^*, y_2^*]^\top \in \mathbb{R}^4$ be the ground-truth box and $\hat{B}(x) = [\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2]^\top \in \mathbb{R}^4$ the box emitted by the base detector $f(x)$. The outward residual vector $s(x, B^*) = [s_0, s_1, s_2, s_3]^\top \in \mathbb{R}^4$ is

$$s_0 = \hat{x}_1 - x_1^*, \quad s_1 = \hat{y}_1 - y_1^*, \quad s_2 = x_2^* - \hat{x}_2, \quad s_3 = y_2^* - \hat{y}_2. \tag{12}$$

Lemma 3 (Exact containment equivalence). *For any scalar margin $q \in \mathbb{R}$, define the outward inflated box $\hat{B} \oplus q = [\hat{x}_1 - q, \hat{y}_1 - q, \hat{x}_2 + q, \hat{y}_2 + q]$. Then the containment event $B^* \subseteq \hat{B} \oplus q$ is equivalent to the scalar inequality*

$$B^* \subseteq \hat{B} \oplus q \iff \max_{j \in \{0,1,2,3\}} s_j(x, B^*) \leq q. \tag{13}$$

Proof. By the definition of box inclusion, $B^* \subseteq \hat{B} \oplus q$ holds if and only if all four boundary conditions hold simultaneously:

$$x_1^* \geq \hat{x}_1 - q \iff \hat{x}_1 - x_1^* \leq q \iff s_0 \leq q, \tag{14}$$
$$y_1^* \geq \hat{y}_1 - q \iff \hat{y}_1 - y_1^* \leq q \iff s_1 \leq q, \tag{15}$$
$$x_2^* \leq \hat{x}_2 + q \iff x_2^* - \hat{x}_2 \leq q \iff s_2 \leq q, \tag{16}$$
$$y_2^* \leq \hat{y}_2 + q \iff y_2^* - \hat{y}_2 \leq q \iff s_3 \leq q. \tag{17}$$

Equations (14)–(17) hold simultaneously if and only if $\max\{s_0, s_1, s_2, s_3\} \leq q$. □

The direct multi-coordinate coverage bound is stated and proved here.

Proposition 4 (Finite-sample multi-coordinate coverage). *Let $\mathcal{D}_{\mathrm{cal},c} = \{(X_i, Y_i) : c_i^* = c\}_{i=1}^{n_c}$ be n_c exchangeable calibration samples of class c with $Y_i = (c_i^*, B_i^*)$, and let (X_{n_c+1}, Y_{n_c+1}) be a fresh test sample exchangeable with $\mathcal{D}_{\mathrm{cal},c}$ conditional on $c_{n_c+1}^* = c$. Define the supremum nonconformity score $S_{\max}(X_i, B_i^*) = \max_{j \in \{0,1,2,3\}} s_j(X_i, B_i^*)$ and let*

$$\hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)} = \mathrm{Quantile}\left(\frac{\lceil (n_c+1)(1-\alpha_{\mathrm{box}}) \rceil}{n_c};\ \{S_{\max}(X_i, B_i^*)\}_{i=1}^{n_c} \right). \tag{18}$$

Then, assuming continuous scores so that ties occur with probability zero,

$$1 - \alpha_{\mathrm{box}} \leq P\left(B_{n_c+1}^* \subseteq \hat{B}(X_{n_c+1}) \oplus \hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)} \;\middle|\; c_{n_c+1}^* = c \right) < 1 - \alpha_{\mathrm{box}} + \frac{1}{n_c+1}. \tag{19}$$

Proof. Because $(X_i, Y_i)_{i=1}^{n_c+1}$ are exchangeable conditional on $c^* = c$, the scalar scores $\{S_1, \ldots, S_{n_c}, S_{n_c+1}\}$ are exchangeable real-valued random variables. By the conformal prediction theorem (Vovk et al., 2005; Angelopoulos & Bates, 2021), the rank of the test score S_{n_c+1} among $\{S_1, \ldots, S_{n_c+1}\}$ is uniform on $\{1, \ldots, n_c+1\}$.

Let $k = \lceil (n_c+1)(1-\alpha_{\mathrm{box}}) \rceil$, so that $\hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)}$ is the k-th smallest value in $\{S_1, \ldots, S_{n_c}\}$. Then $S_{n_c+1} \leq \hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)}$ holds if and only if the rank of S_{n_c+1} is at most k, so

$$P\left(S_{n_c+1} \leq \hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)} \mid c_{n_c+1}^* = c \right) = \frac{k}{n_c+1} = \frac{\lceil (n_c+1)(1-\alpha_{\mathrm{box}}) \rceil}{n_c+1}. \tag{20}$$

Using $z \leq \lceil z \rceil < z + 1$,

$$1 - \alpha_{\mathrm{box}} \leq \frac{\lceil (n_c+1)(1-\alpha_{\mathrm{box}}) \rceil}{n_c+1} < 1 - \alpha_{\mathrm{box}} + \frac{1}{n_c+1}. \tag{21}$$

By Lemma 3, the event $S_{n_c+1} \leq \hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)}$ is the containment event $B_{n_c+1}^* \subseteq \hat{B}(X_{n_c+1}) \oplus \hat{q}_{1-\alpha_{\mathrm{box}}}^{(c)}$, which gives Equation (19). □

Comparison with coordinate-wise Bonferroni splitting. Coordinate-wise calibration (Ries et al., 2026) divides the budget α_{box} into four sub-budgets $\alpha_j = \alpha_{\mathrm{box}}/4$ and computes four marginal quantiles $\hat{q}_j = \mathrm{Quantile}(1 - \alpha_{\mathrm{box}}/4; \{s_j\})$. Spatial coverage then relies on the union bound

$$P\left(\bigcup_{j=0}^{3} \{s_j > \hat{q}_j\} \right) \leq \sum_{j=0}^{3} P(s_j > \hat{q}_j) \leq 4 \cdot \frac{\alpha_{\mathrm{box}}}{4} = \alpha_{\mathrm{box}}. \tag{22}$$

The union bound is tight only when the four violation events are mutually disjoint. The actual miscoverage probability is

$$P(\mathrm{miscoverage}) = \sum_{j=0}^{3} P(s_j > \hat{q}_j) - \sum_{j<k} P\big(\{s_j > \hat{q}_j\} \cap \{s_k > \hat{q}_k\}\big) + \cdots, \tag{23}$$

Table 7: Sample-complexity frontier: minimum calibration sample size $n_{\min}$ required to avoid a vacuous ($+\infty$) quantile, under coordinate Bonferroni and under supremum scalarization. The $D = 4$ column is the axis-aligned box this paper calibrates; the wider columns show how the Bonferroni requirement grows with the number of split coordinates, and carry no claim about those representations.

Risk	**Nominal**	**Coordinate Bonferroni** ($n_{\min}$)			**CertOD**
α	**coverage**	2D box (D=4)	3D box (D=7)	Poly (D=8)	**supremum** (D=1)
0.20	80.0%	19	34	39	**4**
0.10	90.0%	39	69	79	**9**
0.05	95.0%	79	139	159	**19**
0.02	98.0%	199	349	399	**49**
0.01	99.0%	399	699	799	**99**
0.005	99.5%	799	1399	1599	**199**

so overlap among coordinate failures determines the slack in the bound. The measured linear residual correlation ($\bar{\rho} \in [0.028, 0.064]$ on BDD100K) does not by itself identify those tail intersections. Our empirical comparison therefore measures box efficiency directly. Independently of that efficiency comparison, splitting the budget across four coordinates raises the finite-quantile sample threshold by approximately $4\times$; supremum scalarization avoids that multiplicity penalty.

B.3 Sample Complexity of the Geometric Margin

Theorem 5 (Conformal feasibility frontier)**.** *Let $\alpha \in (0, 1)$ be a miscoverage budget. A finite conformal quantile $\hat{q}_{1-\alpha} < +\infty$ can be computed from n calibration instances if and only if*

$$n \geq \left\lceil \frac{1}{\alpha} \right\rceil - 1. \tag{24}$$

Proof. The conformal quantile at level $1 - \alpha$ is the k-th order statistic of the augmented set $\mathcal{S}_{\text{aug}} = \{S_1, \ldots, S_n\} \cup \{+\infty\}$ with $k = \lceil (n + 1)(1 - \alpha) \rceil$. Since $\mathcal{S}_{\text{aug}}$ has $n + 1$ elements whose largest is $+\infty$, a finite quantile requires $k \leq n$. Because n is an integer, $\lceil z \rceil \leq n$ is equivalent to $z \leq n$, hence

$$\begin{aligned} (n+1)(1-\alpha) &\leq n \\ 1 - \alpha &\leq n\alpha \\ n &\geq \frac{1-\alpha}{\alpha} = \frac{1}{\alpha} - 1, \end{aligned} \tag{25}$$

and integrality gives $n \geq \lceil 1/\alpha \rceil - 1$. □

Corollary 6 (Bonferroni multiplicity penalty in D dimensions)**.** *If a perception envelope in D dimensions is calibrated by splitting the budget equally across coordinates, $\alpha_j = \alpha/D$, then the minimum calibration sample size for a finite envelope is*

$$n_{\text{Bonferroni}}(D, \alpha) \geq \left\lceil \frac{D}{\alpha} \right\rceil - 1, \tag{26}$$

whereas supremum scalarization, which projects all D coordinates onto the single score $S_{\max} = \max_{j=1,\ldots,D} s_j$, requires only

$$n_{\text{CertOD}}(\alpha) \geq \left\lceil \frac{1}{\alpha} \right\rceil - 1, \tag{27}$$

a reduction of $\mathcal{R}(D) = n_{\text{Bonferroni}}(D, \alpha)/n_{\text{CertOD}}(\alpha) \approx D$.

Table 7 evaluates both thresholds across the budgets and dimensions relevant here. The arithmetic is unconditional; whether any class of a given dataset falls between the two columns is not, which is why Appendix A.6 measures the window rather than citing the ratio.

Where the window actually falls. The thresholds above are arithmetic; what decides whether they matter is the calibration budget each class receives. Under our protocol a class gets 40% of its matched rows, so on BDD100K exactly one category lands between the two thresholds—motorcycle, with 100 instances against a supremum requirement of 39 and a Bonferroni requirement of 159. Bicycle clears the Bonferroni threshold by 34 instances. The argument is therefore dataset-specific, and on a corpus with more instances per rare class it would carry no weight at all.

- **2D detection ($D = 4$).** Under the split budget of Section 3.5, the pedestrian box level is $\alpha_{\text{box}} = 0.005$: Bonferroni splitting then requires $n_c \geq 799$ calibration instances against $n_c \geq 199$ for one supremum quantile. Spending the whole class budget on the box instead gives 399 and 99. Both conventions appear in the literature; we use the first convention unless specified.
- **Beyond $D = 4$.** Corollary 6 is arithmetic and applies to any coordinatewise budget split, but the reduction to a single supremum quantile is not. It rests on Lemma 3, whose equivalence between box containment and four scalar inequalities is specific to axis-aligned boxes. A representation whose containment event is not a conjunction of coordinate inequalities—an oriented 3D box, for instance—would need an envelope construction we do not supply, so we claim the frontier only for $D = 4$.

B.4 Decoupled Pair Conditional Validity

First, the budget decomposition of Lemma 1 follows directly:

Proof. By Boole's inequality, $P(\mathcal{E}_c^{\text{cls}} \cup \mathcal{E}_c^{\text{box}} \mid c^* = c) \leq P(\mathcal{E}_c^{\text{cls}} \mid c^* = c) + P(\mathcal{E}_c^{\text{box}} \mid c^* = c) \leq \alpha_c^{\text{cls}} + \alpha_c^{\text{box}} = \alpha_c$. Taking complements gives the claim. □

We restate Corollary 2 of the main text and give the complete proof.

Corollary 2 (Decoupled pair conditional validity)**.** *Let $\mathcal{C}_{\text{pair}}(x) = \{(c, \hat{B}_c(x)) : c \in S^+(x)\}$ be the certified pairs produced by CertOD under the budget allocation $\alpha_c = \alpha_c^{\text{cls}} + \alpha_c^{\text{box}}$, and let $M_c \in \{0, 1\}$ be the certifiability mask*

$$M_c = \mathbb{I}\left(n_c \geq \lceil 1/\alpha_c^{\text{cls}} \rceil - 1 \ \wedge \ n_c \geq \lceil 1/\alpha_c^{\text{box}} \rceil - 1\right). \tag{28}$$

For any category $c \in \mathcal{K}$ with $M_c = 1$,

$$P\left((c, \hat{B}_c(x)) \in \mathcal{C}_{\text{pair}}(x) \ \wedge \ B^* \subseteq \hat{B}_c(x) \,\middle|\, c^* = c\right) \geq 1 - \alpha_c. \tag{29}$$

Proof. Let $(x, (c^*, B^*))$ be a test detection with $c^* = c$. By the definition of $\mathcal{C}_{\text{pair}}(x)$, the pair $(c, \hat{B}_c(x))$ belongs to $\mathcal{C}_{\text{pair}}(x)$ if and only if $c \in S^+(x)$, so the safety event and its complement are

$$\mathcal{E}_{\text{safe}} = \left\{c \in S^+(x) \ \wedge \ B^* \subseteq \hat{B}_c(x)\right\}, \qquad \mathcal{E}_{\text{safe}}^{\complement} = \left\{c \notin S^+(x) \ \vee \ B^* \not\subseteq \hat{B}_c(x)\right\}. \tag{30}$$

Applying Boole's inequality conditional on $c^* = c$,

$$P(\mathcal{E}_{\text{safe}}^{\complement} \mid c^* = c) \leq P\left(c \notin S^+(x) \mid c^* = c\right) + P\left(B^* \not\subseteq \hat{B}_c(x) \mid c^* = c\right). \tag{31}$$

We bound the two terms separately.

Classification set error. The class candidate set is obtained by thresholding the softmax probabilities,

$$S(x) = \left\{k \in \mathcal{K} : p_k(x) \geq 1 - \hat{q}_k^{\text{cls}} \ \wedge \ M_k = 1\right\}, \tag{32}$$

with nonconformity score $s_i^{\text{cls}} = 1 - p_c(x_i)$ evaluated on calibration instances with $c_i^* = c$. Split conformal prediction at level $1 - \alpha_c^{\text{cls}}$ gives

$$P\left(1 - p_c(x) \leq \hat{q}_c^{\text{cls}} \mid c^* = c\right) \geq 1 - \alpha_c^{\text{cls}} \implies P\left(c \notin S^+(x) \mid c^* = c\right) \leq \alpha_c^{\text{cls}}, \tag{33}$$

because $S(x) \subseteq S^+(x)$.

Box containment error. The margin $\hat{q}_c^{\text{box}}$ is calibrated by supremum scalarization on instances with $c_i^* = c$, so Proposition 4 gives

$$P\left(B^* \not\subseteq \hat{B}(x) \oplus \mathbf{q}_c^{\text{box}}(x) \mid c^* = c\right) \leq \alpha_c^{\text{box}}. \tag{34}$$

Substituting both bounds into Equation (31) yields $P(\mathcal{E}_{\text{safe}}^{\complement} \mid c^* = c) \leq \alpha_c^{\text{cls}} + \alpha_c^{\text{box}} = \alpha_c$, and taking complements proves Equation (29). □

No test-time oracle is required. Although calibration partitions by true class, inference evaluates only fixed thresholds and margins for candidates in $S^+(x)$. The consumer receives all pairs without knowing the true category; the joint event that the true-class pair is emitted and contains B^* has probability at least $1 - \alpha_{c^*}$.

Non-inferiority under forced argmax. When α_c is permissive, for example $\alpha = 0.20$ for traffic lights, pure conformal thresholding returns an empty candidate set on 4.36% of detections and attains 89.98% coverage, below the raw top-1 accuracy of 99.81% (Table 15); augmenting with the argmax raises class coverage to 99.82%. CertOD augments the set as

$$S^+(x) = S(x) \cup \left\{ k^*(x) : k^*(x) = \arg\max_{k \in \mathcal{K}} p_k(x),\ M_{k^*(x)} = 1 \right\}. \tag{35}$$

Because $S(x) \subseteq S^+(x)$, the membership event is monotone,

$$P(c \in S^+(x) \mid c^* = c) \geq P(c \in S(x) \mid c^* = c) \geq 1 - \alpha_c^{\text{cls}}, \tag{36}$$

and the top-1 label is retained whenever it is certifiable. Thus semantic coverage is no worse than the base detector within the certified label space; this operation does not guarantee geometric containment.

B.5 The Axis-Aligned Envelope of the Pairs

The pair representation reduces certified area only for a consumer that keeps the class label attached to each box. One that insists on a single axis-aligned envelope around every hypothesis recovers nothing, and the reason is geometric rather than statistical.

Proposition 7 (The axis-aligned envelope of the pairs is the single box)**.** *Let $\hat{B}_c(x) = \hat{B}(x) \oplus \mathbf{q}_c^{\text{box}}(x)$ for $c \in S^+(x)$, all sharing the base box $\hat{B}(x)$. The smallest axis-aligned rectangle containing $\bigcup_{c \in S^+(x)} \hat{B}_c(x)$ is obtained by taking the coordinatewise maximum outward margin, and is therefore exactly $\hat{B}_{\text{Set-Max}}(x)$ of Equation* (6)*.*

Proof. Each $\hat{B}_c$ moves edge j outward by $q_{c,j}^{\text{box}}(x) \leq \bigvee_{c'} q_{c',j}^{\text{box}}(x)$, so every $\hat{B}_c$ lies inside the stated rectangle. For each edge j, some class attains the maximum, so any axis-aligned rectangle containing every pair must reach that edge. The rectangle is therefore minimal. The geometric union itself need not be rectangular when the class-specific envelopes are not nested. □

A consumer that requires one axis-aligned envelope therefore pays exactly what the single box costs. The area advantage reported in Section 4.1 is real but conditional on retaining the class-pair structure; a consumer able to act on the non-rectangular union itself is not evaluated in this paper.

B.6 Unified Joint Scoring

We define the unified alternative of Section 3.5, prove its validity, and state its limitations.

Theorem 8 (Unified joint coverage)**.** *Let $\hat{q}_{1-\alpha_c}^{\text{joint}}$ be the $(1 - \alpha_c)$ conformal quantile of $S_{\text{joint}}(x, (c^*, B^*)) = \max\{(1 - p_{c^*}(x))/\tau_{c^*},\ S_{\text{affine}}^{(c^*)}(x, B^*)\}$ over calibration instances of true class c, with γ_c and $\tau_c > 0$ fixed before calibration. Define $S(x) = \{c : (1 - p_c(x))/\tau_c \leq \hat{q}_{1-\alpha_c}^{\text{joint}}\}$ and inflate $\hat{B}_c(x)$ by $\hat{q}_{1-\alpha_c}^{\text{joint}}(\hat{w} + \gamma_c, \hat{h} + \gamma_c, \hat{w} + \gamma_c, \hat{h} + \gamma_c)$. Then $P(c^* \in S(x) \wedge B^* \subseteq \hat{B}_{c^*}(x) \mid c^* = c) \geq 1 - \alpha_c$.*

Proof. Conditional on $c^* = c$ the scalar scores of the calibration instances and a fresh test instance are exchangeable, so the standard split-conformal argument gives $P(S_{\text{joint}} \leq \hat{q}_{1-\alpha_c}^{\text{joint}} \mid c^* = c) \geq 1 - \alpha_c$. Since S_{joint} is a maximum of two terms, that event holds if and only if both terms are below the threshold; the first is membership $c \in S(x)$ and the second is outward containment. The conjunction therefore inherits the bound. □

What the theorem does not give. It certifies the conjunction with one quantile instead of two, removing the union bound. It says nothing about whether the resulting box is *smaller* than the one the Boole split produces, and the construction has a degenerate limit that makes any area comparison meaningless unless the class set is held fixed: as $\tau_c \to \infty$ the class term vanishes, the single quantile relaxes from the $\alpha_c/2$ level to α_c and the box shrinks, while $S(x)$ grows to contain every category. Coverage still passes, because a set containing everything contains the truth.

Section 4 therefore sweeps τ and reports set size beside area at every point. On the six categories whose tuned τ leaves the class set no larger, the unified score is 8.4% *worse* in median area than the split ($+7.7\%$ across all seven certifiable categories, the one tighter category being the one whose set grows); tuned to minimise area without that constraint it reports a 16.3% reduction while mean $|S|$ rises from 2.19 to 6.63 of 8 categories. We know of no argument that recovers a guaranteed area improvement here, and the weak measured co-dependence between confidence and localisation error ($\bar{\rho} \in [0.028, 0.064]$) suggests limited scope for improvement.

B.7 Two-Stage Risk Control and the Recall Ceiling

Conformal guarantees hold conditionally on detection, $P(\text{Covered} \mid \text{Detected}) \geq 1 - \alpha$. Missed objects cannot be recovered post hoc, so we report $\text{Cov} \times \text{Rec}$ as an end-to-end coverage proxy rather than calling it $P(\text{Safe})$: it does not account for background false positives or simultaneous scene-level coverage. We test whether calibrating the proposal threshold with conformal risk control can raise this proxy.

Definition 3 (False negative risk)**.** Let $\mathcal{G} = \{g_1, \ldots, g_M\}$ be the ground-truth objects in a scene. For a proposal threshold $\lambda \in [0, 1]$, let $\mathcal{D}(\lambda) = \{d_j : \max_k p_k(d_j) \geq \lambda\}$ be the surviving proposals. An object g_i is detected if $\max_{d_j \in \mathcal{D}(\lambda)} \text{IoU}(d_j, g_i) \geq \text{IoU}_{\text{thresh}}$. The false negative risk is

$$R_{\text{FN}}(\lambda) = \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M} \mathbb{I}(g_i \text{ is not detected by } \mathcal{D}(\lambda))\right] = 1 - \text{Recall}(\lambda). \tag{37}$$

Theorem 9 (Stage 1: risk control on recall)**.** *The loss $R_{\text{FN}}(\lambda)$ is non-decreasing in λ. Let $\hat{R}_n(\lambda)$ be the empirical false negative risk on calibration images and define*

$$\hat{\lambda}_{\text{prop}} = \sup\left\{\lambda \in [0, 1] : \frac{n}{n+1}\hat{R}_n(\lambda) + \frac{1}{n+1} \leq \beta\right\}. \tag{38}$$

Then, by conformal risk control (Angelopoulos et al., 2024), which extends risk-controlling prediction sets (Bates et al., 2021) and Learn-Then-Test (Angelopoulos et al., 2025) to a monotone loss calibrated by this rule,

$$\mathbb{E}[R_{\text{FN}}(\hat{\lambda}_{\text{prop}})] \leq \beta \implies \mathbb{E}[\text{Recall}(\hat{\lambda}_{\text{prop}})] \geq 1 - \beta. \tag{39}$$

Measured outcome and unimplemented Stage 2. With all 100 DETR queries retained down to $\lambda = 0.01$, pedestrian recall saturates at 66.67%; hence the 95% target is outside the detector's proposal support, and no risk-control rule can select a threshold that attains it. Only the 60% recall target is feasible, at a median threshold of 0.20, where the detector produces 28.9 false positives per image at 29.5% precision over the vulnerable-road-user images the loss is averaged on (22.2 per image over all evaluation images, Table 11). A calibrated false-discovery filter would be a natural second stage, but we neither specify nor implement one here. The reported false-positive load therefore sizes an open problem; it is not an FDR guarantee.

C Algorithmic Specifications

Inference is Algorithm 1 in Section 3.6. This section gives the offline calibration pipeline that produces the thresholds it consumes.

Algorithm 2 CertOD offline calibration and hyperparameter tuning

Require: Calibration dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ with $Y_i = (c_i^*, B_i^*)$, pretrained detector f, per-class risk budgets $\{\alpha_c\}_{c\in\mathcal{K}}$, candidate offset grid $\mathcal{G}_{\text{offset}} = \{0, 10, 25, 50, 100, 200, 400, 800\}$ px, and temperature grid $\mathcal{G}_\tau$.

Ensure: Calibrated parameters $\{\tau_c^*\}$, $\{\gamma_c^*\}$, quantiles $\{(\hat{q}_c^{\text{cls}}, \hat{q}_c^{\text{box}})\}$, joint quantiles $\{\hat{q}_c^{\text{joint}}\}$, certifiability mask $M \in \{0,1\}^K$.

Partition $\mathcal{D}$ into a tuning split $\mathcal{D}_{\text{tune}}$ (40%), a calibration split $\mathcal{D}_{\text{cal}}$ (40%), and a held-out evaluation split $\mathcal{D}_{\text{eval}}$ (20%).
Run f on $\mathcal{D}_{\text{tune}}$ and match detections to ground truth to obtain tuples $\{(\hat{B}_i, p_i, c_i^*, B_i^*)\}$.
for each class $c \in \mathcal{K}$ **do**
  $\mathcal{T}_c \leftarrow \{i \in \mathcal{D}_{\text{tune}} : c_i^* = c\}$.
  For each $g \in \mathcal{G}_{\text{offset}}$, calibrate the $(1-\alpha_c/2)$ quantile of $S_{\text{affine}}^{(c)}$ on $\mathcal{T}_c$.
  $\gamma_c^* \leftarrow \arg\min_{g\in\mathcal{G}_{\text{offset}}}$ median certified area on $\mathcal{T}_c$.
  Let r_c^{split} be class c's inclusion rate on $\mathcal{D}_{\text{tune}}$ under the decoupled class threshold.
  For each $\tau \in \mathcal{G}_\tau$, calibrate the joint score on $\mathcal{T}_c$ and record its median area and class-c inclusion rate $r_c(\tau)$.
  $\tau_c^* \leftarrow \arg\min_{\tau\in\mathcal{G}_\tau : r_c(\tau) \le r_c^{\text{split}}}$ median certified area; use the smallest grid value if no candidate is feasible.
end for
Run f on $\mathcal{D}_{\text{cal}}$ and match to obtain calibration tuples $\{(\hat{B}_i, p_i, c_i^*, B_i^*)\}$.
for each class $c \in \mathcal{K}$ **do**
  $\mathcal{C}_c \leftarrow \{i \in \mathcal{D}_{\text{cal}} : c_i^* = c\}$, with $n_c = |\mathcal{C}_c|$.
  if $n_c \ge \lceil 2/\alpha_c \rceil - 1$ **then**
    $M_c \leftarrow 1$ {class is certifiable}
    Modular Boole calibration:
    $\alpha_c^{\text{cls}} \leftarrow \alpha_c/2,\ \alpha_c^{\text{box}} \leftarrow \alpha_c/2.$
    $\hat{q}_c^{\text{cls}} \leftarrow \text{Quantile}\left(\frac{\lceil (n_c+1)(1-\alpha_c^{\text{cls}})\rceil}{n_c};\ \{1-p_{c,i}\}_{i\in\mathcal{C}_c}\right).$
    $S_i^{\text{aff}} \leftarrow S_{\text{affine}}^{(c)}(X_i, B_i^*; \gamma_c^*).$
    $\hat{q}_c^{\text{box}} \leftarrow \text{Quantile}\left(\frac{\lceil (n_c+1)(1-\alpha_c^{\text{box}})\rceil}{n_c};\ \{S_i^{\text{aff}}\}_{i\in\mathcal{C}_c}\right).$
    Unified joint calibration:
    $S_i^{\text{joint}} \leftarrow \max\left(\frac{1-p_{c,i}}{\tau_c^*},\ S_i^{\text{aff}}\right).$
    $\hat{q}_c^{\text{joint}} \leftarrow \text{Quantile}\left(\frac{\lceil (n_c+1)(1-\alpha_c)\rceil}{n_c};\ \{S_i^{\text{joint}}\}_{i\in\mathcal{C}_c}\right).$
  else
    $M_c \leftarrow 0,\ \hat{q}_c^{\text{cls}} \leftarrow +\infty,\ \hat{q}_c^{\text{box}} \leftarrow +\infty,\ \hat{q}_c^{\text{joint}} \leftarrow +\infty$ {declared uncertifiable}
  end if
end for
return $\{\tau_c^*\}, \{\gamma_c^*\}, \{\hat{q}_c^{\text{cls}}, \hat{q}_c^{\text{box}}, \hat{q}_c^{\text{joint}}\}, M.$

D Implementation Details and Extended Empirical Tables

D.1 Model Architecture and Experimental Protocol

We use a frozen pretrained DETR model (`facebook/detr-resnet-50`) with a ResNet-50 backbone, a 6-layer Transformer encoder, a 6-layer Transformer decoder, and 100 object queries (41.3M parameters). The model is evaluated zero-shot on the BDD100K (Yu et al., 2020) validation split (10,000 images at 1280×720). Ground-truth BDD `det_20` annotations (151,221 boxes) are mapped to 8 COCO classes according to Table 8, and evaluation is conducted across 100 independent image-disjoint splits. Comparisons that need no tuning half use $80/20$ calibration/evaluation; comparisons that fit the affine offset or the joint temperature use $40/40/20$ tune/calibration/evaluation, so a class receives 40% of its matched rows for calibration in those runs. Each table states which it uses.

D.2 Full Grid Sensitivity of the Affine Regularizer

Table 9 sweeps $\gamma \in [0, 800]$ px on pedestrians ($\alpha = 0.01$). The measured trade-off is monotone rather than U-shaped. At $\gamma = 0$ (pure relative scoring), median area is $\times 4.44$, tiny-target coverage is 98.94%, and large-target coverage is 98.24%. Increasing γ leaves marginal coverage near 99.04% but steadily enlarges tiny targets and lowers large-target coverage; at $\gamma = 800$ px the corresponding

Table 8: BDD100K to COCO 8-class mapping and the nominal safety budget assigned to each class.

BDD100K label	**Mapped class**	**Nominal safety budget** $(1-\alpha_c)$
pedestrian, rider	person	99.0% ($\alpha = 0.01$)
car	car	95.0% ($\alpha = 0.05$)
bicycle, bike	bicycle	95.0% ($\alpha = 0.05$)
motorcycle, motor	motorcycle	95.0% ($\alpha = 0.05$)
bus	bus	90.0% ($\alpha = 0.10$)
truck	truck	90.0% ($\alpha = 0.10$)
train	train	90.0% ($\alpha = 0.10$)
traffic light	traffic light	80.0% ($\alpha = 0.20$)
traffic sign	*excluded (no COCO counterpart)*	

Table 9: Every offset in the grid run as a fixed score on pedestrians ($\alpha = 0.01$) under the protocol of the named scores. The trade-off is monotone: no interior optimum, and no tiny-object coverage cliff at $\gamma = 0$. Setting $\gamma = 0$ reproduces `rel_max` to floating point, which is how we check the family.

			Q_0 (tiny)		Q_4 (large)	
Offset γ	**Joint cov.**	**Area (med.)**	Cov.	Area	Cov.	Area
0 px	99.06%	×**4.44**	98.94%	×**5.20**	**98.24%**	×4.33
10 px	99.03%	×4.44	98.99%	×6.29	97.63%	×3.79
25 px	99.04%	×4.89	98.99%	×7.97	97.46%	×3.60
50 px	99.04%	×5.71	98.99%	×10.60	97.29%	×3.62
100 px	99.03%	×6.80	98.99%	×14.28	97.14%	×3.68
200 px	99.03%	×7.81	98.99%	×17.95	97.08%	×3.69
400 px	99.04%	×8.74	98.99%	×21.41	97.02%	×3.74
800 px	99.04%	×9.53	98.99%	×24.31	96.97%	×3.81

values are $\times 9.53$, 98.99%, and 96.97%, with the tiny-object area reaching $\times 24.31$. Thus no positive offset improves the pedestrian row. The held-out tuning procedure selects $\gamma^* = 0$ for pedestrians and reserves positive offsets for classes where they reduce measured area, notably buses and trucks.

D.3 Additional Experimental Tables

This section collects the detailed result tables referenced from Section 4: the comparison between unified joint scoring and the Boole budget split (Table 12), the size-stratified breakdown (Table 10), and the confidence-threshold sweep (Table 11). Appendix A.5 reports the detector comparison (Table 6).

D.4 Downstream Geometric Cost

A certified envelope is useful only if something downstream can act on it. The question this data can answer is geometric: how much of the scene does certification consume, and does a per-class output leave a consumer more room than a single box? We do not attempt a planner study. Turning envelopes into braking decisions needs a pixel-to-metric mapping, a cost functional, time dynamics and an evaluation protocol, none of which we have; asserting collision or braking rates without them would not be a measurement.

Three quantities. *Free space* is one minus the fraction of the frame covered by the union of certified envelopes, clipped to the frame. *Envelope merging* is the fraction of detection pairs whose envelopes overlap while their base boxes do not—the gap a planner would have used, closed. *Lateral extent* is the added horizontal reach in units of the object's own ground-truth width.

Measured (Table 14), certification takes free space from 87.5% to 75.2%, and the representation is worth about one further point. The number we would highlight is the merging rate: 15.0% **of detection pairs that the raw boxes keep apart are joined by their certified envelopes**, against 12.9% under argmax conditioning—the concrete sense in which conservative certificates crowd a scene, stated without dynamics. Per focus class (Table 13) the benefit is uneven and not in the

Table 10: Size-stratified audit on pedestrians, using quintiles computed within the pedestrian class, Q_0 (smallest) to Q_4 (largest), nominal 99%. Boundaries are $89.7/688.7/1260.3/2181.4/4498.3/191065.3\,\mathrm{px}^2$; full-population counts are $1670/1669/1669/1669/1670$, and mean evaluation counts per split are $334.2/330.3/331.7/333.3/332.4$. Absolute scores over-inflate small objects; scale-adaptive scores reduce the spread without establishing stratum validity.

Score		Q_0	Q_1	Q_2	Q_3	Q_4
`abs_max`	Cov.	98.99%	99.86%	99.63%	99.79%	**96.91%**
	Area	×30.47	×17.09	×11.33	×7.48	×4.11
`affine_max`	Cov.	98.97%	99.54%	99.27%	99.49%	**97.95%**
	Area	×**5.76**	×4.70	×4.35	×4.11	×4.08
`rel_max`	Cov.	98.94%	99.46%	99.15%	99.49%	**98.24%**
	Area	×5.20	×4.50	×4.29	×4.18	×4.33

Table 11: Threshold sweep on the proposal cache, $\lambda \in [0.01, 0.50]$. Conformal coverage and certified area are flat across the range; the entire cost of expanded recall is false positives, and $\mathrm{Cov} \times \mathrm{Rec}$ saturates at 66.1%, pinned by the ceiling of Appendix A.8.

λ	FP / img	Prec.	Person rec.	Joint cov.	Area (med.)	$\mathrm{Cov}\times\mathrm{Rec}$
0.01	71.60	12.53%	**66.67%**	99.14%	×4.69	**66.09%**
0.05	45.38	17.97%	65.35%	98.87%	×4.59	64.61%
0.10	32.58	22.80%	63.94%	98.93%	×4.56	63.26%
0.15	26.18	26.42%	62.79%	99.14%	×4.60	62.25%
0.20	22.17	29.40%	61.84%	98.98%	×4.48	61.21%
0.30	17.13	34.36%	60.04%	99.01%	×4.44	59.44%
0.50	11.37	42.86%	56.76%	99.07%	×4.44	56.23%

direction we expected: because padding is $t_c \cdot (\mathrm{dimension} + \gamma_c)$ with class-specific offsets, a bus-only consumer recovers 17.7% of its occupancy and a car-only consumer 3.2%. In absolute terms every one of these is small: the largest room gained by any single-class consumer is 1.31 percentage points of the frame, and a pedestrian-only consumer gains 0.11.

D.5 Complete 8×8 Confusion Matrix

Table 15 gives the confusion between true class and detector top-1 over all $89{,}177$ matched detections. Pedestrians are correctly classified 98.73% of the time, and the residual 1.27% goes mostly to `car` (74 instances), `traffic light` (14), `bicycle` (7) and `motorcycle` (7). The row that matters most for Appendix A.9 is `bus`: the detector labels it correctly only 48.85% of the time, sending 31.68% to `truck` and 18.62% to `car`. That is the mechanism behind the heavy-vehicle coverage collapse under argmax conditioning, and it is a property of the detector rather than of the calibration.

D.6 Further CODA2022 Results

Same-detector external baseline. We implement the unscaled two-step method of Ries et al. (2026) on the same detections: regularized adaptive prediction sets (RAPS, Angelopoulos et al., 2021) with $\lambda_{\mathrm{reg}} = 0.01$, $k_{\mathrm{reg}} = 5$ and non-empty sets, class-wise coordinate Bonferroni, and one safe box formed by the coordinatewise maximum over the RAPS set. The safe envelope meets all four aggregate targets. Replacing only that envelope by per-class pairs reduces median area by $15.2/17.2/3.6/4.5\%$ and p90 area by $33.6/41.4/26.4/13.4\%$ for bus, car, person and truck respectively. The original implementation's shortcut—selecting the class whose quantile vector has the largest single coordinate—is tighter but is not a coordinatewise envelope; its bus and truck point estimates are 89.82% and 89.99%. The scaled baseline requires an aleatoric variance head absent from the frozen DETR, so it is a cross-system extension rather than a same-detector comparison.

Recall and open-world boundary. At 0.01 threshold and 100 proposals per image, proposal recall is 77.9% for person, 85.4% for car, 77.7% for bus and 76.9% for truck. Across all 43 CODA classes

Table 12: Unified joint score against the Boole split. The equal-set-size constraint is per class: a category may not appear in more sets than it does under the split arm. Six categories satisfy it, and the unified score is larger in median area on all six. *No certifiable category is tighter at no larger set size.* Traffic light, reported below the rule, is the one category where the unified score is tighter (-15.8%), and it is also the one whose tuned τ leaves a *larger* class set ($1.19 \to 1.27$): the saving is bought with the class term, not with the margin. Including it, the median change across all seven certifiable categories is $+7.7\%$. The unconstrained line is what tuning τ to minimise area alone produces, and the set size beside it is why that number cannot be read as a margin saving. Areas are on the emitted basis of Table 1, so the Split column reproduces that table's CertOD areas; the two arms do not share a class set, and Excl. reports how their emitted populations differ.

	Joint cov.		$\lvert S\rvert$		**Excl.**		**Area (med.)**		**Change**
Category	Split	S_{joint}	Split	S_{joint}	Split	S_{joint}	Split	S_{joint}	
car	95.01	94.99	2.24	2.17	2.50%	3.09%	×3.07	×3.36	+9.3%
bicycle	95.98	95.47	2.44	2.22	2.33%	3.38%	×4.21	×4.60	+9.2%
person	99.04	99.00	1.95	1.84	0.47%	0.70%	×4.45	×4.86	+9.2%
motorcycle	95.38	95.00	3.00	2.69	2.65%	3.28%	×4.02	×4.33	+7.7%
truck	90.39	90.27	2.15	2.15	4.87%	5.67%	×2.67	×2.85	+6.6%
bus	90.54	89.90	2.34	2.31	5.03%	5.74%	×2.38	×2.42	+2.0%
traffic light	89.72	80.30	1.19	1.27	0.20%	0.20%	×2.72	×2.29	−15.8%
Unconstrained tuning: median -16.3% *area, mean* $\lvert S\rvert$ $2.19 \to 6.63$ *of* 8									

Table 13: What a class-aware consumer gains. A consumer interested in class c avoids the c-pair; a single-box consumer must avoid the whole inflated box on the same detections. The gain is the gap between a class's own margin and the set maximum, which under a scale-normalised score depends on the affine offset and the box scale rather than on α_c alone.

Consumer avoids	**In set**	**Occupancy**	**Room gained**
bus	16.6%	−17.7%	+1.31 pp
traffic light	12.0%	−10.5%	+0.04 pp
truck	47.6%	−9.2%	+1.12 pp
person	16.6%	−5.1%	+0.11 pp
motorcycle	26.5%	−4.9%	+0.18 pp
bicycle	13.7%	−3.6%	+0.06 pp
car	75.6%	−3.2%	+0.66 pp

it is 64.6%; for the 36 novel classes it is only 30.7%. No closed-set certificate can repair those missing proposals. Consistent with that ceiling, a 95% image-level recall target is infeasible under conformal risk control.

D.7 Across-Split Variance and Statistical Stability

Table 16 reports across-split dispersion. It is not uniformly small: the safety class is stable ($\sigma = 0.29$ points) but motorcycle and bicycle are not (3.64 and 2.55 points), reflecting evaluation folds of roughly 50 and 96 instances. Two cautions follow. A category whose coverage sits 0.65 points above nominal with $\sigma = 3.64$ has not been shown to meet its target, only to be consistent with it. This table is the earlier $80/20$ coordinate-Bonferroni audit, not the fitted affine-supremum comparison in Table 1. These 100 splits resample one fixed set of objects, so σ measures split-to-split variation and not sampling error over driving scenes; a dependency-aware resampling at the image or sequence level would be needed for the latter.

To estimate uncertainty over evaluation scenes rather than over repeated partitions of one dataset, we fix one image-disjoint $80/20$ calibration/evaluation split on each dataset and resample the evaluation images $2{,}000$ times. Pair and matched-single outputs are resampled together, so Table 17 gives paired intervals for their difference. These intervals condition on the fixed calibration fold: they quantify

Table 14: What each output format costs the frame. Free space is the complement of the union of certified envelopes; merged pairs are detection pairs separated by their base boxes but joined by their envelopes; lateral extent is in units of the object's own ground-truth width. The minimal axis-aligned envelope of the pairs equals the true-quantile single box exactly, as Proposition 7 requires.

Output format	**Free space**	**Merged pairs**	**Lateral**
Raw detector box	87.5%	0.00%	+0.00
Argmax-conditioned CP	76.1%	12.89%	+0.78
Single box (predicted quant.)	75.2%	14.90%	+0.89
Single box (true quant.)	75.2%	15.02%	+0.90
Envelope of the pairs	75.2%	15.02%	+0.90

Table 15: Confusion between true class and detector top-1 over all matched detections of the BDD100K validation split. Cells are row percentages and n is the row total. The detector confuses bus with truck and car heavily, which is the mechanism behind the heavy-vehicle deficit.

True	person	bicycle	m.cycle	car	bus	truck	train	t.light	n
person	**98.73**	0.08	0.08	0.89	0.00	0.05	0.00	0.17	8,347
bicycle	8.70	**81.57**	5.18	3.93	0.00	0.62	0.00	0.00	483
motorcycle	6.83	4.02	**66.67**	21.29	0.00	0.80	0.00	0.40	249
car	0.27	0.01	0.05	**95.33**	0.15	4.10	0.01	0.08	65,526
bus	0.00	0.00	0.00	18.62	**48.85**	31.68	0.60	0.26	1,171
truck	0.25	0.00	0.00	20.86	2.67	**75.75**	0.25	0.21	2,804
train	0.00	0.00	0.00	0.00	0.00	0.00	**100.0**	0.00	2
traffic light	0.09	0.00	0.00	0.08	0.00	0.01	0.00	**99.81**	10,595

evaluation-scene sampling error, not calibration randomness, and are not a test of conformal validity. This distinction matters for the rare BDD classes—the fixed fold contains only 55 motorcycles and 82 bicycles—but the efficiency result is stable. The p90-area reduction has a non-negative lower endpoint for every BDD class and a strictly positive one for all four CODA classes; for person it is 41.5% $[39.4, 44.1]$ on BDD and 17.4% $[11.7, 27.4]$ on CODA.

D.8 Full Multi-Category Distribution Shift

Table 18 gives the full breakdown. Against argmax conditioning the advantage is large on heavy vehicles, but the pair certificates are themselves below nominal there under lighting shift (83.55% on bus and 80.23% on truck against 90%), so the claim we support is graceful degradation rather than retained validity. Under weather shift most categories do hold their level. Two categories are uncertifiable under the occlusion stratification because the unoccluded calibration pool leaves them below the sample-size floor; those cells are marked uncertifiable.

Table 16: Across-split standard deviations over 100 image-disjoint $80/20$ splits for the coordinate-Bonferroni pair certificate. These are split-to-split dispersion, not sampling error: the splits resample the same objects, so a small σ does not license reading a fraction of a point above nominal as a real margin.

Category	Nominal	Joint cov.	σ (pp)	Area (med.) $\pm\sigma$
person	99.0%	99.08%	0.29	$\times 9.30 \pm 0.34$
car	95.0%	95.25%	0.22	$\times 5.68 \pm 0.10$
traffic light	80.0%	90.44%	0.76	$\times 2.91 \pm 0.05$
truck	90.0%	90.86%	1.40	$\times 3.73 \pm 0.20$
bus	90.0%	90.87%	2.09	$\times 3.13 \pm 0.26$
bicycle	95.0%	95.84%	2.55	$\times 6.52 \pm 0.98$
motorcycle	95.0%	95.65%	3.64	$\times 7.92 \pm 1.86$

Table 17: Paired image-cluster bootstrap on one fixed image-disjoint split per dataset (2,000 resamples; percentile 95% intervals). $e/n/m$ is the number of true-class-emitted rows/matched rows/images in the evaluation fold. Δ joint is pair minus matched-single joint coverage in percentage points; area columns are the percentage reduction from matched single to pair on the e emitted rows. This audit uses the same absolute coordinate-Bonferroni construction as Table 16, independently of the affine-supremum headline comparison. The intervals condition on the calibration fold and therefore do not replace the conformal guarantee.

Dataset	Class	$e/n/m$	Pair joint [CI]	Δ joint [CI]	Median reduction [CI]	p90 reduction [CI]
BDD100K	bicycle	82/82/60	100.00 [100.00, 100.00]	0.00 [0.00, 0.00]	7.4 [0.0, 11.5]	16.9 [0.0, 23.8]
	bus	227/242/199	88.02 [83.72, 92.10]	-2.07 [-3.97, -0.43]	22.1 [12.5, 26.3]	33.8 [23.3, 39.4]
	car	12844/13168/1929	95.40 [95.00, 95.77]	-0.36 [-0.47, -0.27]	15.5 [14.4, 16.3]	29.3 [27.9, 30.4]
	motorcycle	53/55/45	83.64 [72.34, 93.33]	-7.27 [-14.06,-1.64]	49.1 [28.9, 51.8]	50.9 [38.4, 61.7]
	person	1674/1684/542	99.05 [98.52, 99.52]	-0.06 [-0.19, 0.00]	19.8 [13.7, 25.1]	41.5 [39.4, 44.1]
	traffic light	2017/2024/811	90.37 [89.02, 91.70]	-0.59 [-0.94, -0.28]	3.1 [2.1, 5.4]	19.8 [14.4, 26.2]
	truck	549/580/442	90.17 [87.79, 92.40]	0.00 [0.00, 0.00]	0.0 [0.0, 0.6]	9.7 [0.3, 12.9]
CODA2022	bus	187/200/167	89.00 [83.96, 93.23]	-0.50 [-1.66, 0.00]	12.3 [1.4, 20.9]	29.9 [20.5, 33.0]
	car	4768/4885/1524	95.31 [94.68, 95.94]	-0.39 [-0.57, -0.23]	15.9 [13.0, 18.3]	37.0 [34.2, 39.9]
	person	1373/1380/580	98.91 [98.33, 99.44]	0.00 [0.00, 0.00]	2.3 [0.8, 4.5]	17.4 [11.7, 27.4]
	truck	624/653/486	90.35 [87.89, 92.72]	-0.31 [-0.79, 0.00]	4.2 [1.6, 10.5]	16.0 [9.5, 22.0]

Table 18: Distribution shift, all categories. *Cert.* is the fraction of splits on which the category is certifiable at all; where it is zero, a joint coverage of 0.00% is a calibration-scarcity failure and not a coverage measurement. Day-to-night and weather pools share no image; the occlusion pools share $8{,}475$ images, because occlusion is an object attribute, so that block is a stratification rather than a domain shift. Areas are the earlier coordinate-Bonferroni certificate evaluated on all matched rows, not the emitted affine-supremum basis of Table 1; this table is an argmax-versus-pair shift audit and is not compared against a headline number.

Shift	Category	Nom.	Cert.	Argmax	Pair	Area argmax	Area pair
Day → night	person	99.0%	1.00	96.57%	**98.65%**	×5.54	×4.22
	car	95.0%	1.00	81.13%	89.78%	×2.30	×2.50
	bicycle	95.0%	1.00	61.56%	90.77%	×3.50	×3.93
	motorcycle	95.0%	1.00	59.66%	**96.50%**	×3.46	×4.22
	bus	90.0%	1.00	37.00%	83.55%	×1.74	×1.95
	truck	90.0%	1.00	50.48%	80.23%	×1.74	×2.37
	traffic light	80.0%	1.00	76.67%	**85.71%**	×2.45	×2.76
Clear → adverse	person	99.0%	1.00	98.03%	**99.36%**	×8.81	×4.87
	car	95.0%	1.00	92.01%	**95.44%**	×4.00	×3.26
	bicycle	95.0%	1.00	84.24%	**97.51%**	×7.17	×6.42
	motorcycle	95.0%	0.74	66.78%	72.90%	×5.93	×4.53
	bus	90.0%	1.00	40.71%	**91.95%**	×2.57	×2.79
	truck	90.0%	1.00	68.15%	**91.64%**	×2.44	×2.78
	traffic light	80.0%	1.00	81.63%	**91.17%**	×2.41	×2.79
Unocc. → occ.	person	99.0%	1.00	97.44%	98.27%	×8.48	×4.51
	car	95.0%	1.00	93.75%	**96.27%**	×5.07	×3.27
	bicycle	95.0%	**0.00**	76.03%	—	×5.75	—
	motorcycle	95.0%	**0.01**	61.51%	—	×4.58	—
	bus	90.0%	1.00	38.90%	86.75%	×2.80	×2.44
	truck	90.0%	1.00	64.57%	**89.86%**	×2.67	×2.68
	traffic light	80.0%	1.00	72.83%	**84.51%**	×2.16	×2.53