Traffic and Safety Rule Compliance of Humans in Diverse Driving Situations

Michael Kurenkov[1,2,*]**, Sajad Marvi**[1,3,*]**, Julian Schmidt**[1]**, Christoph B. Rist**[1]**,
Alessandro Canevaro**[1]**, Hang Yu**[1]**, Julian Jordan**[1]**, Georg Schildbach**[2]**, Abhinav Valada**[3]

[1] Department of Automated Driving Software Functions, Mercedes Benz AG, Germany
[2] Department of Electrical Engineering in Medicine, University of Lübeck, Germany
[3] Department of Computer Science, University of Freiburg, Germany

Abstract: The increasing interest in autonomous driving systems has highlighted the need for an in-depth analysis of human driving behavior in diverse scenarios. Analyzing human data is crucial for developing autonomous systems that can replicate safe driving practices and ensure seamless integration into human-dominated environments. This paper presents a comparative evaluation of human compliance with traffic and safety rules across multiple trajectory prediction datasets, including Argoverse 2, nuPlan, Lyft, and DeepUrban. By defining and leveraging existing safety and behavior-related metrics such as time to collision, adherence to speed limits, and interactions with other traffic participants, we aim to provide a comprehensive understanding of each dataset's strengths and limitations. Our analysis focuses on the distribution of data samples, identifying noise, outliers, and undesirable behaviors exhibited by human drivers in both the training and validation sets. The results underscore the need for applying robust filtering techniques to certain datasets due to high levels of noise and the presence of such undesirable behaviors.

Keywords: Autonomous Driving, Dataset Analysis

1 Introduction

In recent years, autonomous vehicles (AVs) have gained significant attention due to their potential to reduce traffic fatalities. The widespread adoption of AV technology is contingent not only on technical performance but also on public trust, with concerns centering on safety and potential technological malfunctions [1, 2]. A key factor in improving trust in autonomous systems is the ability to understand and replicate human driving behavior. However, worldwide, road accidents cause over 1.19 million deaths annually, with a majority resulting from human error [3], hence following human driving pattern is not always desired. Since the majority of accidents are caused by human error, analyzing human driving data allows us to identify common mistakes and undesirable driving patterns. This understanding is crucial for training machine learning models, such as those used in behavior cloning, where the goal is to mimic human driving behavior. Identifying undesirable driving patterns is especially useful for achieving a defensive driving behavior, which is proven to play a significant role in increasing passenger comfort and trust in AVs [4].

In summary, our main contributions are:

- Review existing criticality measures and safety violations and define new relevant metrics
- Provide comprehensive overview of dataset characteristics
- Examine critical behavior across datasets and compare findings

---

*These authors contributed equally to this work.

Figure 1: Example of an onboard sensor dataset sample from (a) Argoverse2 [5] and (b) drone data DeepUrban [6].

Dataset	Cities	Recording Type	Total Volume (h)	Year	Access
Argoverse 1 [8]	2	Onboard	320	2019	Open
Lyft L5 [9]	1	Onboard	1118	2020	Open
Waymo [10]	6	Onboard	574	2021	Open
Shifts [11]	-	Onboard	1667	2021	Open
nuPlan [12]	4	Onboard	1500	2021	Open
Argoverse 2 [5]	6	Onboard	763	2023	Open
highD [13]	-	Drone	16	2018	Open
INTERACTION [14]	-	Drone	16	2019	Open
inD [15]	-	Drone	10	2020	Open
MONA [16]	-	Elevated Camera	130	2022	Open
DeepUrban [6]	-	Drone	66	2024	Restricted

Table 1: Naturalistic Driving Studies collection over the years.

2 Related Work

The following sections will provide a summary of the current state of the art in metrics, datasets, and trajectory prediction methodologies.

2.1 Datasets for Trajectory Prediction and Planning

This section explores key datasets containing human-driving data. The quality of machine learning models trained using these datasets (e.g., trajectory prediction models or planner models) relies heavily on the driving behavior and patterns demonstrated in these datasets. Early projects such as the 100-Car Naturalistic Study [7] collected large amounts of data but were primarily suited for perception tasks. More recent datasets are tailored for trajectory prediction, directly providing data on agent and lane level instead of raw sensor data only (Table 1).

Onboard Sensor Datasets: These datasets use multiple sensors (e.g., cameras, LiDAR, Radar) from the ego vehicle to detect its surroundings, capturing real-time measurements like speed and acceleration (Fig. 1(a)). However, onboard sensors are prone to noise and limited in their ability to capture distant traffic conditions. Despite these limitations, onboard sensor data is essential for the full AV stack, including trajectory prediction, planning, perception, and control, due to its ability to provide highly accurate internal vehicle signals.

External Sensor Datasets: External datasets, recorded via drones or elevated cameras, offer a top-down view of traffic as shown in Fig. 1(b). These methods offer an alternative perspective by precisely tracking all traffic participants, including those not visible from the ego vehicle's viewpoint. However, external sensor datasets face limitations such as restricted availability (strict flight regulations) and environmental dependencies (weather and lighting conditions).

2.2 Criticality Measures and Traffic Rule / Safety Violations

Ensuring and demonstrating safe driving behavior is a crucial step for the widespread adoption of AVs, particularly for the certification of SAE Level 5 systems [17], where human intervention is not possible. Both the research and industrial sectors are keen on identifying critical driving situations and maneuvers [18] and ensuring appropriate responses.

Criticality Measures: Criticality measures are essential for analyzing human driving behavior, helping to quantify the risk involved in specific traffic scenarios. Several key metrics are commonly used for this purpose, including Gap, Time to Collision (TTC), and Post Encroachment Time (PET) [19].

Gap: The Gap metric measures the distance between two agents A_i and A_j at a given time t. In a car following scenario, this is described as:

$$\text{GAP}(A_i, A_j, t) = \text{distance}(p_{A_i,\text{front}}(t), p_{A_j,\text{rear}}(t)), \tag{1}$$

where $p_{A_1,\text{front}}(t)$ is the 2D position of the front of agent i at time step t. The Headway (HW) is a related measure that calculates the distance between the fronts of two agents [20].

Time to Collision (TTC): TTC estimates the time until two agents would collide given a dynamic motion model at the current time step [20, 19]. It is widely used as a surrogate safety indicator. The critical threshold for TTC, TTC^*, varies across the literature, with suggested values ranging from 1.5 to 4 seconds [21, 22, 23]. It is calculated as

$$\text{TTC}(A_i, A_j, t) = \min\left(\{\tilde{t} \geq 0 \mid d(p_{A_i}(t + \tilde{t}), p_{A_j}(t + \tilde{t})) = 0\} \cup \{\infty\}\right). \tag{2}$$

There are also different variations of TTC, which have been used for evaluating the quality of planned trajectory. One variation used in CommonRoad [24] takes into consideration the relative velocity, acceleration, and Gap in the car following the scenario.

Post Encroachment Time (PET): PET calculates the time interval between one agent exiting and another agent entering a designated conflict area (CA):

$$\text{PET}(A_i, A_j, \text{CA}) = t_{\text{entry}}(A_j, \text{CA}) - t_{\text{exit}}(A_i, \text{CA}). \tag{3}$$

PET can also be semi-predictive, with $t_{\text{entry}}(A_j, \text{CA})$ being predicted using motion models such as constant velocity.

Traffic Rule Violations: Legal compliance is a critical aspect of analyzing Naturalistic Driving Studies, as most accidents are caused by human error [25]. While criticality measures may detect potential risks, they do not always capture traffic rule violations, which can also increase the risk of accidents. AVs must adhere strictly to traffic regulations, and defining such rules precisely for autonomous systems is essential. Driver errors and rule violations can be classified into three categories: recognition, decision, and reaction errors, with most accidents caused by the first two [26]. In this study, the focus is on analyzing rule violations, particularly according to German law (StVO) [27]. Key traffic rules examined include:

- Boundaries: Prohibition of crossing solid lines (StVO Section 2).
- Overtaking: Maintaining a minimum distance of 1.5 meters when overtaking pedestrians or cyclists in urban areas and 2 meters outside urban areas (StVO Section 5).
- Pedestrian Crossings: Vehicles must stop for pedestrians at crossings and must not block the crossing during congestion (StVO Section 26).

By examining such violations in motion prediction datasets, this analysis helps ensure that AVs comply with traffic regulations to reduce the risk of accidents.

3 Critical Behavior Extraction Method

In this section, first the various metrics used for dataset analysis are discussed. The second part focuses on the implementation of the metrics and details of the dataset analysis.

3.1 Definition of Criticality Measures and Safety Rule Violations

The analysis of human driving behavior, particularly in the context of autonomous driving, requires specific criteria that are generally applicable to all datasets. The following criteria were used in this work:

Speed (VEL): One of the fundamental criteria is the speed of vehicles, which can be calculated from the agents' positions:

$$\text{VEL} = \frac{p_{A_i}(t+1) - p_{A_i}(t)}{\Delta t}, \tag{4}$$

where $p_{A_i}(t)$ is the position of agent i at time t, and Δt is the time step. For this analysis, speeds above 14 m/s (around 50 km/h) are considered critical, based on urban area speed limit regulations of the US (30mph $\sim$ 48kph) and Germany (50kph).

Acceleration (ACC): Acceleration is another key criterion computed from the vehicle's speed:

$$\text{ACC} = \frac{v_{A_i}(t+1) - v_{A_i}(t)}{\Delta t}, \tag{5}$$

where $v_{A_i}(t)$ is velocity of the agent i at time t. Acceleration values below $-6m/s^2$ and above $6m/s^2$ are considered critical, as research has shown that such values can cause discomfort or even whiplash [28].

Gap: The calculation of the Gap is based on Eq.1. We consider the trajectory of the ego as a reference, and if other agents' polygon at each time step are positioned in the future trajectory of the ego, the minimal distance between two agents along the trajectory is considered as GAP value (Fig. 2 (b)).

Time to Collision (TTC): TTC is a key criterion used to assess how long it will take for two vehicles to collide if their current trajectories are maintained. The general formula for TTC is shown in Equation 2. It calculates the minimum time t_{ttc} until a potential collision between two vehicles, A_i and A_j, based on their current trajectories. To make this calculation more general and not just limited to car-following scenarios, the vehicles' future trajectories are used as the reference paths. At each discrete time step $t_{\text{ttc}} \in \{0.5, 1.0, 1.5, \ldots, 39.5, 40\}$, the predicted positions $\hat{p}_{A_i}(t+t_{\text{ttc}}), \hat{p}_{A_j}(t+t_{\text{ttc}})$ for vehicles A_i, A_j are computed based on their current speed v_t and acceleration a_t by interpolating the traveled distance:

$$\Delta X = v_t \cdot t_{\text{ttc}} + \frac{1}{2} \cdot a_t \cdot t_{\text{ttc}}^2$$

over the reference path. The TTC value is determined as the first time step when the vehicles occupy the same position, indicating a potential collision. (Fig. 2(a)).

$$\text{TTC} = min(\{t_{\text{ttc}}|\hat{p}_{A_i}(t+t_{\text{ttc}}) \cap \hat{p}_{A_j}(t+t_{\text{ttc}})\}), \text{ for } t_{\text{ttc}} \in \{0.5, 1.0, 1.5, \ldots, 39.5, 40\}\text{s}, \tag{6}$$

where $\hat{p}$ is the function that outputs the interpolated future position of the agent. Critical TTC thresholds typically vary depending on environmental factors and road conditions. Values in the range of 1.5 to 4 seconds have been suggested in the literature [29, 30]. For this study, a threshold of 2 seconds is chosen to ensure that only the most critical situations are flagged.

Distance to Bicycles (DTB): The Distance to Bicycles (DTB) criterion is used to ensure safe interactions between vehicles and cyclists. The distance from a vehicle A to a cyclist C is calculated at each time step as

$$\text{DTB} = \text{distance}(p_A(t), p_C(t)). \tag{7}$$

This analysis focuses on values within a 5 meter radius. To address specific legal requirements, such as the StVO in Germany, DTB is further divided into longitudinal and lateral components. The longitudinal DTB (LODTB) measures the forward or backward distance along the vehicle's direction of travel relative to the cyclist, while the lateral DTB (LADTB) assesses the lateral distance between the vehicle and cyclist.

A minimum lateral distance of 1.5 meters must be maintained during overtaking in urban areas.
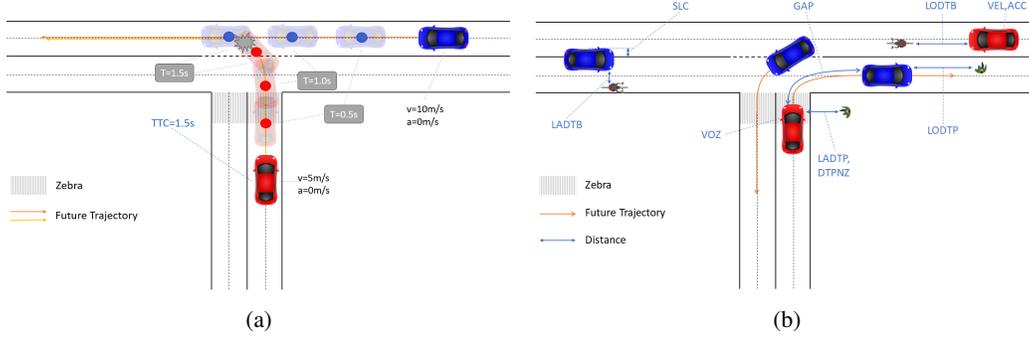
Figure 2: (a) TTC calculation where the red vehicle turns left and the blue vehicle follows the lane, using constant acceleration for future trajectory. (b) Visualization of all metrics except TTC in our framework.

Name	Category	Unit	Requires Map	Criticality Threshold
VEL	Traffic Violation	m/s	No	$VEL > 14\,m/s$
ACC	Criticality Measure	m/s^2	No	$|ACC| > 6\,m/s^2$
GAP	Criticality Measure	m	No	-
TTC	Criticality Measure	s	No	$TTC < 2\,s$
DTB	Traffic Violation	m	No	$DTB < 1.5\,m$
DTP	Traffic Violation	m	No	$DTP < 1.5\,m$
DTPNZ	Traffic Violation	m	Yes	$DTPNZ < 1.5m$
VOZ	Criticality Measure	m/s	Yes	-
SLC	Traffic Violation	m	Yes	$SLC > 0\,m$

Table 2: Metrics that are analyzed in the work with criticality threshold.

Distance to Pedestrians (DTP): Similar to the DTB, the Distance to Pedestrians (DTP) criterion measures the distance between a vehicle A and pedestrians P as

$$DTP = distance(p_A(t), p_P(t)). \tag{8}$$

As with the DTB, this distance is evaluated within a 5 meter range and divided into longitudinal (LODTP) and lateral (LADTP) components to ensure compliance with pedestrian safety regulations. The critical threshold of 1.5 meters is again used as the minimum distance vehicles should maintain from pedestrians in urban areas.

Pedestrian Crossing (DTPNZ): The DTPNZ criterion, a subset of the DTP, measures the proximity of pedestrians to the vehicle when it crosses a pedestrian zone. The criterion, therefore, gives special attention to whether the vehicle complies with traffic rules when approaching crosswalks. The critical threshold of 1.5 meters applies here as well, ensuring pedestrian safety.

Speed at Pedestrian Crossings (VOZ): This criterion measures vehicle speed when crossing pedestrian zones, derived from the general speed criterion (VEL) [31, 32]. Speeds are analyzed to ensure they comply with legal speed limits and safety regulations near crosswalks.

Solid Line Crossing (SLC): Crossing solid lines is a common traffic violation. The distance between the vehicle's outer tire and the solid line is measured as

$$SLC = distance(outer\_tire(A_t), solid\_line). \tag{9}$$

This criterion helps assess whether the vehicle complies with lane discipline rules, as crossing a solid line can increase the risk of collisions. All the investigated criteria and their corresponding critical thresholds are aggregated in Table 2.

3.2 Analysis of Datasets

The analysis of various datasets is a central component of this work. To make the evaluation process more efficient, we used a common internal representation, similar to TrajData [33] of the traffic scene

Metric	Lyft		nuPlan		Argoverse 2		DeepUrban
	Train	Val	Train	Val	Train	Val	
VEL (ms^{-1})	5.78	5.8	7.38	6.51	7.11	7.13	6.39
ACC (ms^{-2})	0.11	0.11	0.14	0.14	0.14	0.14	0.2
GAP (m)	25.84	25.74	20.53	19.45	30.10	20.60	18.40
TTC (s)	6.0	6.0	7.0	7.0	6.0	6.0	7.0
LADTB (m)	2.13	2.04	1.75	1.86	1.91	1.69	2.03
LODTB (m)	7.17	7.81	7.83	7.42	7.17	6.69	7.46
LADTP (m)	2.4	2.39	2.35	2.32	2.16	2.15	2.18
LODTP (m)	0.98	1.81	3.68	3.64	3.75	2.54	3.68
DTPNZ (m)	3.59	3.52	2.61	3.11	3.07	3.0	3.36
VOZ (ms^{-1})	3.86	3.82	4.45	4.5	4.96	5.0	2.88
SLC (m)	0.23	0.22	0.20	0.24	0.22	0.21	0.11

Table 3: Comparison of median values for every metric and dataset.

in our problem formulation. The representation contains key data required for safe autonomous driving, such as:

- *Map data:* Contains information about lane borders, lane types (e.g., bus, bicycle lanes), boundaries (e.g., dashed, solid lines), crosswalks, and restricted areas. A graph-like structure links lanes via predecessors, successors, and neighbors.
- *Agent data:* Tracks the ego vehicle and other agents (e.g., buses, bicycles, pedestrians), including their position, speed, dimensions, and categories.
- *General data:* Stores scenario metadata such as the recording city, dataset name, frame rate, and time of day.

We standardized scenarios to 11 seconds at 10 frames per second, including only relevant map regions within 150 meters of the agents. The framework classifies agents and scenarios as critical or non-critical based on metrics and thresholds, filtering out undesirable behaviors for downstream tasks. It also generates metadata on critical agents or scenarios, as shown in Table 4.

4   Results

The analysis focus on evaluating the performance of Lyft, nuPlan, Argoverse 2 (AV2), and Deep-Urban based on the defined metrics and examining the critical agents. The distribution of each criterion is illustrated in Fig. 3 and their corresponding medians are presented in Table 3 for more detail information.

The velocity distribution across datasets revealed that most vehicles navigate at slower speeds, with a median velocity below 8 m/s in all datasets. The Lyft dataset showed the widest range of velocities and the highest number of outliers, reaching almost 100 m/s, likely due to noise in the data. The highest median velocity was observed in the nuPlan dataset (7.39 m/s). Values for outliers in datasets suggested that these velocities are too high to be valid, except for DeepUrban and nuPlan, where top velocities of 28 m/s and 35 m/s, respectively, are more plausible for their scenarios.

Acceleration distributions center around 0 m/s², except for DeepUrban, which has a narrower range. Lyft has the most extreme values, suggesting that velocity outliers influence acceleration noise in both Lyft and AV2, while DeepUrban has the most precise measurements.

The gap distribution varied significantly between datasets, with AV2 showing the highest median gap (30.21 meters), followed by Lyft (25.9 meters). A notable observation is the difference between AV2's training and validation datasets, where agents in the validation set maintained shorter distances from each other compared to the training set. High gap values are generally safer, but shorter gaps with stationary vehicles are also safe, illustrating that gap alone is not necessarily a measure of criticality, therefore, we suggest using this metric in combination with others. TTC is a crucial metric for assessing potential collisions. The Lyft and AV2 datasets both have a median TTC of 6 seconds, with similar distributions between their training and validation sets. nuPlan and Deep-
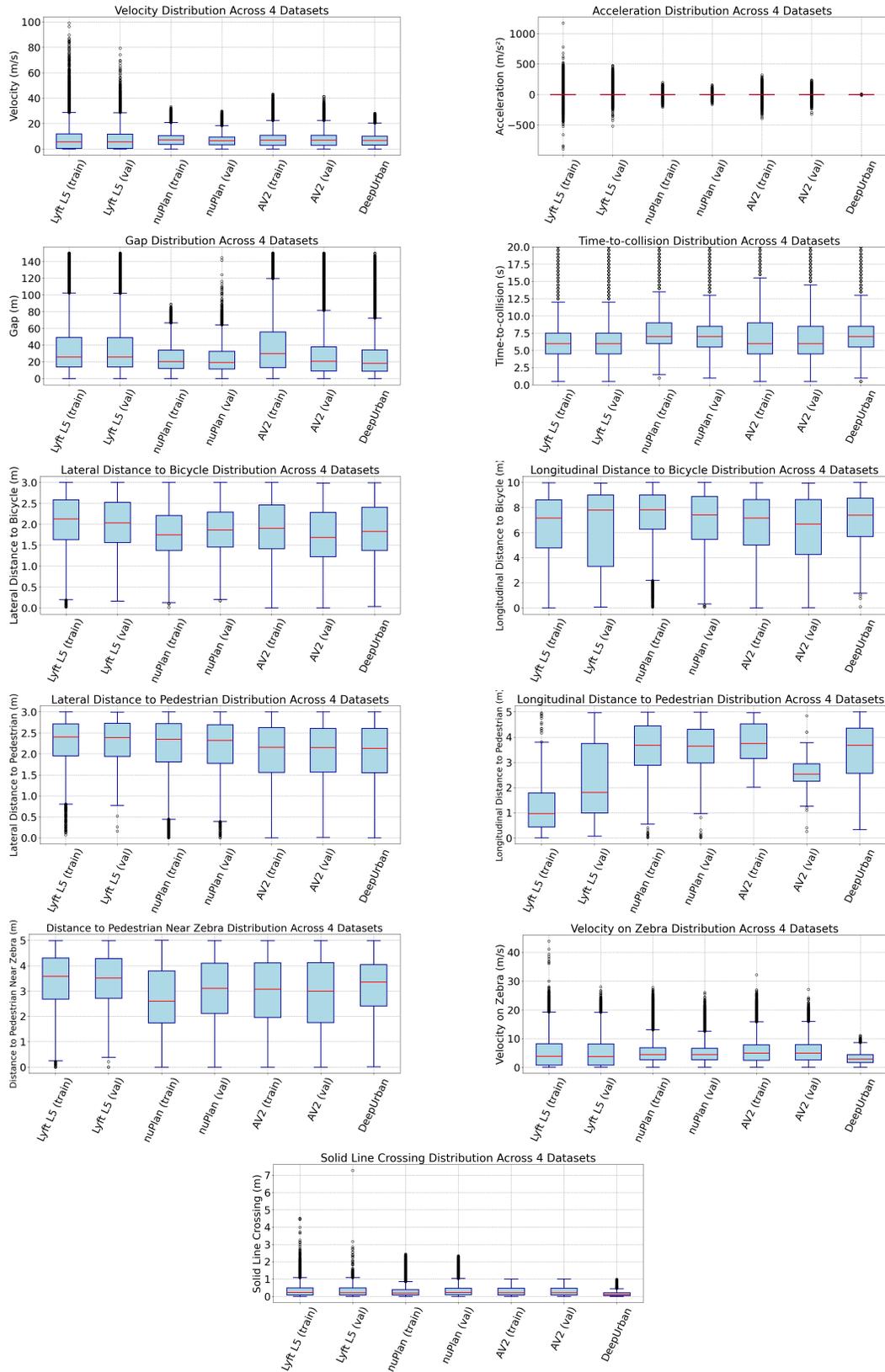
Figure 3: Comparison of distributions of different criteria in datasets.

Dataset	Total Agents	VEL	ACC	TTC	DTPNZ	LADTB	SLC
Lyft (train)	40,938,112	4.48%	9.57%	0.05%	0.00%	0.00%	0.16%
Lyft (val)	4,802,359	4.37%	9.49%	0.06%	0.00%	0.00%	0.15%
nuPlan (train)	12,870,111	1.32%	1.32%	0.00%	0.12%	0.01%	0.88%
nuPlan (val)	2,376,986	0.84%	1.02%	0.00%	0.07%	0.01%	1.76%
AV2 (train)	8,141,840	6.32%	11.66%	0.42%	0.08%	0.01%	0.78%
AV2 (val)	1,020,858	6.34%	11.69%	0.38%	0.10%	0.01%	0.79%
DeepUrban	909,507	3.01%	0.00%	0.00%	0.01%	0.04%	0.04%

Table 4: Comparison of critical agent amount based on different metric thresholds in proposed datasets.
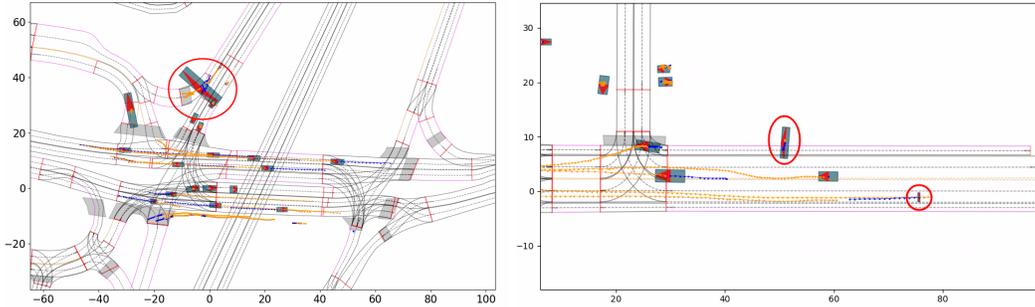
Figure 4: Examples of noisy data in Lyft dataset which have wrong heading angle.

Urban have higher median TTC values of 7 seconds, indicating safer behavior among agents. Both datasets also show a similar distribution, particularly in nuPlan's validation set. Low-range outliers, particularly in nuPlan and DeepUrban, have TTC values as low as 1 second or 0.5 seconds. The Lyft dataset demonstrates a more defensive driving style toward bicycles in LADTB, compare to all other datasets which lower quartiles fell below critical value of 1.5 meters.

Among the four metrics capturing driver-pedestrian interactions, the largest differences appear in longitudinal distance. The Lyft dataset shows more aggressive behavior, while AV2's training set exhibits more desirable behavior with most cases above 2 meters, though its validation set presents more challenging situations. DeepUrban demonstrates a more defensive driving style in VOZ, with lower median and tail values. In SLC, most agents fall between 0 and 1, showing small deviations from the center line. However, Lyft contains unrealistic cases exceeding 3 meters in SLC, likely due to errors in bounding box orientation or size estimation, as shown in Fig. 4.

5 Conclusion

The analysis of human driving behavior across four key datasets revealed significant differences in driving behavior and noise levels. The Lyft dataset exhibited the most noise, with extreme outliers in both velocity and acceleration, highlighting the need for thorough preprocessing to ensure accurate model training. In contrast, the DeepUrban dataset demonstrated more controlled and realistic driving behaviors, with fewer outliers and a consistently defensive driving style, making it a reliable source for analysis. The AV2 shows a significantly higher percentage of critical TTC values, with up to 0.42% in the training set and 0.38% in the validation set, compared to other datasets such as Lyft and nuPlan, which mostly hover around 0%. This higher percentage in AV2 highlights a distinct driving pattern within the dataset. The nuPlan and AV2 datasets showed balanced driving behaviors, representing a middle ground between the more defensive DeepUrban and the outlier-heavy Lyft. The comparative analysis between Lyft and AV2 also revealed that advancements in sensors and post-processing techniques have led to noticeable reductions in noise levels over time. Additionally, datasets captured via drones generally exhibited lower noise levels than those recorded using on-board sensors. This difference is likely due to the broader, less obstructed perspective provided by drones, whereas onboard sensors are limited by the ego vehicle's viewpoint. By employing the proposed framework, we can effectively identify abnormal behaviors across these datasets, facilitating the training of models that focus on more desirable human driving patterns.

References

[1] C. Bieber. 93% Have Concerns About Self-Driving Cars According to New Forbes Legal Survey, 2024.

[2] L. Londoño, J. V. Hurtado, N. Hertz, P. Kellmeyer, S. Voeneky, and A. Valada. Fairness and bias in robot learning. *Proceedings of the IEEE*, 2024.

[3] World Health Organization. Global status report on road safety 2023, 2023. URL https://www.who.int/publications/i/item/9789241565684.

[4] F. Hartwich, C. Hollander, D. Johannmeyer, and J. F. Krems. Improving passenger experience and trust in automated vehicles through user-adaptive hmis: "the more the better" does not apply to everyone. *Frontiers in Human Dynamics*, 3, 2021.

[5] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

[6] C. Selzer and F. B. Flohr. Deepurban: Interaction-aware trajectory prediction and planning for automated driving by aerial imagery. In *IEEE International Conference on Intelligent Transportation Systems*, volume 27. IEEE, 2024.

[7] V. L. Neale, T. A. Dingus, S. G. Klauer, J. Sudweeks, and M. Goodman. An overview of the 100-car naturalistic study and findings. *National Highway Traffic Safety Administration, Paper*, 5:0400, 2005.

[8] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019.

[9] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020.

[10] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. *arXiv preprint arXiv:2104.10133*, 2021.

[11] A. Malinin, N. Band, Ganshin, Alexander, G. Chesnokov, Y. Gal, M. J. F. Gales, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina, V. Raina, Roginskiy, Denis, M. Shmatova, P. Tigas, and B. Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2022.

[12] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2022.

[13] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018.

[14] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, and M. Tomizuka. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019.

[15] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein. The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1929–1934, 2020.

[16] L. Gressenbuch, K. Esterle, T. Kessler, and M. Althoff. Mona: 25th ieee international conference on intelligent transportation systems, itsc 2022. *2022 IEEE 25th International Conference on Intelligent Transportation Systems*, page 2093–2100, 2022.

[17] SAE. J3016_202104: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, Apr. 2021. URL https://www.sae.org/standards/content/j3016_202104/.

[18] J. Schmidt, J. Jordan, D. Raba, T. Welz, and K. Dietmayer. Meat: Maneuver extraction from agent trajectories. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 1810–1816. IEEE, 2022.

[19] L. Westhofen, C. Neurohr, T. Koopmann, M. Butz, B. Schütt, F. Utesch, B. Kramer, C. Gutenkunst, and E. Böde. Criticality metrics for automated driving: a review and suitability analysis of the state of the art. *Archives of Computational Methods in Engineering*, 30 (1):1–35, 2023.

[20] J. C. Hayward. Near miss determination through use of a scale of danger. *Pennsylvania State University University Park*, 1972.

[21] R. Van der Horst. Time-to-collision as a cue for decision-making in braking. *Vision in Vehicles III*, pages 19–26, 1991.

[22] S. Naseralavi, N. Nadimi, M. Saffarzadeh, and A. R. Mamdoohi. A general formulation for time-to-collision safety indicator. *Proceedings of the ICE - Transport*, 166:294–304, 2013.

[23] D. Deveaux, T. Higuchi, S. Uçar, C.-H. Wang, J. Härri, and O. Altintas. Extraction of risk knowledge from time to collision variation in roundabouts. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, page 3665–3672, 2021.

[24] Y. Lin and M. Althoff. Commonroad-crime: A toolbox for criticality measures of autonomous vehicles. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, page 1–8. IEEE, 2023.

[25] S. Singh. Critical reasons for crashes investigated in the national motor vehicle crash causation survey. *Traffic Safety Facts - Crash Stats*, 2018. URL https://trid.trb.org/View/1507603.

[26] A. J. Khattak, N. Ahmad, B. Wali, and E. Dumbaugh. A taxonomy of driving errors and violations: Evidence from the naturalistic driving study. *Accident Analysis and Prevention*, 151:105873, Mar. 2021.

[27] Bundesamt für Justiz. Stvo 2013 - nichtamtliches inhaltsverzeichnis, 2013. URL https://www.gesetze-im-internet.de/stvo_2013/index.html.

[28] L. Masello, G. Castignani, B. Sheehan, M. Guillen, and F. Murphy. Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis and Prevention*, 184:106997, 2023.

[29] M. Benmimoun and L. Eckstein. Detection of critical driving situations for naturalistic driving studies by means of an automated process. *Journal of Intelligent Transportation and Urban Planning*, 2(1):11–21, 2014.

[30] ISO. Intelligent transport systems—forward vehicle collision warning systems—performance requirements and test procedures. In *the International Organization for Standardization*, 2013.

[31] A. Varhelyi. Drivers' speed behaviour at a zebra crossing: a case study. *Accident Analysis and Prevention*, 30(6):731–743, 1998.

[32] M. Budzynski, A. Gobis, L. Guminska, L. Jelinski, M. Kiec, and P. Tomczuk. Assessment of the influence of road infrastructure parameters on the behaviour of drivers and pedestrians in pedestrian crossing areas. *Energies*, 14, 2021.

[33] L. Feng, M. Bahari, K. M. B. Amor, E. Zablocki, M. Cord, and A. Alahi. Unitraj: A unified framework for scalable vehicle trajectory prediction. *arXiv preprint arXiv::2403.15098*, 2024.