Evaluating Pedestrian Fairness: Annotation and Sparsity Fragility

1st Nina Burdorf
CARIAD SE
Berlin, Germany
nina.burdorf@cariad.technology

2nd Christian Medeiros Adriano
Hasso Plattner Institute
Potsdam, Germany
christian.adriano@hpi.de

3rd Holger Giese
Hasso Plattner Institute
Potsdam, Germany
holger.giese@hpi.de

***Abstract*—Fairness evaluation in autonomous driving increasingly relies on demographic subgroup annotations to compare pedestrian detector performance across attributes such as perceived skin tone, age, gender presentation, and mobility-device use. Existing work typically treats subgroup labels and subgroup sample sizes as reliable inputs for disparity estimation. In this paper, we study when this assumption breaks down.**

Using aligned Berkeley DeepDrive 100K (BDD100K) based fairness annotations together with large-scale Zenseact Open Dataset (ZOD) object annotations, we analyze annotation disagreement, subgroup sparsity, statistical power, and visibility-dependent detector evaluation. We show that perceived skin-tone labels exhibit substantial disagreement across annotation schemes, while age and gender annotations are comparatively stable but remain sparse for several safety-critical groups. We further show through statistical power analysis that subgroup sizes comparable to observed wheelchair support in ZOD can only reveal comparatively large miss-rate disparities with conventional statistical confidence. Finally, a pilot YOLOX audit on wheelchair-associated pedestrians demonstrates that visibility filtering increases measured detection performance while simultaneously reducing subgroup support and statistical sensitivity.

Together, our findings expose an overlooked reliability gap in fairness evaluation for rare vulnerable-road-user groups, where annotation ambiguity, sparse subgroup support, and visibility-dependent evaluation choices may obscure safety-critical perception failures.

***Index Terms*—autonomous driving, fairness evaluation, annotation uncertainty, vulnerable road users, statistical power, long-tail distributions**

I. Introduction

Fairness in autonomous-driving perception systems is an important safety-critical problem, particularly for pedestrian detection across demographic groups. Prior work evaluates pedestrian detectors across attributes such as perceived skin tone, age, or gender presentation, reporting subgroup-conditioned miss-rate disparities [1], [2]. These evaluations are motivated by concerns that biased pedestrian detection performance could translate into unequal safety outcomes across demographic groups. However, existing fairness evaluations implicitly rely on two assumptions: demographic labels must be sufficiently reliable for stable subgroup assignment, and demographic groups must contain enough examples for statistically meaningful analysis. Both assumptions become difficult to satisfy in real-world autonomous-driving imagery, where pedestrians are frequently distant, occluded, motion blurred, poorly illuminated, or visually ambiguous. These challenges are particularly severe for rare and safety-critical vulnerable road users (VRUs), such as wheelchair users and children. Here, sparse subgroup support and uncertain demographic annotations can make subgroup-conditioned fairness estimates statistically fragile. In this paper, we therefore study the reliability of fairness evaluation itself. Rather than asking whether a particular detector is fair or unfair, we ask:

When do current fairness evaluations become unreliable?

Our contributions:

- We quantify demographic annotation disagreement across aligned autonomous-driving fairness annotations and analyze how disagreement depends on visual ambiguity.
- We demonstrate through statistical power analysis that sparse subgroup support fundamentally limits the sensitivity of fairness evaluation.
- We conduct a pilot YOLOX audit on wheelchair-associated pedestrians in ZOD and show that visibility filtering trades improved subgroup interpretability against statistical support.

II. Related Work

A. Fairness in Autonomous Driving

Recent work studies demographic disparities in autonomous-driving perception systems by evaluating pedestrian detectors across attributes such as perceived skin tone, age, and gender presentation [1], [2]. These studies report subgroup-conditioned miss-rate disparities and equalized-opportunity-style metrics under fixed benchmark annotations. At the same time, several works show that object detection performance depends strongly on environmental and visibility-related factors including illumination, weather, distance, occlusion, and scene composition [3]. Such findings suggest that demographic fairness cannot be separated from the visual conditions under which pedestrians appear. Existing fairness evaluations, however, largely treat demographic annotations and subgroup definitions as fixed and reliable inputs rather than studying their statistical stability.

B. Annotation Uncertainty and Sparse Support

Demographic fairness evaluation assumes stable group assignment and sufficient group-conditioned support. Recent

work in autonomous driving has begun to study annotation quality and ambiguity explicitly. Llorca et al. [1] introduce protected-attribute annotations for driving datasets and report that several demographic groups, including children and wheelchair users, are extremely underrepresented. More recently, Schwirten et al. [4] show that pedestrian annotations themselves can become ambiguous under realistic driving conditions and argue that annotation ambiguity is an important but underexplored dimension of dataset quality. These challenges are especially relevant for fairness evaluation because demographic labels determine subgroup membership, while fairness metrics themselves depend on sufficient subgroup support. Under long-tailed driving distributions, annotation disagreement and subgroup sparsity therefore interact directly as small numbers of label changes can substantially affect subgroup-conditioned estimates for rare groups. Existing autonomous-driving fairness work rarely studies this interaction explicitly.

C. Mobility-Impaired and Rare Vulnerable Road Users

Mobility-impaired VRUs form a useful stress test for fairness evaluation because they combine sparse support, visibility-related detection difficulty, and annotation ambiguity. Several datasets specifically target mobility-aid users and assistive devices [5], [6]. These datasets provide valuable coverage of wheelchairs, walkers, and related assistive technologies under curated conditions. However, such datasets differ substantially from real-world autonomous-driving imagery, where wheelchair users remain extremely rare and often appear at long distance, under occlusion, or with limited visual detail [7]. These challenges become even more severe for intersectional groups. Existing autonomous-driving datasets rarely provide simultaneous annotations for attributes such as mobility-device use, perceived skin tone, and age, making intersectional fairness evaluation difficult even before statistical sparsity is considered.

D. Statistical Reliability of Fairness Evaluation

Most fairness evaluations report disparity point estimates together with aggregate benchmark metrics [2], [8]. However, comparatively little work examines whether subgroup-conditioned fairness estimates are statistically identifiable under realistic dataset sizes [9]. This issue becomes especially important for rare and intersectional groups [10]. Even when overall datasets contain hundreds of thousands of pedestrian instances, subgroup support may remain too small to reliably detect modest disparities [9]. Our work studies this problem explicitly through statistical power analysis and visibility-aware detector evaluation.

III. Datasets and Experimental Setup

A. Datasets and Annotation Sources

We analyze fairness annotations derived from BDD100K [3] together with large-scale object annotations from the Zenseact Open Dataset (ZOD) [7]. For BDD100K, we compare demographic annotations released by Llorca et al. [1] and Li et al. [2]. Both works annotate perceived demographic attributes including skin tone, age, and gender presentation, while Llorca et al. additionally provide wheelchair-related annotations. After pedestrian alignment (section A), this yields 2,155 matched pedestrian instances with demographic labels from both annotation pipelines. For ZOD, we analyze the full object-detection annotations released by Alibeigi et al. [7]. Across all 100,000 images, we identify 6,895,692 annotated objects and 359,177 pedestrian instances. Unlike the BDD100K fairness annotations, ZOD does not provide demographic labels such as perceived skin tone or age. However, it includes 467 wheelchair-related vulnerable-road-user annotations that we use to study fairness evaluation under rare subgroup support.

B. Pedestrian Alignment

To compare annotations directly, we align pedestrian instances across annotation schemes using image correspondence and bounding-box overlap with one-to-one matching constraints (section A).

C. Controlled Detector Evaluation

To isolate annotation effects from detector effects, we use fixed YOLOX detection outputs released by Li et al. [2]. We select YOLOX from their models because it achieved strong overall performance across their experiments. Detector predictions remain unchanged across all annotation-sensitivity experiments for skin-tone, gender, and age.

For the wheelchair audit in ZOD, we additionally run a pretrained YOLOX pedestrian detector [11] on wheelchair-associated frames. Because wheelchair annotations in ZOD belong to vulnerable-vehicle categories rather than pedestrian categories, we construct a matching pipeline that associates wheelchair annotations with nearby pedestrian annotations. We therefore use spatial overlap and center-distance criteria (section B). We then evaluate whether the matched pedestrian instances are detected by YOLOX under different visibility and bounding-box-size thresholds.

IV. Results

A. Skin-Tone Labels Exhibit Annotation Disagreement

We first examine whether demographic annotations are stable enough to support subgroup-conditioned fairness evaluation. To do so, we compare aligned demographic annotations from Llorca et al. [1] and Li et al. [2]. Llorca et al. collect annotations from five annotators and explicitly allow an *unknown* label when demographic assignment is uncertain. In contrast, annotations by Li et al. are derived from a merged two-annotator pipeline without an equivalent uncertainty category.

After pedestrian alignment, the disagreement rate across demographic attributes Table I for perceived skin tone is 18.4% when uncertain labels are included and 10.1% after excluding them. The observed Cohen's $\kappa = 0.676$ further indicates that perceived skin-tone assignment remains sensitive to annotation protocol and visual ambiguity. In contrast, gender disagreement is only 3.7% with $\kappa = 0.913$, while age disagreement is 0.8%. Annotation uncertainty is therefore strongly attribute-dependent rather than uniform across demographic

labels. Qualitative inspection further shows that disagreement frequently occurs under difficult visual conditions including distance, occlusion, low illumination, motion blur, and small pedestrian size.

These findings suggest that demographic labels in autonomous-driving imagery should not always be treated as fixed ground truth, particularly for visually inferred attributes such as perceived skin tone. At the same time, apparent annotation stability can itself be misleading under sparse subgroup support. While age annotations exhibit very low disagreement, children account for only 32 of the 2,155 aligned pedestrian instances in the annotations by Llorca et al. . Reliable agreement estimates therefore become difficult precisely for the rare demographic groups that are often most safety-critical. There might therefore be an interaction between annotation uncertainty and sparse subgroup support.

TABLE I: Disagreement rates across demographic annotations.

Attribute	Disagreement (incl. unknown)	Disagreement (excl. unknown)
Skin tone	18.4%	10.1%
Gender	4.2%	3.7%
Age	0.9%	0.8%

B. Sparse Support Limits Statistical Reliability

We examine the statistical consequences of sparse subgroup support. While the previous section showed that demographic labels can be uncertain under difficult visual conditions, sparse demographic groups additionally limit whether fairness disparities can be detected reliably at all. To study this effect, we perform a two-proportion statistical power analysis and estimate the minimum detectable miss-rate gap as a function of subgroup size. We evaluate subgroup sizes ranging from extremely sparse groups ($n = 10$) to larger demographic subsets ($n = 5000$), including the observed wheelchair support in ZOD ($n = 467$). We further compare two dataset regimes: KITTI [1], [12], representing a comparatively small pedestrian dataset ($n = 6{,}336$ pedestrians), and ZOD, representing a large-scale pedestrian dataset ($n = 359{,}177$ pedestrians). Minimum miss-rate disparity Figure 1 can be detected with 80% statistical power at $\alpha = 0.05$. We use low baseline miss rates between 0.2% and 1.0% to reflect modern pedestrian detector performance, where even small absolute miss-rate differences may correspond to meaningful safety disparities.

The results show that sparse subgroup support severely limits the statistical power of fairness evaluation even in very large datasets. For example, consider a detector with an overall pedestrian miss rate of 0.5%. For a subgroup size comparable to the observed wheelchair support in ZOD ($n = 467$), a conventional two-proportion test with 80% power can only reliably detect subgroup miss rates of approximately 1.83% or higher. In other words, the subgroup miss rate would need to be more than three times larger than the baseline miss rate of 0.5% before the disparity becomes statistically detectable with high confidence. Smaller miss-rate differences remain statistically difficult to distinguish from sampling variability. This limitation is particularly important for rare and safety-critical vulnerable road users. Even when fairness disparities exist, sparse subgroup support may prevent them from being detected reliably. At the same time, the effective number of usable examples could be further reduced by annotation ambiguity, low visibility, subgroup construction challenges or simply visibility-related uncertainty. In practice, the effective number of reliably usable subgroup examples may be substantially smaller than nominal annotation counts suggest.

C. Wheelchair-Associated Pedestrians as a Stress Test for Fairness Evaluation

The previous analysis treated subgroup size as fixed. As discussed above, however, nominal subgroup counts may substantially overestimate the number of reliably usable evaluation examples. We use wheelchair-associated pedestrians as an illustrative stress test for fairness evaluation under extreme subgroup sparsity. To maintain consistency with prior fairness evaluations, we use the YOLOX pedestrian detector [11], as it was also employed by Li et al. [2] and, as mentioned, achieved strong overall performance across their demographic fairness experiments. Across all ZOD pedestrian annotations, we identify 467 wheelchair-related instances, corresponding to approximately 0.13% of all pedestrians. In the BDD100K dataset, Llorca et al. annotate only 8 instances of pedestrians using a wheelchair, corresponding to approximately 0.08% of pedestrian labels. These findings already indicate that mobility-impaired vulnerable road users remain extremely sparse even in large-scale autonomous-driving datasets. However, as mentioned in subsection IV-A and subsection IV-B nominal subgroup counts alone do not capture whether these examples are actually suitable for reliable fairness evaluation. To investigate this issue, we manually reviewed wheelchair-related annotations from the aligned BDD100K annotations released by Llorca et al. [1]. As shown in section C, several examples correspond to visually difficult cases or other mobility-related objects such as strollers or walkers rather than unambiguous wheelchair users. We additionally identify one duplicate annotation. Random visual inspection of wheelchair $n = 50$ annotations in ZOD reveals similar challenges caused by distance, occlusion, illumination, and small object size. Motivated by these findings, we next analyze wheelchair-associated pedestrians in ZOD under explicit visibility constraints. As wheelchair annotations in ZOD are represented as vulnerable-vehicle categories rather than pedestrian categories, we first associate wheelchair annotations with pedestrian annotations using spatial overlap and center-point proximity (section B). Applying the matching criteria described in section B reduced the initial 467 wheelchair annotations to 287 matched wheelchair-associated pedestrian instances. We use pedestrian bounding-box area as a proxy for visibility. We progressively restrict the analysis to larger bounding boxes to study the effect on detection performance.

Detection performance increases after restricting evaluation to larger pedestrian boxes, suggesting that visibility and prox-

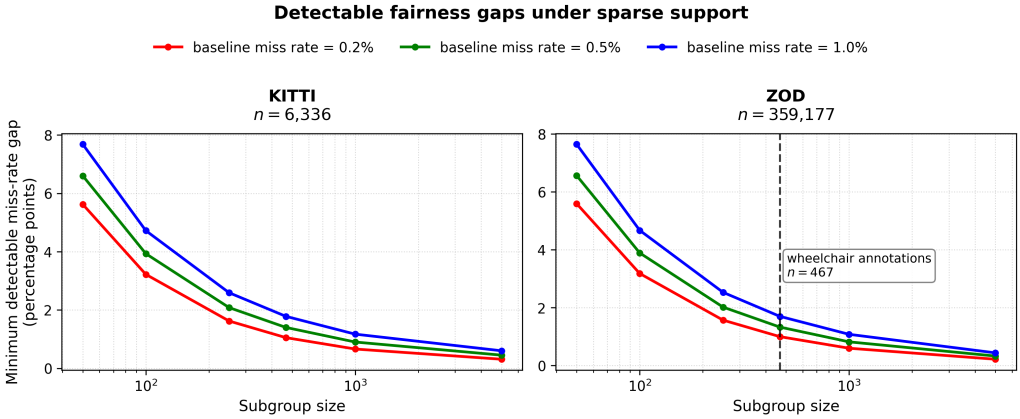


Fig. 1: Minimum detectable miss-rate gaps under varying subgroup sizes for KITTI and ZOD-scale pedestrian datasets. Sparse demographic groups require substantially larger disparities before statistically reliable detection becomes possible.

TABLE II: YOLOX detection performance for wheelchair-associated pedestrians under increasing visibility thresholds.

Minimum box area	Samples	Detection rate
All matched samples	287	8.0%
Area $\geq$ 2000	157	14.6%
Area $\geq$ 4000	108	21.3%
Area $\geq$ 6000	67	34.3%
Area $\geq$ 8000	49	46.9%
Area $\geq$ 10000	41	53.7%

imity strongly affect whether wheelchair-associated pedestrians can be reliably evaluated at all. At the same time, progressively stricter visibility thresholds drastically reduce subgroup support, shrinking the evaluation subset from 287 matched instances to only 41 pedestrians for bounding-box areas above $10{,}000$ pixels. This result directly connects to the statistical power analysis in the previous section. Increasing evaluation reliability through stricter visibility filtering simultaneously reduces statistical support and therefore limits the sensitivity of fairness analysis. In practice, improving subgroup quality may come at the cost of making subgroup-level conclusions increasingly underpowered. At the detector level, the observed detection rates also remain low. Detection performance reaches only 8.0% across all instances and approximately 53.7% even for the largest pedestrian boxes. While this experiment is exploratory and depends on the constructed subgroup definition, the results nevertheless suggest that detector behavior may vary substantially across different sensitive attributes and visibility conditions. In particular, a detector that performed strongly in prior fairness evaluations across perceived skin tone, age, and gender presentation does not necessarily generalize reliably to mobility-impaired vulnerable road users.

V. Discussion and Conclusion

Findings – Our results show that fairness evaluation in autonomous-driving perception systems may depend strongly on annotation reliability and subgroup support. While perceived skin-tone annotations exhibit disagreement under difficult visual conditions, age and gender annotations appear more stable but remain sparsely represented for several safety-critical groups such as children. Our statistical power analysis further shows that sparse subgroup support limits the sensitivity of fairness evaluation, even in large-scale datasets. The wheelchair experiments demonstrate that nominal subgroup counts can overestimate effective evaluable support. Restricting evaluation to larger and less ambiguous pedestrian instances increases measured detection performance, but simultaneously reduces subgroup support and therefore statistical power. Together, these findings suggest that fairness evaluation for rare VRU groups may depend not only on detector behavior, but also on annotation conventions, visibility constraints, subgroup-construction decisions, and statistical support.

Limitations – Further evaluation across multiple pedestrian detectors and broader multi-annotator studies is needed to assess the robustness and generalizability of our findings. Additionally, pedestrian alignment across annotation pipelines as well as wheelchair-pedestrian matching is approximate and may introduce residual matching uncertainty. Our analysis also focuses on perceived demographic attributes rather than self-identified demographics. Finally, intersectional fairness analysis still remains limited by missing demographic annotations in existing autonomous-driving datasets.

Implications and Future Work – Our findings suggest that fairness evaluation itself should be treated as a reliability problem rather than purely a detector-performance problem. Future work should therefore focus on targeted collection of mobility-impaired VRU data, more standardized and uncertainty-aware annotation protocols, and fairness metrics that remain robust under sparse or ambiguous labels. Another promising direction is the combination of real-world and synthetic data sources to reduce both annotation ambiguity and long-tailed subgroup sparsity [13]. Finally, mechanistic interpretability methods over models trained on sparse and safety-critical data [14] may help better understand detector behavior on underrepresented vulnerable-road-user distributions and safety conditions.

REFERENCES

[1] D. Fernández-Llorca, P. Frau, R. Izquierdo, and E. Gómez, "Attribute annotation and bias evaluation in visual datasets for autonomous driving," *Journal of Big Data*, vol. 11, 09 2024.

[2] X. Li, Z. Chen, J. M. Zhang, F. Sarro, Y. Zhang, and X. Liu, "Bias behind the wheel: Fairness testing of autonomous driving systems," 2024. [Online]. Available: https://arxiv.org/abs/2308.02935

[3] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," 2020. [Online]. Available: https://arxiv.org/abs/1805.04687

[4] L. Schwirten, J. Scholz, D. Kondermann, and J. Keuper, "Ambiguous annotations: When is a pedestrian not a pedestrian?" 2024. [Online]. Available: https://arxiv.org/abs/2405.08794

[5] S. Dávila-Soberón, A. Morales-Díaz, and M. Castelán, "A novel image dataset for detecting and classifying mobility aid users," *Expert Systems with Applications*, vol. 293, p. 128697, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417425023152

[6] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard, "Deep detection of people and their mobility aids for a hospital robot," in *Proc. of the IEEE Eur. Conf. on Mobile Robotics (ECMR)*, 2017. [Online]. Available: http://ais.informatik.uni-freiburg.de/publications/papers/vasquez17ecmr.pdf

[7] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 178–20 188.

[8] B. Pathiraja, C. Liu, and R. Senanayake, "Fairness in autonomous driving: Towards understanding confounding factors in object detection under challenging weather," 2024. [Online]. Available: https://arxiv.org/abs/2406.00219

[9] A. Ferrara, F. Cozzi, A. Perotti, A. Panisson, and F. Bonchi, "Size-adaptive hypothesis testing for fairness," 2026. [Online]. Available: https://arxiv.org/abs/2506.10586

[10] M. Molina and P. Loiseau, "Bounding and approximating intersectional fairness through marginal fairness," 2023. [Online]. Available: https://arxiv.org/abs/2206.05828

[11] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] L. Hennicke, C. M. Adriano, H. Giese, J. M. Koehler, and L. Schott, "Mind the gap between synthetic and real: Utilizing transfer learning to probe the boundaries of stable diffusion generated data," 2024. [Online]. Available: https://arxiv.org/abs/2405.03243

[14] N. Burdorf and C. M. Adriano, "Weight decay shapes representation geometry: Towards a more nuanced understanding of sparse autoencoders in vision transformers," in *ICML 2026 Workshop on Mechanistic Interpretability*, 2026, under revision.

Appendix A
Pedestrian Annotation Alignment

To compare demographic annotations across annotation schemes, we aligned pedestrian instances between the annotations released by Li et al. [2] and Llorca et al. [1]. As both works annotate pedestrians independently, direct comparison requires identifying corresponding pedestrian instances across datasets.

A. Image-Level Matching

We first matched annotations at the image level using shared image identifiers and file naming conventions. Only images available in both annotation releases were retained for subsequent analysis.

B. Bounding-Box Alignment

Within each matched image, pedestrian instances were aligned using bounding-box overlap and coordinate agreement. For each pedestrian annotation in the Llorca et al. dataset, all pedestrian annotations from the corresponding image in the Li et al. dataset were considered as candidate matches.

For every candidate pair, the intersection-over-union (IoU) between the two bounding boxes was computed. A candidate pair was retained if either:

- the IoU was at least 0.5, or
- the absolute difference between the corresponding bounding-box coordinates (x_0, y_0, x_1, y_1) was at most 5 pixels for all four coordinates.

To obtain one-to-one correspondences, retained candidate pairs were matched greedily based on IoU. For each Llorca annotation, the candidate match with the highest IoU was selected. If multiple Llorca annotations were assigned to the same Li et al. annotation, only the pair with the highest IoU was retained.

C. Handling Ambiguous Cases

Pedestrian instances for which no candidate satisfied the matching criteria described above were excluded from the agreement analysis. We additionally excluded uncertain labels from the agreement statistics reported in the main paper when explicitly stated.

This alignment procedure yielded matched pedestrian instances with demographic labels from both annotation pipelines, enabling direct comparison of perceived skin-tone, age, and gender annotations.

Appendix B
Wheelchair Annotation Matching in ZOD

ZOD [7] does not annotate wheelchair users directly as pedestrians. Instead, wheelchair-related instances are represented as vulnerable-vehicle annotations. Evaluating pedestrian detectors for wheelchair users therefore requires constructing associations between wheelchair annotations and nearby pedestrian annotations.

A. Wheelchair Frame Extraction

We parsed all ZOD object-detection annotation files and extracted frames containing vulnerable-vehicle annotations with subclass `Wheelchair`. This produced 467 wheelchair-related annotations across the dataset.

B. Pedestrian Candidate Selection

For each wheelchair-associated frame, all pedestrian annotations were extracted from the same annotation file. Pedestrian candidates were represented using axis-aligned bounding boxes.

C. Wheelchair–Pedestrian Matching

Wheelchair annotations were matched to pedestrian annotations within the same image. For each wheelchair annotation, we computed the intersection-over-union (IoU) and center-point distance to every pedestrian annotation in the corresponding frame. Candidate matches were ranked by descending IoU and ascending center-point distance. The highest-ranked pedestrian annotation was selected as the corresponding match.

To obtain a conservative evaluation subset, we retained only matches with an IoU greater than 0.05 or a center-point distance below 75 pixels. Applying these criteria reduced the initial set of 467 wheelchair annotations to 287 matched wheelchair-associated pedestrian instances.

D. Visibility Filtering

Visual inspection revealed that many matched instances were extremely small, visually ambiguous, partially occluded, or corresponded to other mobility-related objects such as strollers.

To investigate how visibility influences detector evaluation, we further filtered the matched pedestrian instances using minimum pedestrian bounding-box area thresholds. Detection performance is reported under multiple area thresholds to quantify the trade-off between improved visibility and reduced subgroup support. We report detector results under multiple minimum-area thresholds to study how visibility affects measured detection performance.

E. YOLOX Evaluation

A pretrained YOLOX pedestrian detector [11] was run on all matched wheelchair-associated frames. We used YOLOX because it was also used by Li et al. [2] and achieved strong overall performance across their experiments.

For each matched pedestrian instance, we computed:

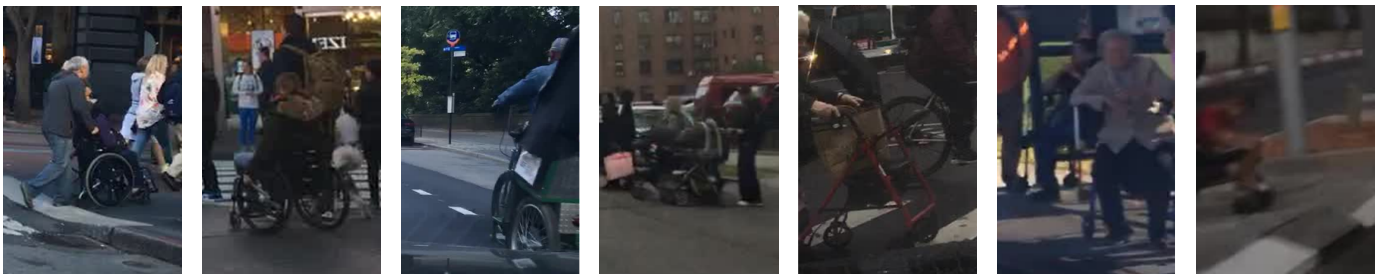
- maximum IoU with predicted pedestrian detections,
- maximum detection confidence score,
- number of pedestrian predictions per frame.

A matched pedestrian was considered detected if the best predicted pedestrian box exceeded the specified IoU threshold.

Appendix C
Qualitative Review of Wheelchair-Related Annotations

To better understand the practical challenges of fairness evaluation for rare mobility-impaired vulnerable road users, we manually inspected wheelchair-related annotations from the aligned BDD100K subset released by Llorca et al. [1]. The reviewed examples illustrate several recurring issues including visually ambiguous cases, small and distant objects, mobility devices that are difficult to distinguish from wheelchairs, and occasional annotation inconsistencies.

Representative examples are shown below.



Clear wheelchair | Clear wheelchair | Ambiguous | Ambiguous | Walker / PMD | Walker / PMD | Stroller

Fig. 2: Representative wheelchair-related annotations from the aligned BDD100K subset. Several examples correspond to visually difficult cases or other mobility-related objects rather than unambiguous wheelchair users.