When a Closed-Vocabulary Detector Responds with "Car": Diagnosing Autorickshaw–Car Overlap under Cityscapes-to-IDD Adaptation

Gunita Singh Khurana[1], Avnish Panwar[2], and Indrajit Ghosh[2]

[1] Thapar Institute of Engineering and Technology, Patiala, Punjab, India
gunitasinghkhurana@gmail.com
[2] Indian Institute of Technology Roorkee, Roorkee, Uttarakhand, India
avnish_p@mfs.iitr.ac.in
indrafce@iitr.ac.in

Abstract. A road detector with a closed vocabulary cannot explicitly name an object outside its source label space. We study how such detectors handle one safety-relevant target-private category under Cityscapes-to-IDD transfer: does an autorickshaw GT receive no qualifying known-class prediction, or does its region overlap a source-known prediction such as *car*? We use complementary GT-anchored and fixed-region diagnostics on an image-disjoint, autorickshaw-focused 1000-image IDD subset containing 2196 autorickshaw GT boxes. Across source-only Faster R-CNN, Standard DA-Faster R-CNN, and two RoI-gated stress-test runs, diagnosis-subset known-class AP, response coverage, misses, and absolute as-car counts produce different point-estimate rankings. The conditional class-overlap pattern is nevertheless stable: 97.08–98.00% of autorickshaw GTs with any qualifying known-class match also have a car match. Standard DA has fewer as-car matches than source-only (932 versus 977) but more misses (1236 versus 1191), whereas RoI-gated DA has the highest diagnosis-subset AP (15.32), fewer misses, and the largest as-car count (1028). A separately trained supervised autorickshaw localization probe provides a common-region check on frozen student outputs. The main insight is therefore coverage-confounded car overlap: absolute target-private overlap counts must be interpreted jointly with detector response and silence, not as evidence that one adaptation method is universally better.

Keywords: Domain adaptation · Open-set detection · Road perception

1 Introduction

Autonomous-driving perception is commonly evaluated on structured urban datasets, yet deployment scenes contain locally common objects that may be absent from the source vocabulary. Indian road scenes, for example, include autorickshaws, handcarts, animals, and roadside encroachments that do not map

cleanly to the categories of standard European road-scene datasets. Cityscapes-to-IDD transfer is therefore not only a domain-shift problem; it is also a *target-private object* problem.

Cityscapes defines the detection classes {person, rider, car, truck, bus, train, motorcycle, bicycle}, but not *autorickshaw*. A closed-vocabulary Cityscapes detector evaluated on IDD cannot output "autorickshaw" or an explicit "unknown" label. We describe its observable response to this missing category as **target-private handling**. Under the stated score and IoU criteria, an autorickshaw GT may remain *silent*, with no qualifying known-class prediction, or its region may overlap one or more source-known predictions. We call the latter a *known-class overlap response*. When the overlapping class is *car*, we use *car absorption* only as shorthand for the operational GT–prediction overlap event defined in Sec. 3; it is not a claim of unique semantic assignment.

This distinction matters because silence and an overlapping known-class prediction communicate different states to the rest of a perception stack. In the former, the detector provides no qualifying known-class output. In the latter, a familiar source-domain label can cover the target-private GT region and conceal that the category is outside the trained vocabulary. We treat this as a perception-trustworthiness concern, not direct evidence of collision risk. Its downstream effect remains stack-dependent: geometry and tracking may still support obstacle avoidance, whereas class-conditioned motion priors, shape assumptions, interaction models, or rule-based logic may use the car label differently.

DA-Faster R-CNN [1] aligns source and target features, but is designed primarily for a shared label space. With target-private categories, adaptation can alter whether an object receives a known prediction, remains unmatched, or crosses a high-confidence threshold. Open-world road methods such as ORDER [10] instead modify detectors to identify and learn unknown objects. Our goal is complementary: we probe frozen detectors with neither an autorickshaw nor an explicit unknown output.

Our analysis uses two complementary views. *GT-anchored diagnosis* measures final post-NMS behaviour at each true autorickshaw location. *Fixed-probe-region diagnosis* compares frozen students on the same likely autorickshaw regions recovered by a supervised post-hoc localization probe. The views use different anchors, answer related but non-identical questions, and are not numerically interchangeable.

The central observation is that diagnosis-subset AP and target-private response metrics form separate evaluation axes. Standard DA has fewer car overlaps than source-only but more misses, while a confidence-weighted RoI stress test has the highest diagnosis-subset AP and response coverage but more car matches. Because car accounts for 97.08–98.00% of known-matched GTs in every run, the strongest result is not a changed conditional semantic tendency; it is that absolute as-car counts co-vary with whether the detector responds at all. All comparisons are single-run descriptive observations, not causal or seed-robust method rankings.

1.1 Scope and Contributions

This is a focused failure-analysis and diagnostic study. It does not propose a complete open-set detector, a deployment-ready autorickshaw detector, or evidence of increased collision risk.

(1) Target-private car-overlap response. We define autorickshaw–car overlap as a class-specific existential GT–prediction event under Cityscapes-to-IDD transfer, using "car absorption" only as operational shorthand.

(2) Joint response–silence evaluation. We report known-matched coverage, miss rate, absolute as-car counts, the conditional fraction of known-matched autorickshaws also matched by car, and high-confidence car overlap. This prevents a reduction in the as-car count from being mistaken for improvement when it coincides with increased detector silence.

(3) Complementary post-hoc diagnostics. We combine GT-anchored final-output analysis with a fixed-region view supplied by a localization probe. The two views use different anchors but yield the same descriptive point-estimate ordering of high-confidence car overlap across the evaluated students.

(4) Metric disagreement under adaptation. Across the evaluated single-run students, known-class AP, target-private coverage, and car-overlap counts rank models differently. A RoI-gated variant is used as a diagnostic stress test that exposes this disagreement, not as a proposed mitigation.

2 Related Work

Road-scene detection and domain adaptation. Faster R-CNN [8] is a natural base detector for this study because its proposal and RoI-classification stages make it possible to separate localization from semantic assignment. Cityscapes [2] represents structured European urban scenes, while IDD [11] captures less structured Indian driving conditions. DA-Faster R-CNN [1] aligns source and target features with image- and instance-level adversarial objectives, but its standard formulation assumes a shared source-target label space.

Open-set adaptation and open-world detection. Open-set adaptation asks what happens when target samples do not belong to source-known categories. Saito *et al.* [9] emphasize that unknown target samples should not simply be aligned with source-known samples. In object detection, Dhamija *et al.* [3] formalize an open-set detection protocol and Wilderness Impact, while ORE [5] evaluates unknown-to-known errors and incrementally learns new classes. Unknown Sniffer [6] targets unknowns missed by standard proposals, and UniDetector [12] broadens detection across heterogeneous label spaces. ORDER [10] adapts the open-world perspective to road scenes. These methods modify detectors to identify or learn unknown objects; we instead diagnose final outputs of a detector with neither an unknown nor an autorickshaw class.

Uncertainty and diagnostic probes. Softmax confidence may remain high on out-of-distribution inputs, and energy-based scoring has been used for unknownness estimation [7]. Probing studies commonly keep evaluated models frozen and use controlled readouts to test a defined capability [4]. Our use of *probe* is narrower and behavioural: we do not train a readout on student features or claim to recover information from latent representations. Because exploratory objectness-, entropy-, and energy-style cues did not provide clean object-level localization, we separately train a supervised single-class autorickshaw detector to supply common regions for post-hoc comparison. It is independent only of the students—not of target-private supervision—and is neither part of adaptation nor an unknown detector.

3 Problem Formulation and Metrics

3.1 Target-Private Handling

Let the source domain be Cityscapes and the target domain be IDD. The student detector predicts

$$\mathcal{Y}_s = \{\text{person}, \text{rider}, \text{car}, \text{truck}, \text{bus}, \text{train}, \text{motorcycle}, \text{bicycle}\}. \tag{1}$$

Autorickshaw is a target-private category, denoted by $u \notin \mathcal{Y}_s$. Because the evaluated student has neither an autorickshaw output nor an explicit unknown output, its final observable handling of an autorickshaw is either *silence*—no qualifying known prediction—or a response expressed through one or more source-known classes. We call this a *known-class overlap response*. For continuity with open-set error terminology, we use *absorption* as a compact operational name for the existential overlap event defined below, not as proof of a unique object-level interpretation.

3.2 GT-Anchored Absorption

Let u_i be an autorickshaw ground-truth box and $p_j = (b_j, c_j, s_j)$ be a post-NMS student prediction with box b_j, class $c_j \in \mathcal{Y}_s$, and score s_j. We define class-specific absorption as

$$\text{Absorb}(u_i, c) = \mathbf{1}[\exists p_j : c_j = c,\ \text{IoU}(b_j, u_i) \geq \tau,\ s_j \geq \gamma]\,. \tag{2}$$

In the main GT-level analysis, $\tau = 0.5$ is the IoU threshold and $\gamma = 0.5$ is the score threshold. Because autorickshaw is absent from the student's output vocabulary, we use "absorption" as shorthand for a source-known prediction overlapping the target-private GT rather than ordinary within-vocabulary misclassification or a unique semantic assignment. Autorickshaw-to-car absorption is the special case $c = \text{car}$, with count

$$N_{\text{as-car}} = \sum_i \text{Absorb}(u_i, \text{car}). \tag{3}$$

These are existential, GT-centric overlap counts. Each autorickshaw GT contributes at most once to a given class-specific metric, but the counts are non-exclusive: the same GT may overlap qualifying predictions from multiple known classes. Accordingly, *car absorption* means that an autorickshaw GT box has at least one qualifying overlapping car prediction; it does not establish a unique or highest-scoring object-level semantic assignment. We use "car absorption" as shorthand for this operational overlap event throughout the paper.

3.3 Known Match and Miss Rate

An autorickshaw is known-matched when any known-class prediction passes the same score and IoU thresholds:

$$\text{KnownMatch}(u_i) = \mathbf{1}[\exists p_j : c_j \in \mathcal{Y}_s,\ \text{IoU}(b_j, u_i) \geq \tau,\ s_j \geq \gamma]\,. \tag{4}$$

We report

$$\text{KnownPredRate} = \frac{1}{N_u} \sum_i \text{KnownMatch}(u_i), \tag{5}$$

and

$$\text{MissRate} = 1 - \text{KnownPredRate}\,. \tag{6}$$

3.4 Complementary Diagnostic Views

Let $\mathcal{T}$ be the fixed set of probe-localized autorickshaw regions. GT-anchored diagnosis measures final detector behaviour at the true object location. Probe-anchored diagnosis asks a complementary question: whether a high-confidence student car prediction overlaps the same independently localized region across all frozen students. A probe-localized region is counted as a high-confidence car-overlap response when a post-NMS student car prediction overlaps it at IoU at least 0.5 with score at least 0.5.

The two views are distinguished by their anchors: known-matched, as-car, and miss rate use autorickshaw GT boxes, whereas fixed-region analysis uses probe boxes. Because these anchors differ geometrically, their counts answer different questions and are not directly interchangeable.

4 Datasets and Experimental Setup

The held-out diagnosis subset contains 1000 IDD images, each with at least one autorickshaw annotation, and 2196 autorickshaw GT boxes in total. The autorickshaw-only GT file was generated by filtering the unknown-object annotations associated with the 1000-image known-class diagnosis file; all 2196 autorickshaw annotations map to images in that diagnosis file. The 21,638-image

Table 1: Diagnostic metric dictionary. All student predictions are post-NMS. GT-level quantities are binary per autorickshaw GT; the probe-box quantity is binary per recovered probe region.

Quantity	Anchor	Score rule	IoU rule	Denominator
Known match	Autorickshaw GT	Any known $s \geq 0.5$	≥ 0.5	2196 GT
As-car	Autorickshaw GT	Car $s \geq 0.5$	≥ 0.5	2196 GT
Miss	Autorickshaw GT	No qualifying known prediction	–	2196 GT
Probe recovery	Autorickshaw GT	Probe $s \geq 0.5$	≥ 0.5	2196 GT
High-conf. probe-box car	Probe box	Car $s \geq 0.5$	≥ 0.5	1680 regions

Table 2: Data usage and split roles. The diagnosis set is confirmed image-disjoint from target adaptation and supervised probe training. IDD labels are never used in student optimization.

Partition	Images	Boxes	Student-label use	Purpose
Cityscapes train	2975	52,467	Yes	Supervised source training
Cityscapes val	500	10,180	Evaluation only	Source sanity check
IDD target adaptation	21,638	–	No	Unlabeled adaptation
Probe train	23,030	25,900	No	Supervised probe training
Held-out diagnosis	1000	2196	No	Final overlap diagnosis

unlabeled target-adaptation split was generated by removing every diagnosis-set filename from the broader IDD known-class COCO file. A separate image-ID/filename audit confirms zero overlap between the diagnosis set and both target adaptation and the 23,030-image supervised probe-training split. The subset is autorickshaw-focused and is not claimed to reproduce the full IDD validation distribution. All retained annotations for the eight Cityscapes-aligned classes are used for diagnosis-subset AP, whereas autorickshaw is excluded from mAP and evaluated through response, miss, confidence, and overlap metrics. For this study, the diagnosis subset was inherited from the predefined 1000-image annotation JSON used by the earlier held-out experiment; it was not resampled during the present analysis. Sequence-level and near-duplicate-frame separation were not audited.

All students use Faster R-CNN with a ResNet-50-FPN backbone in Detectron2 0.6, initialized from a COCO-pretrained checkpoint. Each model uses a fixed 3000-iteration budget, batch size 1, input resize min 512/max 1024, base learning rate 2.5×10^{-4}, and DA weight $\lambda = 0.1$. A source-only iteration consumes one labeled Cityscapes image; a DA iteration consumes one labeled source image and one unlabeled IDD image, with target labels removed. Source labels supervise detection, while both domains contribute domain-adaptation losses. Nominal image draws correspond to $3000/2975 \approx 1.01$ source-split passes and $3000/21638 \approx 0.14$ target-split passes. The final `model_final.pth` checkpoint was fixed by this budget, not selected using diagnosis metrics; no learning-rate schedule or checkpoint was tuned using the held-out diagnosis metrics. Training completed without numerical instability, but no checkpoint sweep, longer-schedule study, or multi-seed convergence analysis was performed. We

therefore interpret all comparisons as single-run, fixed-budget diagnostics rather than convergence-validated rankings. For diagnostic exports, retained post-NMS probe and student predictions use an explicitly set Detectron2 RoI score threshold of $s \geq 0.05$; the same export floor is used across the probe and all evaluated students. Other inference settings follow the resolved model configuration and are not modified by the diagnostic scripts.

5 Models and Diagnostic Framework

5.1 Students: Source-Only, Standard DA, and RoI-Gated DA

The **source-only** student is trained on Cityscapes and evaluated directly on IDD. The **Standard DA** student follows DA-Faster R-CNN [1], using source detection loss plus image- and instance-level adversarial alignment:

$$\mathcal{L} = \mathcal{L}_{det}^{src} + \lambda \left(\mathcal{L}_{img}^{src/tgt} + \mathcal{L}_{ins}^{src/tgt} + \mathcal{L}_{cons} \right). \tag{7}$$

The **RoI-gated DA** student is included as a foreground-confidence-gated alignment stress test, not as a proposed open-set solution. Standard instance-level DA aligns target RoIs without explicitly separating source-known-like regions from possible target-private regions. A natural diagnostic modification is therefore to weight the target instance-alignment loss by the student's maximum known-class foreground confidence. This diagnostic variant relatively emphasizes target RoIs that appear known-like under the student while downweighting lower-confidence RoIs. It is used to examine whether such a confidence-weighted adaptation objective can produce a different AP–car-overlap operating point.

For each target RoI, let $p_i = \mathrm{softmax}(z_i)$ and $m_i = \max_{c \in \mathcal{Y}_s} p_i(c)$, excluding background. The target instance-domain loss is weighted by

$$w_i = \max(\alpha_{\min}, m_i) \tag{8}$$

and

$$\mathcal{L}_{ins}^{tgt} = \frac{1}{N} \sum_{i=1}^{N} w_i \, \mathrm{BCEWithLogits}(d_i, y_i). \tag{9}$$

We evaluate $\alpha_{\min} = 0$ and $\alpha_{\min} = 0.1$. Because the weighted loss is averaged over N, rather than normalized by $\sum_i w_i$, the variant changes both the relative emphasis among target RoIs and the effective magnitude of target instance alignment. It is therefore a composite confidence-weighted stress test, not a loss-scale-matched ablation that isolates RoI selectivity or a proposed mitigation.

5.2 Supervised Post-hoc Localization Probe

We separately train a supervised single-class Faster R-CNN autorickshaw localization probe using labeled IDD autorickshaw data. It is independent of the evaluated students, not of target-private supervision: training occurs only after

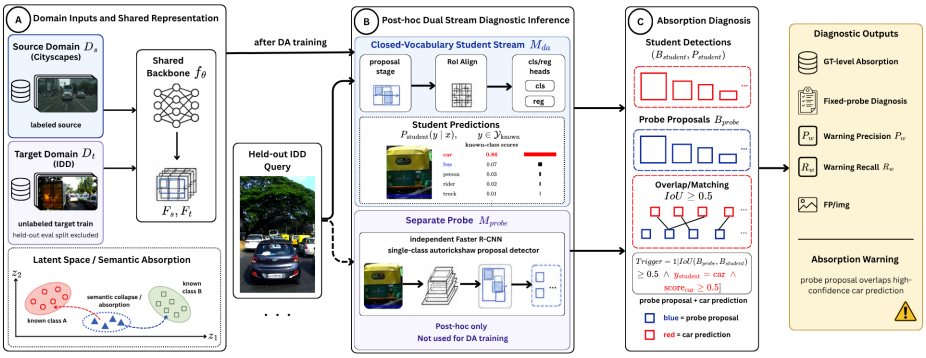

Fig. 1: Complementary post-hoc diagnostic views. (A) Student training uses labeled Cityscapes and, for DA students, unlabeled IDD; the supervised localization probe is absent from optimization. (B) A held-out image is processed separately by a frozen student and the post-hoc probe. (C) GT and probe-box matching yield the response–silence and common-region high-confidence car analyses. All reported quantities use final boxes and stated overlap thresholds rather than latent-space readouts.

student configurations are fixed, and the probe neither reads student features nor contributes to student optimization, adaptation, or checkpoint selection. Its training set is disjoint from the diagnosis subset, and its role is limited to providing common likely-autorickshaw regions for post-hoc comparison of frozen outputs.

The probe is a Faster R-CNN R50-FPN initialized from the source-only weights and trained on 23,030 images containing 25,900 autorickshaw boxes. At probe score and probe–GT IoU thresholds of 0.5, it recovers 1680 of 2196 held-out GT instances (76.50% recall). Probe recovery is GT-anchored and existential: for each autorickshaw GT box, the evaluation searches all retained probe predictions in the same image and records the highest-scoring prediction with IoU at least 0.5. The GT is counted as recovered when this score is at least 0.5. No global one-to-one matching between probe boxes and GT boxes is enforced. Fixed-region conclusions are therefore conditional on this recovered subset and matching rule.

Fig. 1 summarizes the training separation and the two post-hoc diagnostic anchors. GT-level diagnosis asks whether a qualifying final prediction overlaps the true autorickshaw box. Fixed-probe diagnosis asks whether a high-confidence student car prediction overlaps the same supervised-probe region across frozen students.

Table 3: Score-threshold sensitivity of GT-anchored target-private handling on the held-out autorickshaw diagnosis subset. The student prediction threshold γ is varied while GT–prediction IoU is fixed at 0.5. "Car among matched" is $N_{\text{as-car}}/N_{\text{known}}$. Absolute as-car counts must be interpreted jointly with response coverage and detector silence.

Model	γ	Known-matched	Known rate	Miss rate	As-car	Car among matched
Source-only	0.3	1245	56.69%	43.31%	1185	95.18%
Source-only	0.5	1005	45.77%	54.23%	977	97.21%
Standard DA	0.3	1202	54.74%	45.26%	1154	96.01%
Standard DA	0.5	960	43.72%	56.28%	932	97.08%
RoI-gated, no-min	0.3	1215	55.33%	44.67%	1169	96.21%
RoI-gated, no-min	0.5	1007	45.86%	54.14%	982	97.52%
RoI-gated DA	0.3	1283	58.42%	41.58%	1233	96.10%
RoI-gated DA	0.5	1049	47.77%	52.23%	1028	98.00%

6 Experiments and Results

6.1 RQ1: How Does a Source-Trained Detector Handle Autorickshaws?

The source-only detector provides a Cityscapes validation sanity check of AP 21.10, AP50 40.63, and AP75 18.72. On the held-out autorickshaw diagnosis subset, its known-class AP is 15.15. More importantly for target-private handling, 1005 of 2196 autorickshaw GT boxes receive a qualifying known-class match, while 1191 receive none. Of the matched GTs, 977 are also matched by car. The baseline behaviour is therefore split between detector silence and a dominant car-overlap response; neither outcome is visible in known-class AP alone.

6.2 RQ2: Does Adaptation Change the Balance between Response and Silence?

Tab. 3 evaluates the response–silence comparison at student score thresholds $\gamma \in \{0.3, 0.5\}$ while keeping GT–prediction IoU fixed at 0.5. At the primary threshold of 0.5, Standard DA has a lower as-car point estimate than source-only, 932 versus 977, but also more misses, 1236 versus 1191. A lower as-car count must therefore not be read alone; it may reflect a shift from a qualifying car-overlap response to no qualifying known prediction.

Score-threshold sensitivity. Lowering γ from 0.5 to 0.3 increases known-match coverage and decreases misses for every run, while also increasing the number of autorickshaw GT boxes with a qualifying overlapping car prediction. The conditional response remains strongly car-dominated: car accounts for 95.18–96.21% of known-matched GTs at $\gamma = 0.3$, compared with 97.08–98.00% at $\gamma = 0.5$. Standard DA continues to have lower response coverage and fewer absolute as-car matches than RoI-gated DA at both thresholds. Thus, the response–silence

Table 4: Known-class performance on the autorickshaw-focused held-out diagnosis subset. Autorickshaw is excluded from mAP. These values are conditional diagnostic quantities and are not full-IDD benchmark results. Higher is better.

Model	AP ↑	AP50 ↑	AP75 ↑
Source-only	15.15	**28.08**	**14.60**
Standard DA	14.74	27.19	13.91
RoI-gated, no-min	14.74	27.67	13.79
RoI-gated DA	**15.32**	28.06	14.23

Table 5: Known-class AP and GT-anchored autorickshaw–car overlap counts. Both overlap columns use autorickshaw GT boxes as anchors. The high-confidence column is the subset of all-GT as-car matches occurring among the 1680 GT instances recovered by the probe. Bold is used only for AP extrema; overlap counts are presented neutrally.

Model	AP ↑	Car AP ↑	As-car count, all GT	High-conf. as-car, recovered GT (count/rate)
Source-only	15.15	35.33	977	959 (57.08%)
Standard DA	14.74	35.35	932	916 (54.52%)
RoI-gated, no-min	14.74	35.19	982	964 (57.38%)
RoI-gated DA	**15.32**	**35.66**	1028	1002 (59.64%)

conclusion is not specific to the single 0.5 operating point: absolute as-car counts continue to co-vary with whether a detector produces a qualifying known-class response.

6.3 RQ3: Do Conventional AP and Target-Private Handling Rank Models Consistently?

Tab. 4 gives the conventional view on the held-out autorickshaw diagnosis subset. RoI-gated DA has the highest overall AP point estimate, 15.32, while source-only retains the highest AP50 and AP75. Tab. 5 places this beside GT-anchored car-overlap counts. The resulting rankings disagree: the run with the highest AP is not the run with the lowest as-car count or lowest high-confidence car rate.

Across models, 97.08–98.00% of autorickshaw GT boxes matched by any known-class prediction are also matched by car. Thus, the conditional class-overlap pattern is stable and strongly car-dominated, while the absolute number of as-car cases changes with known-match coverage and confidence. The fixed-region analyses in RQ4 provide a complementary check on whether the high-confidence ordering persists when the comparison is restricted to common recovered instances.

Image-bootstrap uncertainty. Tab. 6 reports $B = 10{,}000$ percentile-bootstrap intervals obtained by resampling held-out images with replacement using seed 42, rather than resampling individual GT boxes. The intervals overlap substantially, so they are not used to claim statistically distinct model behaviour. Instead, they show that the reported ordering is a descriptive point-estimate ordering

Table 6: Image-bootstrap uncertainty for GT-anchored target-private handling at the primary student score threshold $\gamma = 0.5$ on the held-out 1000-image diagnosis subset. We use $B = 10{,}000$ image-level bootstrap resamples with replacement and seed 42, carrying all GT boxes and predictions from each sampled image together. Values are point estimates with percentile 95% intervals from the 2.5th and 97.5th percentiles. The intervals quantify evaluation-image sampling variability only; they do not represent training-seed, checkpoint, split-construction, or adversarial-optimization uncertainty.

Model	Known-matched count	Known rate	As-car count
Source-only	1005 [945, 1069]	45.77% [43.64%, 47.96%]	977 [917, 1040]
Standard DA	960 [899, 1023]	43.72% [41.61%, 45.91%]	932 [874, 994]
RoI-gated, no-min	1007 [946, 1072]	45.86% [43.75%, 48.01%]	982 [922, 1047]
RoI-gated DA	1049 [986, 1115]	47.77% [45.67%, 49.95%]	1028 [965, 1094]

for the fixed trained checkpoints. Known-match coverage and absolute as-car count also move together across the evaluated runs, consistent with the coverage-confounding interpretation. These intervals do not account for training-seed, checkpoint, split-construction, or adversarial-optimization variability.

All model comparisons in this section are therefore descriptive rather than claims of seed-robust superiority. The central evidence is the divergence among AP, miss rate, and car-overlap counts under one controlled protocol, not a universal ranking of adaptation methods.

6.4 RQ4: Does the High-Confidence Ordering Persist across Complementary Anchors?

GT-level final-output counts combine whether a qualifying prediction reaches the autorickshaw and which source-known predictions overlap it after post-processing. The localization probe provides a complementary common-region view without changing any student. It recovers 1680 of 2196 held-out autorickshaw GT boxes at threshold 0.5, and every frozen student is evaluated on this same recovered subset.

Tab. 7 first uses GT boxes as anchors. The real-car control comprises all 3756 car GT boxes in the same 1000-image diagnosis subset; the autorickshaw rows use the 1680 probe-recovered autorickshaw GT boxes. We report the fraction with at least one post-NMS car prediction of score ≥ 0.5 and IoU ≥ 0.5. Under RoI-gated DA, 60.01% of real-car GT boxes and 59.64% of probe-recovered autorickshaw GT boxes meet this criterion. This is a confidence control on GT anchors, not evidence that the prediction geometry is caused exclusively by the annotated object.

Tab. 8 then switches to probe-box anchors. A recovered probe region is assigned to "high-conf. car" when any post-NMS student car prediction has score ≥ 0.5 and probe–student IoU ≥ 0.5; otherwise it is assigned to "no high-conf. car." This binary decomposition uses the verified high-confidence threshold of

Table 7: GT-anchored high-confidence car control. "High-conf. car" denotes at least one post-NMS car prediction with score ≥ 0.5 and IoU ≥ 0.5 to the GT box. Real-car rows contain all 3756 car GT boxes in the same diagnosis subset; autorickshaw rows are restricted to the 1680 probe-recovered GT boxes.

Model	GT region	N	High-conf. car
Source-only	Real car	3756	57.37%
Source-only	Autorickshaw	1680	57.08%
Standard DA	Real car	3756	58.36%
Standard DA	Autorickshaw	1680	54.52%
RoI-gated, no-min	Real car	3756	58.41%
RoI-gated, no-min	Autorickshaw	1680	57.38%
RoI-gated DA	Real car	3756	60.01%
RoI-gated DA	Autorickshaw	1680	59.64%

Table 8: Probe-box-anchored high-confidence decomposition of the same 1680 recovered regions. The two categories are mutually exclusive and exhaustive. "No high-conf. car" includes every region without a qualifying car prediction at score and IoU thresholds of 0.5.

Model	High-conf. car	No high-conf. car
Source-only	904 (53.81%)	776 (46.19%)
Standard DA	870 (51.79%)	810 (48.21%)
RoI-gated, no-min	930 (55.36%)	750 (44.64%)
RoI-gated DA	967 (57.56%)	713 (42.44%)

0.5; the underlying retained predictions were exported with the common score floor $s \geq 0.05$.

The high-confidence counts in Tab. 7 and Tab. 8 differ because they use different anchors. The first matches student car predictions to autorickshaw GT boxes; the second matches them to supervised-probe proposal boxes. Probe boxes are not identical to GT boxes, so a prediction can meet IoU ≥ 0.5 with one anchor but not the other.

Despite this geometric difference, both anchors yield the same descriptive point-estimate ordering on the recovered subset: Standard DA is lowest, followed by source-only, the no-minimum gated run, and RoI-gated DA. This agreement is conditional on the 1680 recovered instances; it is neither a formal significance result nor a measurement of internal representations.

Fig. 2 shows two selected cases in which the Standard-DA run has no qualifying overlapping car prediction while one or both gated runs do. These examples illustrate individual output differences at the fixed score/IoU operating point and are not used to estimate prevalence or causal effects.

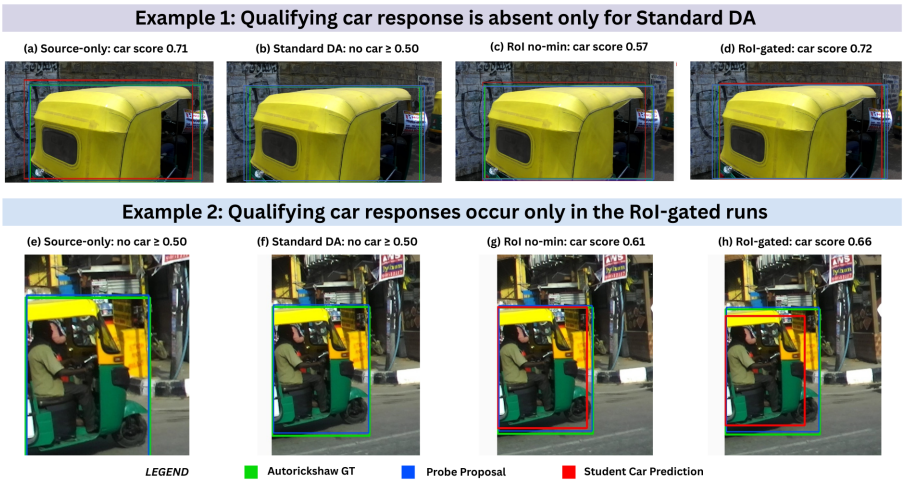

Fig. 2: Selected instance-level outputs from the four evaluated fixed-budget runs. Green denotes autorickshaw GT, blue the supervised localization-probe proposal, and red a student car prediction. In these examples, the Standard-DA run has no qualifying overlapping car prediction, whereas one or both gated runs do. The examples illustrate individual output differences under the fixed score/IoU operating point and are not evidence of causal method effects or statistical prevalence.

7 Discussion

Target-private handling is a joint pattern, not a scalar score. Reporting the as-car count alone can reward detector silence, while known-class AP omits autorickshaw entirely. Since 97.08–98.00% of known-matched GTs also have a car match in every run, absolute as-car counts are strongly coverage-confounded. Standard DA has the lowest as-car point estimate but the highest miss rate; RoI-gated DA has the highest diagnosis-subset AP and response coverage but the highest as-car counts. These single-run observations establish response–silence–overlap disagreement, not a causal trade-off or seed-robust ranking.

Complementary anchors provide conditional convergent evidence. GT anchors evaluate final outputs at true object locations, whereas probe boxes compare retained predictions on common independently localized regions. Their counts are not interchangeable because the boxes differ geometrically, yet both produce the same high-confidence point-estimate ordering. The probe therefore strengthens the behavioural diagnosis on the 1680 recovered instances, but it controls regions rather than reading out internal student representations.

Safety interpretation remains stack-dependent. A car prediction overlapping an autorickshaw GT may still support generic obstacle avoidance when planning relies mainly on geometry and tracking. Conversely, class-conditioned motion priors, shape assumptions, interaction models, and rule-based logic may

conceal target-private uncertainty. We therefore treat the car-overlap response as a trustworthiness signal, not direct evidence of collision risk.

8 Limitations and Future Directions

Scope and optimization budget. The study covers one target-private category, one source–target pair, and one detector family. Each student uses one seed, one backbone, and a fixed 3000-iteration budget (about 1.01 nominal source passes and 0.14 target passes). We did not perform a longer schedule, checkpoint sweep, or multi-seed analysis; bootstrap intervals capture evaluation-image sampling only. Diagnosis-subset AP is conditional on an autorickshaw-focused set. The gated stress test is not loss-scale matched to Standard DA, so isolating RoI selectivity would require normalization by $\sum_i w_i$ or an equivalent control.
Measurement. The absorption shorthand denotes an existential post-NMS GT–prediction overlap, not a unique assignment; it combines localization, calibration, and semantic output, is non-exclusive by class, and may include nearby-object geometry. Probe-box results depend on proposal geometry, the GT-anchored existential recovery rule, and 76.50% probe recall. The probe controls regions but does not read student representations. Retained diagnostic predictions were exported with the verified common score floor $s \geq 0.05$, while primary high-confidence analyses use $s \geq 0.5$.
Data and operational scope. The diagnosis set is confirmed image-disjoint from adaptation and probe training. It was inherited from the predefined 1000-image annotation JSON used by the earlier held-out experiment rather than newly sampled for this analysis. Sequence-level and near-duplicate separation were not audited. Tracking, planning, and control are not evaluated. The primary response–silence conclusion was checked at student score thresholds of 0.3 and 0.5 with IoU fixed at 0.5; a denser score and IoU sweep remains future work. Future work should also address multi-seed convergence, scale-matched gating, one-to-one geometry audits, broader categories, and temporal analysis.

9 Conclusion

We studied final-output handling of target-private autorickshaws in Cityscapes-to-IDD transfer. The principal finding is that absolute as-car overlap is coverage-confounded. Across runs, 97.08–98.00% of known-matched autorickshaw GTs also have a car match, so absolute as-car differences largely track whether a detector responds at all. Standard DA has fewer as-car matches but more misses, whereas the RoI-gated stress-test run has the highest diagnosis-subset AP and response coverage but more car matches. The supervised post-hoc probe yields the same descriptive high-confidence point-estimate ordering on a common recovered subset. Target-private evaluation should therefore report response, silence, and class-specific overlap jointly, rather than rely on known-class mAP or absolute as-car overlap alone.

References

1. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3339–3348 (2018). https://doi.org/10.1109/CVPR.2018.00352
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3223 (2016). https://doi.org/10.1109/CVPR.2016.350
3. Dhamija, A., Gunther, M., Ventura, J., Boult, T.: The overlooked elephant of object detection: Open set. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1021–1030 (2020). https://doi.org/10.1109/WACV45572.2020.9093355
4. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21795–21806 (2024). https://doi.org/10.1109/CVPR52733.2024.02059
5. Joseph, K.J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5826–5836 (2021). https://doi.org/10.1109/CVPR46437.2021.00577
6. Liang, W., Xue, F., Liu, Y., Zhong, G., Ming, A.: Unknown sniffer for object detection: Don't turn a blind eye to unknown objects. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3230–3239 (2023). https://doi.org/10.1109/CVPR52729.2023.00315
7. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Adv. Neural Inform. Process. Syst. vol. 33, pp. 21464–21475. Curran Associates, Inc. (2020)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
9. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: Eur. Conf. Comput. Vis. pp. 156–171 (2018). https://doi.org/10.1007/978-3-030-01228-1_10
10. Singh, D.K., Rai, S.N., Joseph, K.J., Saluja, R., Balasubramanian, V.N., Arora, C., Subramanian, A., Jawahar, C.: New objects on the road? no problem, we'll learn them too. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1972–1978 (2022). https://doi.org/10.1109/IROS47612.2022.9981886
11. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1743–1751 (2019). https://doi.org/10.1109/WACV.2019.00190
12. Wang, Z., Li, Y., Chen, X., Lim, S.N., Torralba, A., Zhao, H., Wang, S.: Detecting everything in the open world: Towards universal object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11433–11443 (2023). https://doi.org/10.1109/CVPR52729.2023.01100