

Autonomous Vehicle's Impact on Traffic: Empirical Evidence From Waymo Open Dataset and Implications From Modelling

Xiangwang Hu¹, Zuduo Zheng², Danjue Chen, and Jian Sun³

Abstract—Previous empirical behavior analysis on Autonomous Vehicles (AV) mainly focused on vehicles with Adaptive Cruise Control (ACC) system due to the lack of high-level AV dataset. Recently released SAE Level-4 AV datasets such as the Waymo Open Dataset provide great opportunities to evaluate their behavioral impact on traffic flow. In this study, we aim to characterize the empirical Car Following (CF) behaviors of the Waymo autonomous vehicle and compare its feature with human-driven Vehicles (HV), and capture such behavioral differences using the IDM CF model. Our main findings include: (a) AV is much safer than HV, based on our analysis using surrogate safety measures, as time headways and jam spacings of the AV are significantly larger than HV; (b) the response time of AV is also significantly larger than that of HV in response to various types of stimuli; (c) despite the short length of trajectories in the Waymo Open Dataset, we have confirmed that these trajectories are suitable for calibrating some of the IDM parameters; and the calibration results of IDM are consistent with our empirical analysis. Moreover, the modelling results, reveal that the proportion of string unstable behavior of AV is less than that of HV; and (d) for HV, there is generally no significant difference between following AV and following HV except a smaller jam spacing when following AV. Overall, we conclude that currently AV behaves in a conservative way to ensure its safety at the cost of traffic efficiency.

Index Terms—Autonomous vehicle, car following, traffic safety, traffic efficiency, wavelet analysis.

I. INTRODUCTION

ALTHOUGH research on car following (CF) behavior of traditional Human-driven Vehicles (HV) is extensive, very few studies are oriented to Autonomous Vehicles (AV)

partly due to the deployment of AV is still in its infant stage. For a long period, most researchers investigated the impact of AV on traffic flow by simulation method where many assumptions need to be made such as the perception delay/error, time headway, and reaction time, see [1] for a thorough review.

It was not until recent years that field experiments on AV were carried out. A review of empirical analysis on AV related field experiment data is presented in Table I. Using a two-vehicle platoon, Milanés and Shladover [2] tested the factory Adaptive Cruise Control (ACC), the IDM based ACC and a proposed Cooperative ACC (CACC) algorithm on homogeneous Infiniti M56s and found that the factory ACC was string unstable due to delay and overshoot but CACC could overcome this issue. Knoop et al. [3] created a platoon of 7 vehicles (from 4 different makes) equipped with ACC and operated the experiment on public roads for about 500km. They concluded that it was difficult to maintain a platoon of more than three to four vehicles due to the disturbances from busy traffic condition and the instabilities in the CF behavior. The string stability of commercially implemented ACC systems were also assessed through CF model calibrated from experiment data and results revealed that all the ACC systems were string unstable [4], [5]. Many more field experiment campaigns in which various platoon sizes (2-ACC, 5-ACC, 10-ACC platoons) under public roads and closed test tracks were organized by the Joint Research Center of European Commission and the collected data are available in the OpenACC dataset [6]. Empirical analysis on the dataset demonstrated that the response times of the equipped ACC systems are comparable to Human-driven Vehicles (HV) and instability in the vehicle-platoon was also displayed [7], [8]. A recent ACC experiment by Li et al. [9], [10] focused more on the behavior of ACC in different ACC settings, traffic conditions and stimuli. They concluded that for a single ACC vehicle the ACC response could amplify or dampen an oscillation but for long platoons the oscillation amplitude tended to exacerbate very quickly.

However, most empirical evidence mentioned above are based on data acquired from SAE Level 2 AVs, a CF behavior analysis of high level (SAE Level 4-5) AV is missing. Recently, several new datasets pertaining to SAE Level 4 AV have been released such as Argo Dataset [11], Lyft Level 5 AV Dataset [12], nuScenes Dataset [13] and Waymo Open Dataset [14], [15]. These data are usually collected

Manuscript received 6 March 2022; revised 17 August 2022 and 1 December 2022; accepted 12 December 2022. Date of publication 22 March 2023; date of current version 31 May 2023. This work was supported in part by the Australian Research Council under Grant DP210102970, in part by the National Natural Science Foundation of China under Grant 52125208, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, in part by NSF under Award 1826162 and Award 1932921, and in part by the China Scholarship Council (CSC). The Associate Editor for this article was H. Farah. (Corresponding author: Zuduo Zheng.)

Xiangwang Hu and Jian Sun are with the Department of Traffic Engineering, and the Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji University, Shanghai 201804, China (e-mail: huxiangwang@tongji.edu.cn; sunjian@tongji.edu.cn).

Zuduo Zheng is with the School of Civil Engineering, The University of Queensland, Brisbane, QLD 4072, Australia (e-mail: zuduo.zheng@uq.edu.au).

Danjue Chen is with the Department of Civil and Environmental Engineering, University of Massachusetts Lowell, Lowell, MA 01854 USA (e-mail: danjue_chen@uml.edu).

Digital Object Identifier 10.1109/TITS.2023.3258145

1558-0016 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

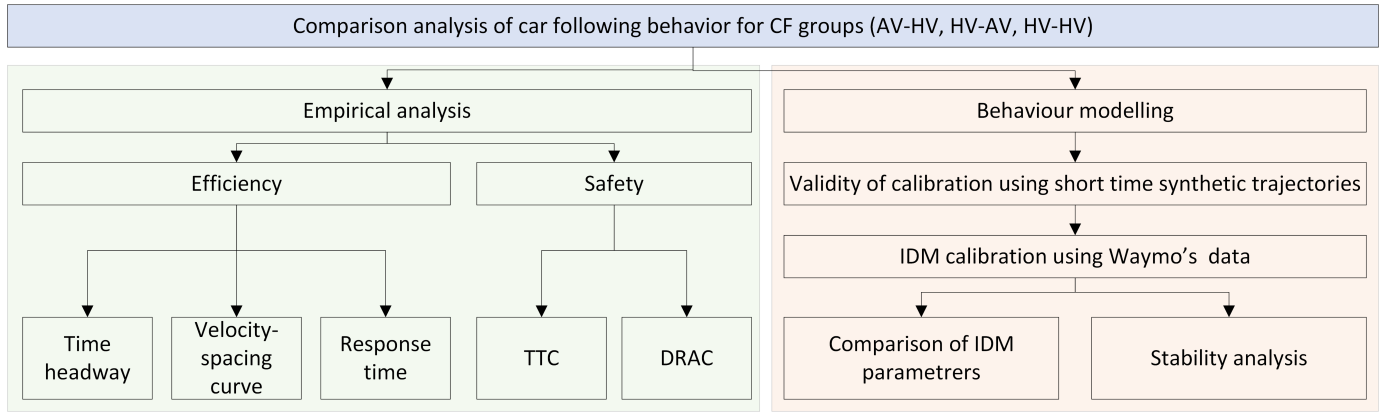


Fig. 1. Flowchart of CF behavior comparison analysis.

TABLE I
REVIEW OF EMPIRICAL ANALYSIS ON AV-RELATED FILED EXPERIMENT DATA

Author and Year	Test vehicles	Platoon size	Empirical analysis
Milanés & Shladover (2014)	Infiniti M56	2 & 4	Speed/acceleration profile, time gap, string stability
Knoop et al. (2019)	4 makers (BMW, Mercedes, Audi, Tesla)	7	ACC usage rate, platoon stability, fuel consumption
Gunter et al. (2020)	1 model	2	OVM calibration, string stability
Makridis et al. (2020a)	1 model	2	Response time, time gap, IDM calibration
Makridis et al. (2020b)	Various makers and models	5	Response time, time headway, string stability
Gunter et al. (2021)	2 makers, 7 models	2 & 8	String stability
Makridis et al. (2021)	17 makers, 27 models	2,5,10	Time headway, string stability, traffic hysteresis
Li et al. (2021)	3 models	3	Response time, oscillation analysis

by onboard sensors (long range Lidar) where the radius of the detection range is 75 meters or more [14]. Subsequently, the trajectories of vehicles and other road users (cyclists and pedestrians) can be obtained by continuous multi-object tracking algorithm based on Lidar points. Compared to aerial photography based and GPS based data acquisition systems, the onboard Lidar trajectory collection method is advantageous considering its excellent scalability, higher accuracy and less penetration issue because of the larger detection range. These high-resolution and large-scale data collected under various traffic conditions provide great opportunities to evaluate the AV's behavioral impact on traffic flow.

In the literature, most studies focused on the safety performance of an individual AV while more or less ignoring AV's impact on the overall traffic flow efficiency. Since AV is likely to co-exist with other vehicle types in the foreseeable future, safety and efficiency of mixed traffic are two crucial but conflicting goals which need to be evaluated simultaneously. Moreover, current efforts on assessing the impact of AV on the whole traffic rely on oversimplistic simulations with strong assumptions regarding AV features and how AV and HV would interact with each other. Therefore, by exploiting the newly released Waymo Open Dataset the aim of this study is twofold. First, as one of the first studies in the literature, we provide empirical evidences on the CF behaviors of AV and its impact on surrounding vehicles both in terms of traffic safety using surrogate safety measures and efficiency using various measures (e.g., time headways, jam spacings, responses, etc.). Second, this study answers a critical question: if AV's impacts on CF behavior is significant, can such impacts be captured by some of the parameters of existing CF models? Before we

TABLE II
PROPORTION OF DANGEROUS TTC AND DRAC

Thresholds	AV-HV	HV-AV	HV-HV
TTC: 1s	0.00%	1.68%	0.11%
TTC: 2s	0.00%	10.41%	3.42%
TTC: 4s	33.08%	49.90%	39.27%
DRAC: $2m/s^2$	1.64%	3.40%	2.55%
DRAC: $3.35m/s^2$	0.026%	0.28%	0.26%
DRAC: $4m/s^2$	0.00%	0.04%	0.04%

can answer this question, one additional question needs to be addressed, that is, will the short-length trajectories (i.e., 20 s) in the Waymo Open Dataset be suitable for calibrating CF models?

To this end, the remainder of this paper is organized as the flowchart shown in Fig. 1. Specifically, the next section introduces the paired CF trajectories utilized in this study. Section III is the empirical analysis consisting of four parts: part A is the safety evaluation, part B/C/D are about the efficiency evaluation including time headway, velocity-spacing relationship and response time. Then in Section IV CF behavior modelling using Intelligent Driver Model (IDM) is first validated with short synthetic trajectories and then implemented on the Waymo's trajectory data, followed by statistical tests on the calibration results and stability analysis. Conclusion is presented in Section V.

II. PAIRED CAR FOLLOWING TRAJECTORIES

The Waymo Open Dataset (<https://waymo.com/open/>) consists of two parts: (a) Perception Dataset for 3D object detection and tracking; and (b) Motion Dataset for

motion/interaction prediction. Exploration on these two datasets show that the AV's trajectory covers all the frames recorded in each segment in Perception Dataset. However, in many segments from Motion Dataset, the AV's trajectories were involved in complex interactions (lane changing, merging, unprotected left turn, etc.) or covering only part of the frames. In this paper, only the Perception Dataset is utilized. The Waymo's Open Dataset refers to the Perception Dataset hereafter.

The Perception Dataset includes large-scale and high-resolution sensor data collected by Waymo autonomous vehicles in multiple cities in US (i.e., San Francisco, Phoenix, and Mountain View). A total of 1000 segments (scenarios) were originally released in 2019, and this number is continuously growing. The driving conditions covered in this dataset is diverse in terms of road types (urban streets, freeways), weather (sunny, rain), and time of day (dawn, day, dusk, night). The distributions of driving segment environment features (namely time of day, weather and segment road type) are presented in Fig. 2. Obviously, in most segments the drivers were in conditions of daytime, sunny weather and urban streets. Only 0.7% segments are in rainy days and only 3.5% segments were on freeways. In general, the driving environments are quite homogeneous.

The sensor data were collected by 5 Lidars (1 mid-range and 4 short-range) and 5 cameras (front and sides), where the Lidars and cameras were calibrated and synchronized. In addition, a large number of 3D ground truth bounding boxes (labels) for Lidar data was manually annotated by Waymo for the purpose of object tracking. This dataset can be very valuable for the research community because: (a) the data volume is large; (b) the time resolution is high, i.e., 0.1 seconds; and (c) the data quality is high and is better than NGSIM dataset [16]. The authors have previously processed, assessed, and enhanced the Waymo Open Dataset for driving behavior research [16]. The processed dataset has also been shared with the public (<https://data.mendeley.com/datasets/wfn2c3437n/2>).

In the stage of preliminary data processing, the original dataset is re-structured and transformed to a user-friendly tabular format trajectory data with 25 essential attributes, including the segment environment information (time of day, weather, etc.), object features (object type, length, etc.) and object tracking trajectory (position, speed, heading, etc.). Camera videos and trajectory view animations are generated for qualitative verification. Then the CF pairs are extracted manually by recording the IDs of the leader and the follower based on trajectory view videos. Great effort has been dedicated to ensuring that each paired CF trajectory is in a proper CF state. The impact of large vehicles (i.e., bus and heavy truck), lane changing, and traffic signals or stops signs has been excluded. To investigate AV CF behaviors, the paired trajectories are classified into 3 groups: an AV follows an HV (AV-HV), an HV follows an AV (HV-AV), and an HV follows an HV (HV-HV). Unfortunately, there is no AV-AV pair since there is at most one AV in one segment. The sample size for each type of CF group is 196 for AV-HV, 274 for HV-AV, 1032 for HV-HV. Other influencing factors (such as time of day, road types and weather) in car following behavior are not considered in

TABLE III
MEAN TIME HEADWAY UNDER DIFFERENT SPEED RANGES

Speed range (m/s)	AV-HV	HV-AV	HV-HV
0.5-5.5	5.35	4.01	4.30
5.5-10.5	4.04	2.93	2.67
10.5-15.5	3.25	2.54	2.13
15.5-20.5	2.41	2.26	1.77

this paper since the sample size would be too small to draw statistically conclusions.

Comprehensive assessing of the data quality is then implemented on the extracted and paired trajectory data. consistency analysis shows that the dataset itself is not internally consistent, i.e., the differentiation of positions yields inconsistent speeds and accelerations. Jerk value analysis reveals that some proportion of anomalies still exist in the data. Moreover, a pattern recognition algorithm is adopted to assess the trajectory completeness [17]. Specifically, different driving regimes (e.g., following, acceleration, deceleration, etc.) in the trajectories are objectively and automatically identified based on Dynamic Time Warping and Bottom-Up algorithms. The related codes can be downloaded from this website (<http://www.connectedandautonomoustransport.com/reproducible-research.html>). Our analysis suggests that the trajectories in the Waymo Open Dataset are all incomplete.

Furthermore, the extracted trajectory data are enhanced by using an optimization-based outlier removal method and a wavelet denoising method [16]. The linear programming optimization model in the outlier removal method can be implemented efficiently and ensure that the resulted trajectory is outlier-free. A wavelet denoising method is applied on the data to filter out noise. Overall, the quality of the trajectory data utilized in this study is adequately high to support our empirical analysis.

III. EMPIRICAL ANALYSIS

A. Safety Evaluation

To evaluate AV's safety performance during car following using the Waymo Open Dataset, surrogate measures of safety (SMS) are used. SMS are microscopic traffic flow measures that can relate to crash risk, and often used in the road safety literature to diagnose traffic conflicts or near misses (situations in which two or more road users are sufficiently close in space and time for their trajectories to cross unless evasive action is taken) which occur more frequently than crashes—for a short period of time, rather than waiting for long time periods to accrue a large number of crashes. In terms of CF safety analysis, two widely used surrogate measures are Time to Collision (TTC) and Deceleration Required to Avoid Crash (DRAC). TTC is the time remaining until a collision will occur between two vehicles if the collision course and speed difference are maintained [18], as shown in Equation (1); thus, a smaller TTC value indicates a more dangerous scenario. And DRAC was defined by Cooper and Ferguson [19] as the minimum deceleration rate required by the following vehicle to come to a timely stop (or match the leading vehicle's speed)

and hence avoid a crash, as shown in Equation (2); thus, a higher DRAC value indicates a more dangerous CF scenario.

$$TTC = \begin{cases} \frac{d}{v_f - v_l}, & \text{if } v_f > v_l \\ \infty, & \text{otherwise} \end{cases} \quad (1)$$

$$DRAC = \begin{cases} \frac{(v_f - v_l)^2}{d}, & \text{if } v_f > v_l \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where v_f and v_l are the speeds of the follower and the leader vehicle, respectively; and d is the gap between the follower and the leader vehicle.

Since no SMS is perfect and each has its own limitations, to increase the reliability of our safety analysis, both TTC and DRAC are adopted in this study to compare the CF safety performances between the defined three CF groups. Since it is not meaningful to use TTC as a surrogate safety measure when the speed is zero, and TTC is not capable of identifying dangerous cases when speed is too small [20]. In this study trajectories with speed smaller than 1m/s are removed before the TTC analysis. The Empirical Cumulative Distribution Function (ECDF) for TTC and DRAC are presented in Fig. 3 and Fig. 4, respectively. By looking at Fig. 3, it is clear that AV following HV is the safest while HV following AV is the most dangerous, among the three groups. However, the pattern is not clear by visually checking Fig. 4. Therefore, to quantitatively evaluate the CF safety, the proportions of dangerous TTC and DRAC values for each group are calculated by using two commonly used thresholds in the road safety literature, i.e., 2 s as the threshold for TTC and 3.35 m/s^2 as the threshold for DRAC [21], [22]. Meanwhile, to test the sensitivity of the thresholds two more values are also analyzed for TTC (1s, 4s) and DRAC (2 m/s^2 , 4 m/s^2), respectively. The results are shown in Table II.

Table II clearly shows that among the three CF groups, the AV-HV group has the smallest proportion (0%) of dangerous TTC, while the HV-AV group has the largest proportion (nearly 10%) of dangerous TTC (threshold 2s). The same pattern can be found by comparing the dangerous DRAC proportions across three CF groups (0.026% for AV-HV group and 0.28% for HA-AV group, under threshold 3.35 m/s^2). If we consider the cases where both TTC and DRAC values are dangerous, the proportions are 0%, 0.28% and 0.19% for AV-HV, HV-AV and HV-HV group, respectively. Similar phenomenon can be observed from other thresholds. Moreover, K-S tests on the distributions of TTC and DRAC values show significant difference among CF groups with very small p-values. Therefore, we can conclude that regarding CF safety, AV's behavior is obviously the safest, indicating that AV is designed to be more conservative than HV in terms of safety. Interestingly, HV tends to behave more dangerously when following an AV, compared with when following another HV.

B. Time Headway

Time headway is an important parameter in car following because it is related to flow rate, safety, and can act as a bridge

connecting macroscopic traffic characteristics and microscopic driving behavior. Therefore, it would be interesting to compare the difference of time headway distribution between AV and human, as shown in Fig. 5. The time headways are fitted to the typical lognormal distribution, where the μ and σ values are: $\mu = 1.25$, $\sigma = 0.43$ for AV; $\mu = 0.99$, $\sigma = 0.49$ for human. Normally, one would expect a relative smaller variance or some simple patterns of time headway distribution for AV compared to human, since AV is of the same vehicle type and is controlled by algorithms while human drivers tend to be more unpredictable thus display larger heterogeneity in terms of desired time headway. For instance, the time headway distributions from OpenACC Dataset have two distinctive peaks which correspond to the maximum and minimum settings of the ACC controllers, respectively [6]. However, in Fig. 5 it is not obvious that the AV is following any constant desired headways. Moreover, the variance in the time headway of AV is comparable to that of human drivers, which indicates that the controller in Waymo is sophisticated enough to generate significant behavioral heterogeneity in the time headway.

Previous research demonstrated that time headway distribution varies among different speed ranges [23]. In this study, we also adopt four speed ranges (i.e., 0.5-5.5, 5.5-10.5, 10.5-15.5, 15.5-20.5, unit:m/s) to comprehensively compare the time headway distributions among the three CF groups, as is shown in Fig. 6. In general, the time headway increases as the speed decreases, which is similar to the results in [23]. Moreover, regardless of speed range, the time headway of the AV-HV group tends to be larger than those of the HV-AV and HV-HV group, although this discrepancy becomes smaller as the speed ranges increase. Additionally, the difference of time headway distribution is not significant between HV-AV group and HV-HV group. Quantitatively, the mean time headways under the selected speed ranges are shown in Table III, and similar phenomenon can be found consistently. Overall, we can conclude that the time headway of AV-HV CF group is larger than that of the other two groups, indicating that the introduction of current AV will bring negative influence on traffic flow efficiency because of the reciprocal relationship between time headway and traffic flow rate.

C. Velocity-Spacing Curve

The velocity-spacing (v-s) relationship is essential to many widely used car following models, e.g., it is explicitly used in Optimal velocity Model [24], General Force Model [25], Full Velocity Difference Model [26]; and implicitly used in Newell's simplified car following models [27], and IDM [28]. To construct v-s curve would require homogeneous equilibrium traffic flow condition, which is not possible in our study due to the short length of the trajectories and limited sample sizes. Nevertheless, even with traffic data under inhomogeneous condition the comparison of v-s curve between the predefined three CF groups would still be meaningful in light of its vital role in delineating car following behavior.

The shape of v-s relationship in this study is assumed to be a piecewise linear curve composed of a free flow part and a congested part. As pointed out by Newell [27], it is certainly

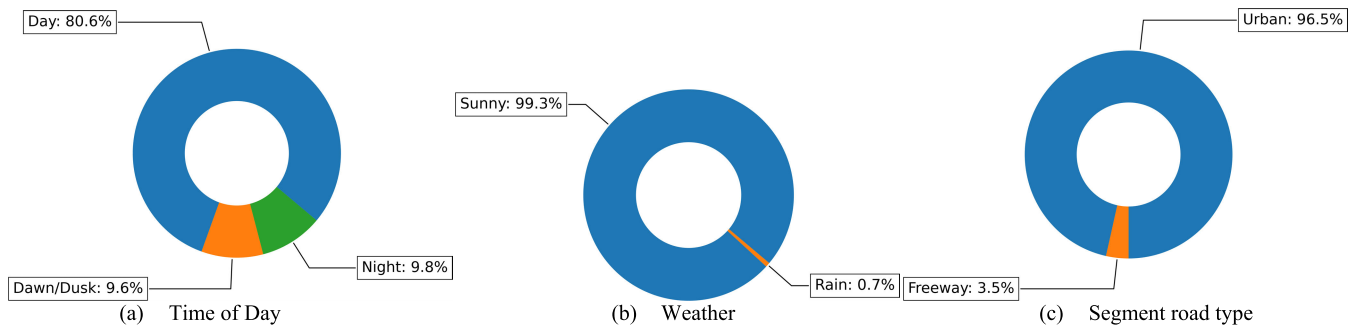


Fig. 2. Driving segment environment features of the Waymo Open Dataset.

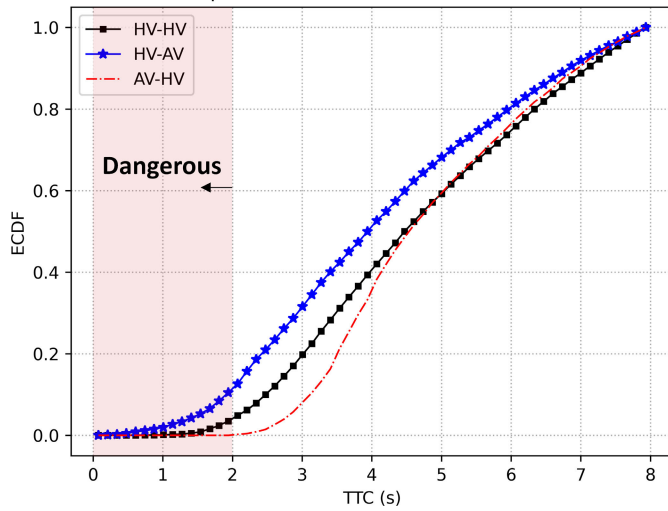


Fig. 3. TTC empirical cumulative distribution function.

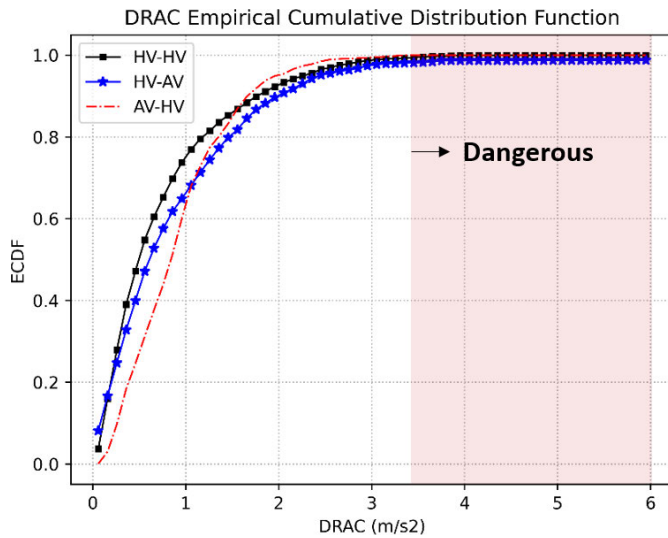


Fig. 4. DRAC empirical cumulative distribution function.

not obvious to assume a non-linear v - s relation according to the empirical observations in previous research. On the other hand, with a piecewise curve the critical traffic state between free flow and congested flow can be pinpointed. Under such hypothesis, the piecewise curve fitting problem would be an optimization problem in which the number of line segments is

two and the breakpoint is unknown. A recently released Python library (<https://pypi.org/project/pwlf/>) called *pwlf* is adopted to solve this problem, in which the users can conveniently fit a piecewise linear curve with or without known breakpoint locations, fit for a specified number of line segments, and force a fit through data points and so on [29]. Additionally, in order to obtain a robust optimization solution and make curve fitting results more comparable, we add the following three constraints:

- The slope of free flow part is 0
- The free flow speed v_f is set to be the same for the three CF groups
- The congested part will go through the average jam spacings s_j calculated from the empirical data

The first and the second constraint is a simplification on the free flow part in light of the complexities that arise from various road speed limits and maximum desired speed from different drivers. Since in this Python library users cannot specify a constraint on the slope of a line segment, we work around by forcing the curve pass through (S_1, v_f) and (S_2, v_f) , where S_1 and S_2 are large values of spacing to prevent inaccurate breakpoint (critical state). Meanwhile, we also force the fit pass through the point $(s_j, 0)$. Specifically, the values of S_1 , S_2 , and v_f are 200m, 300m, and 15m/s, respectively. And the s_j values are 10.0m for AV-HV group, 6.73m for HV-AV group, 7.88m for HV-HV group. The global optimization algorithm used in this library is efficient enough to solve our problem within seconds.

The v - s scatter plots and the curve fitting results are presented in Fig. 7. The primitive v - s curve for each CF pair is directly plotted so that some extreme cases can be clearly demonstrated. Compared to averaging speed over spacings or averaging spacing over speeds analysis methods, the curve fitting method is able to avoid bias near the standstill regime and critical point. Considerable dispersion is observed in the v - s plots of each group. By analyzing empirical HV CF data, Jiang et al. [30] attributed such phenomenon to two driving behaviors: (a) In certain range of spacing, the drivers are not sensitive to the changes in spacing if the velocity differences are small; (b) During the driving process, the driver change their preferred spacing intentionally or unintentionally resulting in unfixed spacing at given velocity. The feature of the scatter plot in Fig. 7(a) implies that the AV might behave similarly. The parameters of the curve fitting are given in

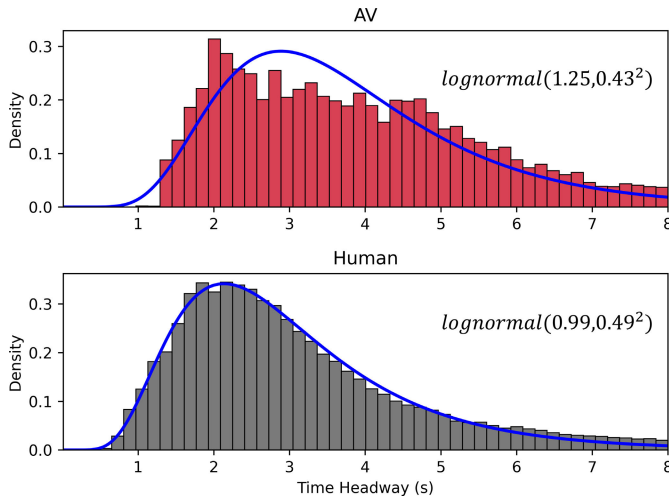


Fig. 5. Comparison of time headway distribution between AV and human.

Table IV. The AV-HV group has the largest jam spacing and critical spacing (the intersection between the free flow branch and the congested branch, i.e., the turning point), the smallest slope for the congested branch, consistently indicating that the AV is more conservative while following a leader, as depicted in Fig. 7(d). Compared to the HV-HV group, the HV-AV group shows similar critical spacing, but a smaller congested branch slope and larger jam spacing.

D. Response Time

Driver response time is an important measure that is closely related to traffic safety such as rear-end collision [31], dilemma zones at signalized intersections [32], cut-off scenario [33] and take-over control of an automated vehicle [34]. Meanwhile, driver response time is also a contributing factor in traffic flow efficiency (i.e., discharge rate) [35].

Due to the inconsistencies of terminology and definitions in the literature, the theoretically and practically justified definition of driver response time proposed by Sharma et al. [36] is adopted in this study: Response time is defined as “the time taken by a driver to adjust his/her speed against a stimulus, with or without deliberately delaying his/her decision.” In this definition, the response time is divided into two parts: (a) Latent Response Time (LRT) which includes reaction time (sensation, perception, decision, initiation) and delay (intentionally produced by the driver) and (b) Observable Response Time (ORT) which is the time between when the foot movement starts and when it ends.

1) *Methodology*: Regarding the methodologies for estimating response time, Sharma et al. [36] provided a thorough review and demonstrated that wavelet-based energy distribution (WED) method is advantageous to existing methods. Suppose the speed time series is a continuous function $v(t)$, and the selected wavelet function is $\psi(t)$, then the wavelet transform coefficients can be formulated as:

$$T(s, x) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} v(t) \psi\left(\frac{t-x}{s}\right) dt \quad (3)$$

TABLE IV
VELOCITY-SPACING CURVE FITTED RESULTS

	AV-HV	HV-AV	HV-HV
Jam spacing (m)	10.0	6.73	7.88
Congested branch slope (s^{-1})	0.44	0.57	0.66
Critical spacing (m)	43.9	33.1	30.5

TABLE V
MEAN RESPONSE TIME FOR DIFFERENT TYPES OF DRIVING STATE CHANGE

State change type	AV-HV (t_0)	HV-AV (t_1)	HV-HV (t_2)
AF	2.51 (25)	1.88 (45)	2.05 (191)
AD	2.59 (22)	1.90 (27)	1.94 (174)
SA	1.62 (52)	1.48 (53)	1.34 (253)
FD	2.13 (20)	1.97 (38)	1.79 (128)
DS	3.59 (36)	2.03 (52)	2.06 (164)

where s and x are the scale and translation parameter, respectively, $T(s, x)$ is the corresponding wavelet coefficient at scale s and location x . Subsequently, the wavelet-based energy at x can be expressed as:

$$E_x = \frac{1}{\max(s)} \int_0^{+\infty} |T(s, x)|^2 ds \quad (4)$$

The principle of WED method is that an abrupt change in speed profile will produce a peak in the temporal wavelet energy distribution profile, where the driving state change points normally locate.

Despite the capabilities of WED method in detecting singularities, one problem Sharma et al. [36] did not emphasize enough is the boundary effect of wavelet transform. When computing wavelet decomposition coefficients, boundary effect will be introduced due to finite size of the signal length, as pointed out by Zheng et al. [37], which is characterized by large wavelet transform coefficients at both ends of the signal range. This occurs because the speed outside the signal range is assumed to be zero. Large wavelet transform coefficients are obtained at the boundaries where the signal shifts from zero to an actual speed value, causing inaccurate detection of singularities (driving state change points in our case) near the boundary. Normally the wavelet coefficients at the boundary are discarded due to this effect, and this is not a problem if the data length is long enough, which is the case (the trajectory length is 180s) in [36]. Yet in Waymo's dataset such information loss is unacceptable since the length of the trajectories is only 20s. Hence, in this study signal extension is implemented on every trajectory by adding 100 trajectory points on each side of the boundary. Note that such extension of the signal at each end will not distort WT analysis of the original signal since the wavelet coefficients from the extended portions are discarded. The acceleration values of the extended trajectories are constant, which are equal to the corresponding real acceleration at the boundaries. In this way, the driving state change points near the boundary can also be accurately identified while no new singularities are introduced.

Moreover, while implementing WED method, the selection of maximum decomposition scale can lead to early or late

TABLE VI
T TEST P VALUES AMONG DIFFERENT TYPES OF STIMULI

	AF	AD	SA	FD	DS
AF	-	-	-	-	-
AD	0.328	-	-	-	-
SA	<0.001	<0.001	-	-	-
FD	0.005	0.072	<0.001	-	-
DS	0.026	0.005	<0.001	<0.001	-

detection of driving state change point [37]. To remedy this issue, wavelet-based Local Maxima Lines (LML) method is adopted in this study. Specifically, the modulus maxima is defined as any point (s_0, x_0) such that $|T(s_0, x)| < |T(s_0, x_0)|$ when x belongs to either a right or the left neighborhood of x_0 , and $|T(s_0, x)| \leq |T(s_0, x_0)|$ when x belongs to the other side of the neighborhood of x_0 [38]. And a local maxima line is formed by connecting all nearest modulus maxima points in the scale space (s, x) . Then the accurate location of a state change point can be found using the corresponding local maxima line at the finest scale [39]. LML method is more precise than WED method because the original signal is scrutinized at all decomposition scales, and is less sensitive to the designated maximum decomposition scale.

The settings of wavelet-based LML method designed in this study are specified as below. First, the wavelet function adopted here is Mexican hat, as suggested in Zheng and Washington [39] and Sharma et al. [36]. Second, with regard to the maximum decomposition scale, we have conducted sensitivity analysis using synthetic trajectory data (thus the ground truth location of the state change point can be obtained). Results show that scales between 10 to 50 are trustworthy for both WED and LML methods. In our study, this number is set to be 32. Third, the scale to calculate response time is set to be 15, which is the middle scale under the selected maximum decomposition scale, to avoid too sensitive (insensitive) detections at small (large) scales.

2) *Empirical Results*: As mentioned above, the response time is defined as the time taken by a driver to adjust his/her speed against a stimulus, which needs to be classified in detail. In this study, the type of driving state change point is adopted to distinguish different stimuli. In car following regimes, the driving states can be divided into 4 types: following the leader at a constant speed (F), accelerating behind a leader (A), decelerating behind a leader (D), and standing behind a leader (S) [17], [40]. Accordingly, 5 types of stimulus points are generated: AF, AD, SA, FD, DS. Note that the FS is excluded since the vehicle is not able to change its state directly (deceleration or acceleration is required) from following to standstill, and vice versa. Using the aforementioned estimation methodology, for each CF group the average response time of each type of stimulus is given in Table V (numbers in parentheses are sample sizes). Across all the stimulus types, the average response times of AV-HV, HV-AV, and HV-HV groups are 2.42s, 1.84s, 1.80s, respectively.

Further analysis on response time is focused on three questions:

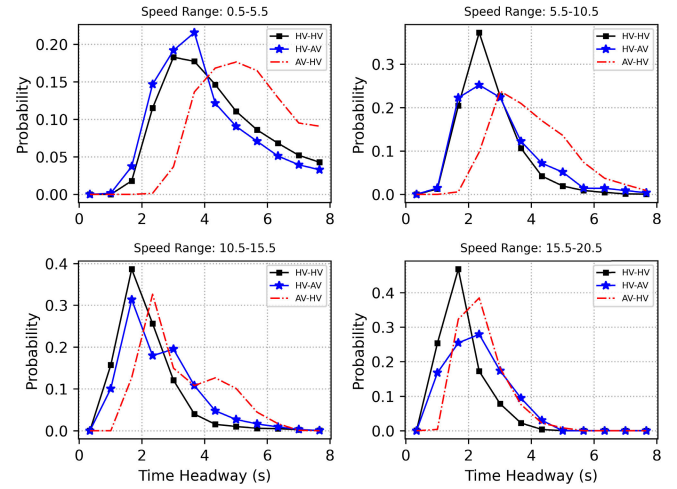


Fig. 6. Speed dependent time headway distribution.

TABLE VII
T TEST (WMW TEST) P-VALUES FOR RESPONSE TIME COMPARISON AMONG CF GROUPS

State change type	CF group comparison	Alternative hypothesis	p-value
AF	AV-HV vs HV-HV	$t_0 > t_2$	<0.001 (<0.001)
	HV-AV vs HV-HV	$t_2 > t_1$	0.084 (0.047)
AD	AV-HV vs HV-HV	$t_0 > t_2$	0.003 (<0.001)
	HV-AV vs HV-HV	$t_2 > t_1$	0.417 (0.319)
SA	AV-HV vs HV-HV	$t_0 > t_2$	<0.001 (<0.001)
	HV-AV vs HV-HV	$t_1 > t_2$	0.041 (0.020)
FD	AV-HV vs HV-HV	$t_0 > t_2$	0.060 (0.066)
	HV-AV vs HV-HV	$t_1 > t_2$	0.125 (0.146)
DS	AV-HV vs HV-HV	$t_0 > t_2$	<0.001 (<0.001)
	HV-AV vs HV-HV	$t_2 > t_1$	0.454 (0.211)

- Is the response time significant different among various stimulus points?
- Is the response time of AV significantly different from HV (AV-HV vs HV-HV)?
- For HV, is the response time of following an AV significantly different from following an HV (HV-AV vs HV-HV)?

For the first question, our finding is that the difference of response times between different types of stimuli should not be ignored. For instance, the response time of SA stimuli is apparently smaller than other types of stimuli. To further validate this observation, unpaired t-tests are conducted and the pair-wise p-values are shown in Table VI. Apart from the AD-AF pair and FD-AD pair, all other p-values are less than 0.05, indicating that the response times among all types of stimuli are significantly different. Thus, it is necessary to treat each type of response time separately.

For the second and third question, p-values of statistical tests are presented in Table VII. Note that our samples are mostly not in normal distribution and many researchers would adopt nonparametric test such as Wilcoxon-Mann-Whitney (WMW) test rather than parametric test such as t test. However, according to the research by Lumley et al. [41], Fagerland [42] and Skovlund and Fenstad [43], the following rules are more preferable: (a) if the sample size is modestly large enough, t test should always be adopted, regardless of the distribution;

(b) if the sample size is small and the distribution is long tailed and skewed, WMW test should be used. In our study, for most samples the sample sizes are modestly large, so t test is more desirable although the distributions are not normal. Nevertheless, the p-values of both t test and WMW test are provided to make the analysis more holistic and rigorous.

In Table VII, the t test p-values are directly given and WMW test p-values are given in parentheses. Overall, the results from t test and WMW test are quite consistent. For the second question, we can conclude that the response time of AV is significantly (marginally significant in the case of FD) different from HV. Specifically, the response time of AV under all types of stimuli is consistently larger. For the third question, when the follower is an HV, the response time difference between following an AV and an HV is not significant (almost marginally significant in the case of SA).

According to the aforementioned definition of response time which is composed of LRT and ORT, the ORT time of AV would be smaller than HV since there is no foot movement. In terms of LRT which includes reaction time and delay time, the reaction time (sensation, perception, decision, initiation) of AV is similar to HV [7]. Since the overall response time of AV is larger, we can deduce that the delay time of AV is larger than HV. In other words, AV is deliberately delaying its action against a stimulus. This is reasonable considering the conservative behavior of AV.

Some previous study on ACC and CACC systems revealed that the observable response times of ACC system range from 1.3s to 2.5s and the corresponding values for CACC are 0.1s to 0.5s [44]. Compared to the average response times of AV (2.4s) and HV (1.8s) in this study, it is found that: (a) CACC system has undoubtedly the smallest response time; (b) in general the ACC system and HV have comparable response times; and (c) the AV has the largest response time. In conclusion, among these four types of vehicles, AV is the most conservative.

IV. CF BEHAVIOR MODELLING

In this section, we use the IDM car following model to model car following behavior of each vehicle group in the Waymo Open Dataset. IDM is a widely used car following model which is known for its capability to replicate many real-world traffic flow phenomena and its straightforward interpretation of each model parameter [28], [45]. Previous studies also adopted the IDM model to simulate the behavior of CACC vehicle and AV [2], [46]. Although IDM might not be able to precisely replicate the behavior of both HV and AV, it is a reasonable tool for us to detect any significant differences in AV's and HV's car-following behavior, given that: (a) IDM has been frequently used in the literature to model both HV car-following and AV car-following; and (b) in our study, results from IDM calibrations are only used as a supplement to our empirical analysis. A common formulation of IDM is:

$$a_n(t) = a \left[1 - \left(\frac{v_n(t)}{v_0} \right)^4 - \left(\frac{s_n^*(t)}{s_n(t)} \right)^2 \right] \quad (5)$$

$$s_n^*(t) = s_0 + T v_n(t) - \frac{v_n(t) \Delta v_n(t)}{2\sqrt{ab}} \quad (6)$$

where a is the maximum acceleration (m/s^2), v_0 is the desired speed (m/s), s_0 is the minimum gap (m), T is the desired time gap (s), b is the comfortable deceleration (m/s^2), $s_n^*(t)$ is the desired gap, $a_n(t)$, $v_n(t)$, $s_n(t)$, $s_n^*(t)$ are the follower's acceleration, speed, spacing (vehicle length not included) and desired spacing, respectively, $\Delta v_n(t)$ is the speed difference (the leader's speed minus the follower's speed).

The impact of trajectory incompleteness on CF model calibration error has been investigated in [40] and [47]. However, both studies used very long trajectories (large than 180s) and the most incomplete trajectory still included acceleration, deceleration and following (AFD). In our case, the driving regimes could consist of only two states or only one state in extreme situations. With such short trajectories, it is unlikely to expect all the calibrated IDM parameters are reliable simultaneously. Nevertheless, whether part of the IDM model parameters can be calibrated under certain corresponding driving situations is still an open question. If this is true, calibration results from the Waymo's trajectory data can still be used since our purpose is to compare the difference of CF behavior, not to precisely replicate the behavior. Thus, before implementation on Waymo's dataset it is imperative to verify the effectiveness of calibrating IDM with rather short trajectories (20s in Waymo's case).

A. Validation of CF Model Calibration With Short Synthetic Trajectories

In order to investigate the impact of short trajectories on IDM calibration, synthetic trajectories generated with ground truth model parameters have been used. In principle, efforts have been made to force the synthetic trajectories to be as similar as the actual Waymo's data. The following rules are designed to generate the paired synthetic trajectory data:

- The length and resolution of the synthetic trajectory is the same as that in Waymo's data (20s, 0.1s)
- 8 types of driving state combinations are considered: F, AF, FD, DS, AFS, AFD, FDS, AFDS, which are based on the analysis in Hu et al. [16]
- The leader trajectories are directly obtained from the actual Waymo's paired trajectory data. For each type of driving regime, 10 typical paired trajectories are selected
- For each chosen leader trajectory, generate 10 set of model parameters and their corresponding follower trajectories
- Each generated follower trajectory has the same initial state as the actual follower

Examples of the generated synthetic trajectories and the corresponding IDM parameter values are shown in Fig. 8. As can be observed, these trajectories are rather similar to the actual Waymo's trajectories which have been shared in Hu et al. [16]. If the model parameters (at least some of the parameters) of these trajectories can be reproduced by calibration, then it is also valid to use the trajectories in Waymo Open Dataset to calibrate IDM model.

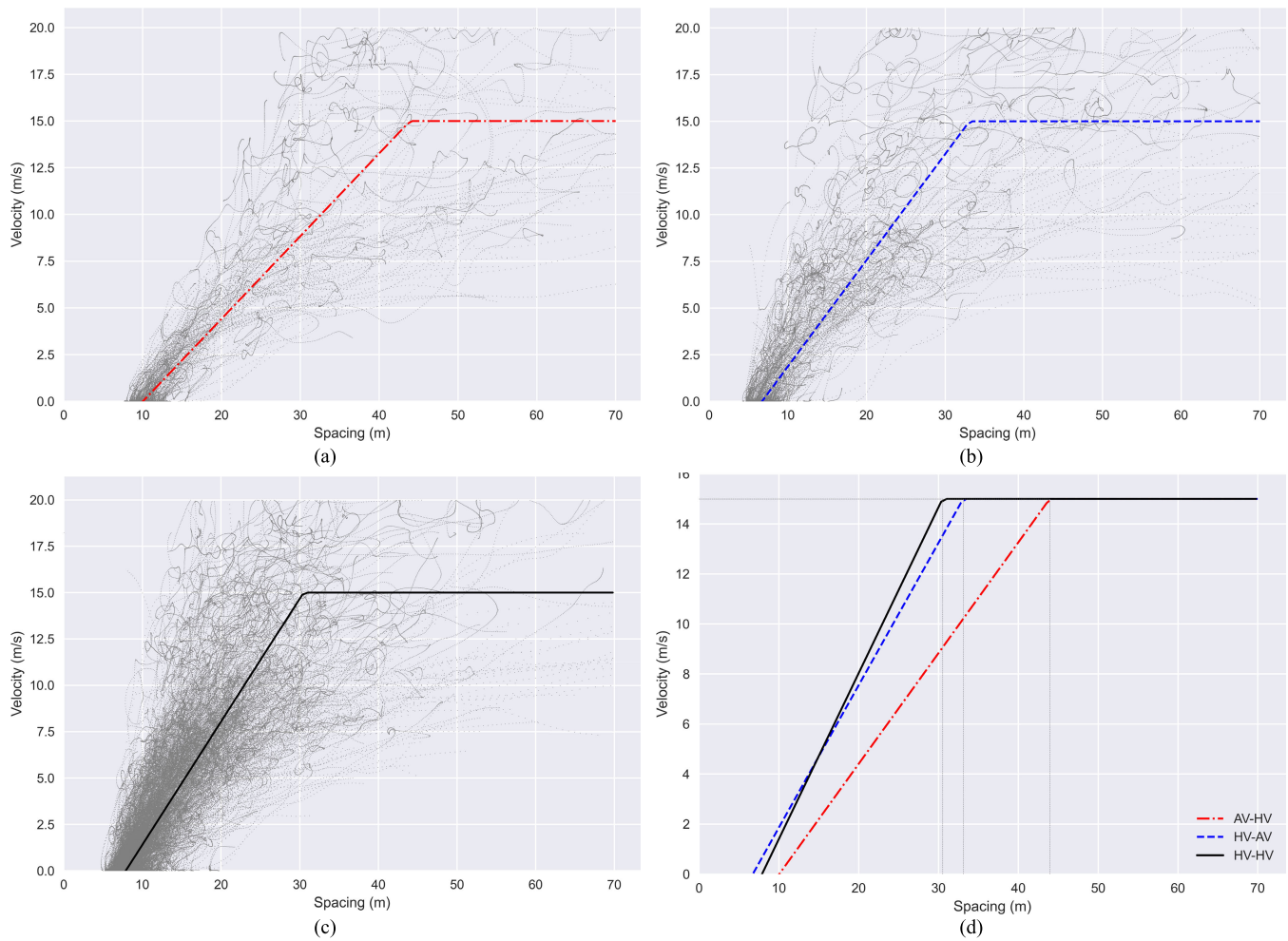


Fig. 7. Velocity-Spacing relationship scatter plots and curve fitting results: (a) AV-HV group; (b) HV-AV group; (c) HV-HV group; (d) Comparison.

TABLE VIII
CALIBRATION ERROR WITH SYNTHETIC TRAJECTORIES

Driving regime	Error type	v_0	T	s_0	a	b
F	MAE	0.37	0.17	1.33	0.20	0.29
	MAPE	2.1%	11.0%	100.9%	10.4%	14.6%
AF	MAE	0.05	0.04	0.23	0.04	0.07
	MAPE	0.3%	2.6%	16.7%	1.7%	2.8%
FD	MAE	0.01	0.02	0.02	0.02	0.02
	MAPE	0.1%	0.7%	0.8%	1.0%	0.8%
DS	MAE	0.00	0.02	0.03	0.20	0.07
	MAPE	0.0%	2.6%	1.5%	13.6%	3.8%
AFS	MAE	0.22	0.04	0.00	0.07	0.18
	MAPE	2.2%	3.1%	0.3%	3.9%	10.8%
AFD	MAE	0.02	0.02	0.07	0.04	0.02
	MAPE	0.1%	1.5%	7.0%	3.0%	0.9%
FDS	MAE	1.41	0.01	0.03	0.13	0.12
	MAPE	14.1%	1.2%	1.8%	6.8%	6.8%
AFDS	MAE	2.01	0.01	0.01	0.02	0.08
	MAPE	20.1%	0.9%	0.5%	1.1%	3.8%

By largely following the guidelines in [48], the calibration settings in this study are described below:

- Global calibration is adopted in which the whole original trajectory is compared with the simulated trajectory rather than each data point.

- The measure of performance is spacing instead of speed or acceleration.
- The goodness of fit function is the RMSE of spacing. Although this is not the optimal choice according to [48], the results from this function are reasonable with a small computational cost.
- Interior point method is selected to solve the optimization problem.
- The ranges of parameters are set as: v_0 [10], [30], T [0.1]3, s_0 [0.1]10, a [0.5]5, b [0.5]5.

The errors (Mean Absolute Error, MAE, and Mean Absolute Percentage Error, MAPE) between the calibrated parameters and the ground truth values with synthetic trajectories are presented in Table VIII. Surprisingly, in most cases the overall calibration errors are not large even with 20s trajectories. It is obvious that if only F regime is present, the calibration results are not trust worthy. Also, in some cases, certain parameters are not precisely calibrated (MAPEs larger than 10% are in bold): s_0 in AF, a in DS, b in AFS, v_0 in FDS and AFDS. However, for all driving regime combinations (except F in which T is still extracted reasonably well), the calibration results of desired time gap T (the principal parameter in IDM [16], [49]) are always reliable. Additionally, if standstill regime is present,

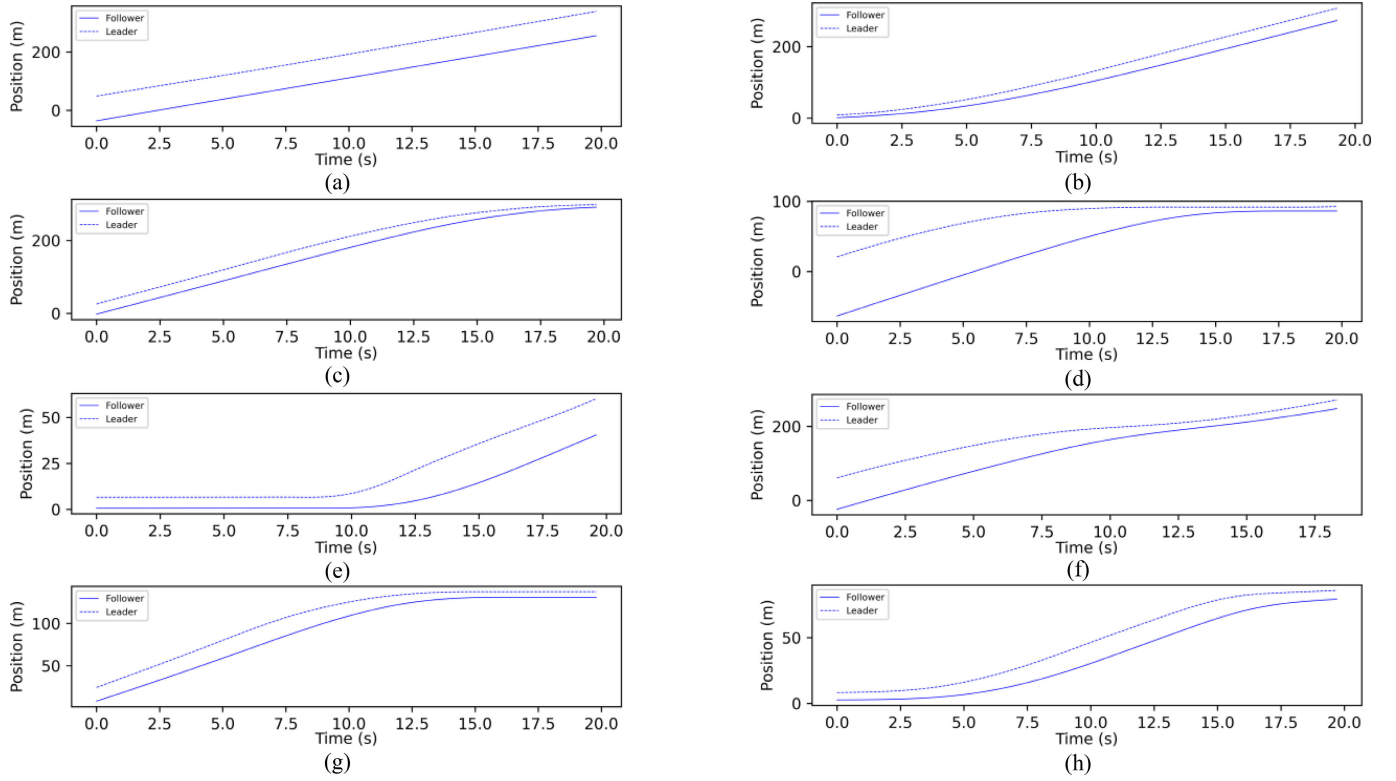


Fig. 8. Examples of the generated synthetic trajectories and the corresponding IDM parameter values (v_0, T, s_0, a, b): (a) **F** (15, 1.8, 2.5, 2.7, 2.1); (b) **AF** (19, 0.6, 1.4, 1.8, 2.8); (c) **FD** (19, 0.6, 0.8, 2.6, 1.6); (d) **DS** (12, 0.8, 1.5, 1.2, 1.7); (e) **AFS** (10, 2.0, 1.8, 1.1, 2.1); (f) **AFD** (21, 1.1, 2.3, 1.9, 2.5); (g) **FDS** (12, 0.8, 2.0, 2.8, 2.5); (h) **AFDS** (10, 1.2, 0.9, 1.4, 2.0).

the accuracy of s_0 will generally be improved. In conclusion, this outcome verifies that it is feasible to compare the CF behavior of each vehicle group by calibrating IDM using the trajectories in Waymo Open Dataset.

B. CF Model Calibration With the Waymo's Dataset

This subsection compares the CF behavior difference among the three CF groups by calibrating the IDM model using the trajectories in the Waymo Open Dataset. Each model parameter is analyzed separately, where v_0 is excluded since it might be influenced by the road speed limits and thus not a good indicator to distinguish CF behavior. Trajectory data from 196 AV-HV pairs, 274 HV-AV pairs and 1032 HV-HV pairs are used in calibrating IDM separately. However, based on the aforementioned findings from using synthetic trajectories results from the pairs with the only F regime are excluded; and minimum gap s_0 results from those with the AF regimes and maximum acceleration a results from those with the DS regimes, and comfortable deceleration b results from those with the AFS regimes are discarded as well. In addition, to avoid bias, boundary values from the calibration results (a common issue in CF model calibration) are also removed from the comparison analysis.

The mean values of IDM parameters with their corresponding sample sizes are shown in Table IX. To quickly assess the accuracy of the calibration results, the value of minimum gap s_0 can be analyzed because of its clear physical meaning, i.e., the jam spacing with the vehicle length

	T	s_0	a	b
AV-HV (group 0)	1.55 (52)	5.22 (51)	1.87 (44)	1.02 (45)
HV-AV (group 1)	1.13 (63)	2.24 (63)	1.58 (50)	1.19 (52)
HV-HV (group 2)	1.17 (317)	3.28 (304)	1.84 (271)	1.29 (250)

TABLE X
T TEST (WMW TEST) P-VALUES FOR CALIBRATION RESULTS COMPARISON AMONG CF GROUPS

IDM parameter	CF group comparison	Alternative hypothesis	p-value
T	AV-HV vs HV-HV	$T_0 > T_2$	<0.001 (<0.001)
	HV-AV vs HV-HV	$T_2 > T_1$	0.291 (0.248)
	AV-HV vs HV-HV	$s_0^0 > s_0^2$	<0.001 (<0.001)
s_0	HV-AV vs HV-HV	$s_0^2 > s_0^1$	<0.001 (<0.001)
	AV-HV vs HV-HV	$a_2 > a_0$	0.444 (0.364)
a	HV-AV vs HV-HV	$a_2 > a_1$	0.029 (0.084)
	AV-HV vs HV-HV	$b_2 > b_0$	0.019 (0.452)
b	AV-HV vs HV-HV	$b_2 > b_0$	0.019 (0.452)
	HV-AV vs HV-HV	$b_1 > b_2$	0.278 (0.566)

AV-HV	HV-AV	HV-HV
5.8%	20.3%	16.6%

excluded. Recall that the average empirical jam spacings (vehicle length included) are shown in Table III. If we subtract the minimum gap s_0 from the jam spacings for each CF group, the corresponding values are 4.78m, 4.49m, 4.6m,

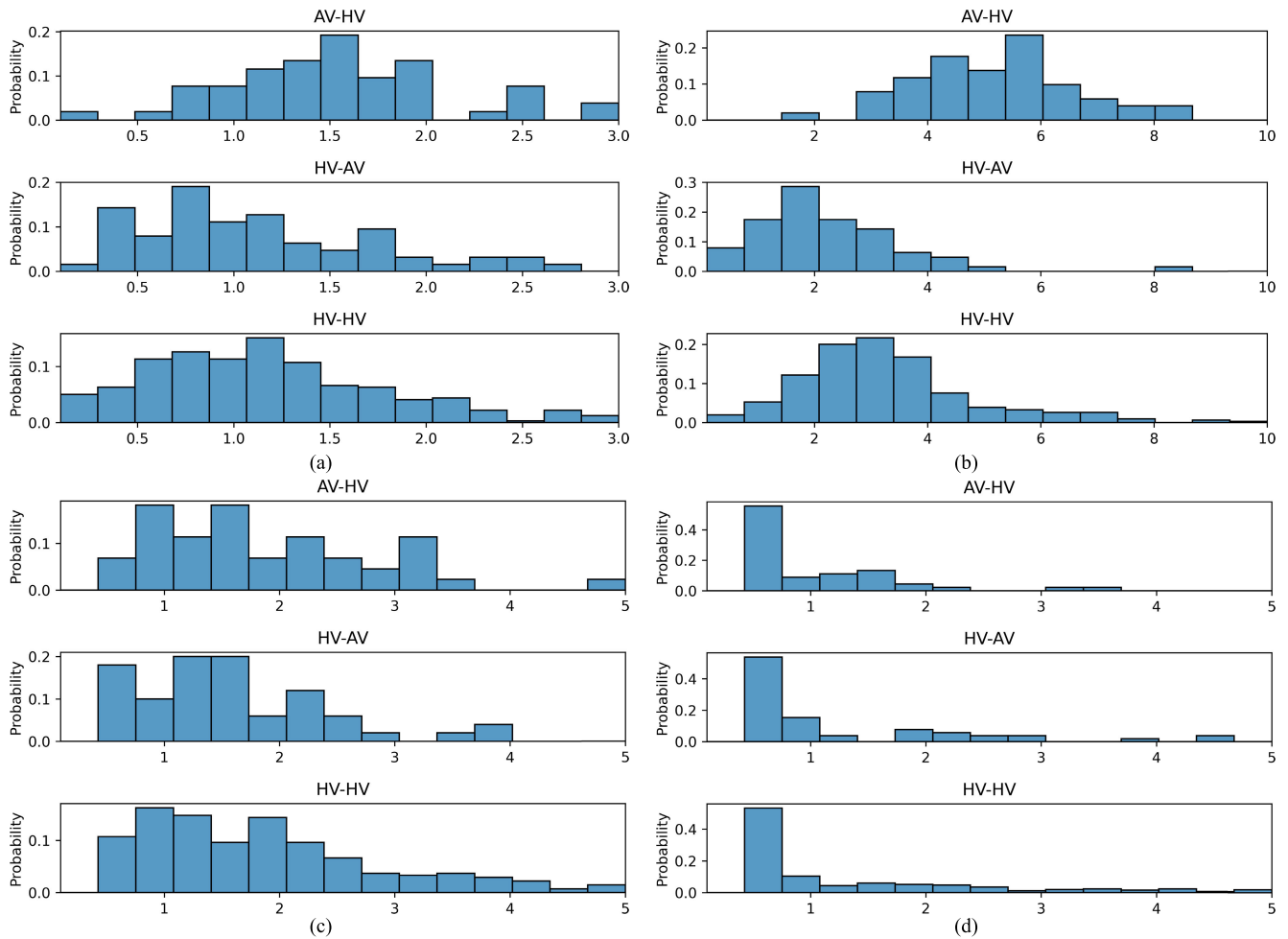


Fig. 9. Calibrated parameters distribution: (a) desired time gap T ; (b) minimum gap s_0 ; (c) maximum acceleration a ; (d) comfortable deceleration b .

which turns out to be very close to the typical average length of light-duty vehicles (4.5m), which indicates that our calibration results are accurate. In addition, the calibrated parameters from HV-HV group are compared with previous IDM calibration research on pure HV interactions [40], [45], [50], results also demonstrate that our calibrated parameters are in reasonable ranges.

OAs shown in Table IX, compared to the other two groups the AV-HV group has the largest mean desired time gap T and minimum gap s_0 . Compared to the HV-HV group, the HV-AV group shows similar calibration results except that the minimum gap s_0 is smaller. Meanwhile, the distribution for each model parameter is presented in Fig. 9. To draw a more rigorous conclusion, statistical tests are carried out using both unpaired t test and WMW test, and results are shown in Table X, where p values from WMW test are given in parenthesis. In most cases, the p-values of t test and WMW test are consistent (assume that the confidence level is 95%). However, with regard to the comfortable deceleration b , these two tests give contradictory conclusion when comparing the AV-HV and HV-HV group. we believe the result from WMW test should be preferred in this case because the sample size is not large and the distribution of b is long tailed and skewed in all CF groups.

From Fig. 9 and Table X, the difference of parameters maximum acceleration a and comfortable deceleration b are not significant or only marginally significant, either in comparing AV-HV and HV-HV group or comparing HV-AV and HV-HV group. Although there's no strict one to one matching between IDM parameters and their corresponding driving regimes [47], in general a and b are related to the acceleration and deceleration behavior. In this sense, we can roughly conclude that the difference of acceleration and deceleration characteristics in car following between AV and HV is not significant. In contrast, T and s_0 values of the AV-HV group are both significantly larger than those of the HV-HV group. Such results are consistent with the empirical analysis section in this paper. Compared to the HV-HV group, the HV-AV group has a significantly smaller s_0 value, but the difference in T values is not significant between these two groups. One possible explanation is that out of curiosity, the HV drivers might tend to keep a smaller jam spacing in order to observe the AV more closely.

Furthermore, once the model calibration results are obtained, CF string stability analysis can be carried out subsequently. String stability analysis investigates the evolution of small perturbations (i.e., speed and spacing deviation from the equilibrium state) in driving behavior over a platoon

of vehicles. If a perturbation amplifies (diminishes) along space, then the platoon is unstable (stable). In general, the empirical L_∞ instability analysis [51] should be implemented if the dataset includes a platoon of vehicles' trajectories with equilibrium states, perturbations and their propagations. In this study we do not have such trajectories and a stability analysis method based on calibrated IDM is adopted. Suppose a general form of CF model $a_n(t) = f(s_n, v_n, \Delta v_n)_t$, according to [52] the measure of string stability can be defined as:

$$S = \frac{1}{2} - \frac{f_{\Delta v}}{f_v} - \frac{f_s}{f_v^2} \quad (7)$$

where S is the value of stability measure, $f_s = \frac{\partial f}{\partial s}|_e$, $f_v = \frac{\partial f}{\partial v}|_e$, $f_{\Delta v} = \frac{\partial f}{\partial \Delta v}|_e$ are the Taylor expansion coefficients of the acceleration function at the steady state after first order linearization.

Under the equilibrium speed v_e and equilibrium spacing s_e , the coefficients of IDM are constructed as follows:

$$f_s = \frac{2a}{s_e} \left(\frac{s_0 + Tv_e}{s_e} \right)^2 \quad (8)$$

$$f_v = -a \left[\frac{4}{v_0} \left(\frac{v_e}{v_0} \right)^3 + \frac{2T(s_0 + Tv_e)}{s_e^2} \right] \quad (9)$$

$$f_{\Delta v} = \sqrt{\frac{a}{b}} \frac{v_e}{s_e} \frac{s_0 + Tv_e}{s_e} \quad (10)$$

Thus, for each pair of CF trajectories, a value of stability measure S can be calculated. A positive value of S indicates that the CF behavior is string stable. And a negative value of S can be interpreted as different severity levels of traffic oscillations as defined by Sun et al. [50].

The proportion of string unstable CF behavior for each CF group is presented in Table XI. Clearly, the AV-HV group has the least string unstable CF behavior, which mostly can be attributed to the larger desired time gap T . According to Sun et al. [50], a higher stability can alleviate the traffic oscillation (stop/slow and go waves) severity. In fact, previous studies also corroborated that a larger time gap is required to successfully dampen the traffic oscillation [53], [54], [55]. Therefore, this result indicates that the current AV technology can potentially mitigate traffic oscillations thanks to its conservative behavior.

V. CONCLUSION

By utilizing the Waymo Open Dataset this paper has completed two goals: a comprehensive and rigorous CF behavior comparison analysis between AV and HV; and capturing important behavioral differences between AV and HV using the IDM car-following model. For the first goal, AV's impact on traffic safety and efficiency is revealed through a four-part empirical analysis. First, two safety surrogate measures TTC and DRAC are adopted to evaluate the CF safety of different vehicle pairs. Second, the time headways under various speed ranges are analyzed. Third, the velocity-spacing relationship is approximated by using a piecewise linear curve fitting method. Fourth, the response times under different stimuli are estimated with a wavelet-based Local Maxima Lines method.

For the second goal, we have first confirmed the feasibility of using the short-length trajectories (i.e., 20 s) to calibrate IDM, using synthetic trajectory data, and then calibrated IDM using vehicle pairs contained in the Waymo Open Dataset. Based on the modelling result, AV's implication on string stability has also been assessed.

Regarding the safety evaluation, results from both TTC and DRAC consistently show that AV is often observed to have a larger time headway, jam spacing, critical spacing and response time, and thus is much safer than HV during the car-following process. However, AV's conservative CF behavior indicates negative influence on traffic flow efficiency, which is further confirmed by result from IDM modelling that the desired time gap T and the minimum gap s_0 of AV is significantly larger than HV. Further string stability analysis suggests that the proportion of string unstable CF behavior of AV is smaller than that of HV. This feature of AV can be potentially beneficial for mitigating traffic oscillation.

Meanwhile, results from the comparison analysis of HV-AV group and HV-HV group show that when HV is the following vehicle, there is no significant difference in its CF behavior regardless whether the lead vehicle is AV or not, with only one exception, i.e., the jam spacing is significantly smaller when the leader vehicle is an AV, which is likely caused by the curiosity of the HV driver.

Nevertheless, there are several open questions that need to be further investigated. First, note that this study only analyzes the case of a single AV rather than a platoon of AVs due to the limitation of the dataset. In such a case, it is not surprising to see that AV ensures its safety at the cost of efficiency since there is no coordination between AV and other vehicles. In the future, it is possible that a platoon of AVs with the cooperative driving capability is capable of improving their safety and efficiency simultaneously. Second, also note that the Waymo data used in this study were all collected in US. Although we believe the main conclusions drawn in this study will hold in a different country, specific details are likely to change when IDM models are re-calibrated using a new dataset. It would be interesting to check those differences between different countries if more data are available. Third, due to lack of hardware-level microscopic data (e.g., the error and delay of the actuation or control system), the safety evaluation of this study is limited to surrogate safety measures. It is important to pinpoint the underlying causal factors that can lead to a collision in the context of designing AV car-following algorithms in the future.

In conclusion, the implications of the main findings in this study are threefold: (a). currently AV has significantly larger TTC and smaller DRAC values, which is much safer than HV; (b) compared to HV, ACC and CACC vehicles, trajectories from AV show that they have larger average time headways, response times and jam spacings, indicating that AV is less efficient as other vehicle types; and (c) for HV, in general, there is no significant difference between following AV and following HV. Overall, we conclude that currently AV behaves in a conservative way to enhance its safety at the cost of traffic efficiency.

ACKNOWLEDGMENT

The authors thank all the anonymous reviewers for their insightful comments which significantly improved this article.

REFERENCES

- [1] H. Yu et al., "Automated vehicle-involved traffic flow studies: A survey of assumptions, models, speculations, and perspectives," *Transp. Res. C, Emerg. Technol.*, vol. 127, Jun. 2021, Art. no. 103101.
- [2] V. Milanés and S. E. Shladover, "Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data," *Transp. Res. C, Emerg. Technol.*, vol. 48, pp. 285–300, Nov. 2014.
- [3] V. L. Knoop, M. Wang, I. Wilmink, D. M. Hoedemaeker, M. Maaskant, and E.-J. Van der Meer, "Platoon of SAE level-2 automated vehicles on public roads: Setup, traffic interactions, and stability," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 9, pp. 311–322, Sep. 2019.
- [4] G. Gunter, C. Janssen, W. Barbour, R. E. Stern, and D. B. Work, "Model-based string stability of adaptive cruise control systems using field data," *IEEE Trans. Intell. Veh.*, vol. 5, no. 1, pp. 90–99, Mar. 2020.
- [5] G. Gunter et al., "Are commercially implemented adaptive cruise control systems string stable?" *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 6992–7003, Nov. 2021.
- [6] M. Makridis, K. Mattas, A. Anesiadou, and B. Ciuffo, "OpenACC. An open database of car-following experiments to study the properties of commercial ACC systems," *Transp. Res. C, Emerg. Technol.*, vol. 125, Apr. 2021, Art. no. 103047.
- [7] M. Makridis, K. Mattas, and B. Ciuffo, "Response time and time headway of an adaptive cruise control. An empirical characterization and potential impacts on road capacity," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1677–1686, Apr. 2020.
- [8] M. Makridis et al., "Empirical study on the properties of adaptive cruise control systems and their impact on traffic flow and string stability," *Transp. Res. Rec.*, vol. 2674, no. 4, pp. 471–484, 2020.
- [9] T. Li, D. Chen, H. Zhou, J. Laval, and Y. Xie, "Car-following behavior characteristics of adaptive cruise control vehicles based on empirical experiments," *Transp. Res. B, Methodol.*, vol. 147, pp. 67–91, May 2021.
- [10] T. Li, D. Chen, H. Zhou, Y. Xie, and J. Laval, "Fundamental diagrams of commercial adaptive cruise control: Worldwide experimental evidence," *Transp. Res. C, Emerg. Technol.*, vol. 134, Jan. 2022, Art. no. 103458.
- [11] M.-F. Chang et al., "Argoverse: 3D tracking and forecasting with rich maps," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8748–8757.
- [12] R. Kesten et al. (2019). *Lyft Level 5 AV Dataset 2019*. [Online]. Available: <https://level5.lyft.com/dataset>
- [13] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11621–11631.
- [14] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2446–2454.
- [15] S. Ettinger et al., "Large scale interactive motion forecasting for autonomous driving: The Waymo open motion dataset," 2021, *arXiv:2104.10133*.
- [16] X. Hu, Z. Zheng, D. Chen, X. Zhang, and J. Sun, "Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research," *Transp. Res. C, Emerg. Technol.*, vol. 134, Jan. 2022, Art. no. 103490.
- [17] A. Sharma, Z. Zheng, and A. Bhaskar, "A pattern recognition algorithm for assessing trajectory completeness," *Transp. Res. C, Emerg. Technol.*, vol. 96, pp. 432–457, Nov. 2018.
- [18] J. C. Hayward, "Near miss determination through use of a scale of danger," in *Proc. 51st Annu. Meeting Highway Res. Board*, Washington, DC, USA, 1972, pp. 24–34.
- [19] D. F. Cooper and N. Ferguson, "Traffic studies at T-junctions. 2. A conflict simulation record," *Traffic Eng. Control*, vol. 17, no. 7, pp. 306–309, 1976.
- [20] K. Mattas et al., "Fuzzy surrogate safety metrics for real-time assessment of rear-end collision risk. A study based on empirical observations," *Accident Anal. Prevention*, vol. 148, Dec. 2020, Art. no. 105794.
- [21] J. Archer, "Indicators for traffic safety assessment and prediction and their application in micro-simulation modelling: A study of urban and suburban intersections," Ph.D. dissertation, KTH, Stockholm, Sweden, 2005.
- [22] Y. Kuang, X. Qu, and S. Wang, "A tree-structured crash surrogate measure for freeways," *Accident Anal. Prevention*, vol. 77, pp. 48–137, Apr. 2015.
- [23] X. Chen, L. Li, and Y. Zhang, "A Markov model for headway/spacing distribution of road traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 773–785, Dec. 2010.
- [24] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama, "Dynamical model of traffic congestion and numerical simulation," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, no. 2, p. 1035, 1995.
- [25] D. Helbing and B. Tilch, "Generalized force model of traffic dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 58, no. 1, p. 133, 1998.
- [26] R. Jiang, Q. Wu, and Z. Zhu, "Full velocity difference model for a car-following theory," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 64, no. 1, 2001, Art. no. 017101.
- [27] G. F. Newell, "A simplified car-following theory: A lower order model," *Transp. Res. B, Methodol.*, vol. 36, no. 3, pp. 195–205, 2002.
- [28] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 62, no. 2, p. 1805, 2000.
- [29] C. Jekel and G. Venter, "A Python library for fitting 1D continuous piecewise linear functions," Tech. Rep., 2019, doi: [10.13140/RG.2.2.28530.56007](https://doi.org/10.13140/RG.2.2.28530.56007).
- [30] R. Jiang, M.-B. Hu, H. M. Zhang, Z. Y. Gao, and Q.-S. Wu, "On some experimental features of car-following behavior and how to model them," *Transp. Res. B, Methodol.*, vol. 80, pp. 338–354, Oct. 2015.
- [31] B. Cheng, "Analysis of driver response to collision warning during car following," *JSAE Rev.*, vol. 23, no. 2, pp. 231–237, Apr. 2002.
- [32] T. Gates, D. Noyce, L. Laracuente, and E. Nordheim, "Analysis of driver behavior in dilemma zones at signalized intersections," *Transp. Res. Rec., J. Trans. Res. Board*, vol. 2030, no. 1, pp. 29–39, 2007.
- [33] S. Dinakar, J. W. Muttart, D. E. Edewgaard, M. Giannone, and C. Dickson, "Driver response time in cut-off scenarios from the second strategic highway research program naturalistic database," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2676, no. 2, pp. 706–717, 2021.
- [34] C. Deng, S. Cao, C. Wu, and N. Lyu, "Modeling driver take-over reaction time and emergency response time using an integrated cognitive architecture," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 12, pp. 380–390, Dec. 2019.
- [35] D. Chen, S. Ahn, J. Laval, and Z. Zheng, "On the periodicity of traffic oscillations and capacity drop: The role of driver characteristics," *Transp. Res. B, Methodol.*, vol. 59, pp. 117–136, Jan. 2014.
- [36] A. Sharma, Z. Zheng, J. Kim, A. Bhaskar, and M. M. Haque, "Estimating and comparing response times in traditional and connected environments," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2673, no. 4, pp. 674–684, Apr. 2019.
- [37] Z. Zheng, S. Ahn, D. Chen, and J. Laval, "Applications of wavelet transform for analysis of freeway traffic: Bottlenecks, transient traffic, and traffic oscillations," *Transp. Res. B, Methodol.*, vol. 45, no. 2, pp. 372–384, 2011.
- [38] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 617–643, Mar. 1992.
- [39] Z. Zheng and S. Washington, "On selecting an optimal wavelet for detecting singularities in traffic and vehicular data," *Transp. Res. C, Emerg. Technol.*, vol. 25, pp. 18–33, Dec. 2012.
- [40] M. Treiber and A. Kesting, "Microscopic calibration and validation of car-following models—A systematic approach," *Proc.-Social Behav. Sci.*, vol. 80, pp. 922–939, Jun. 2013.
- [41] T. Lumley, P. Diehr, S. Emerson, and L. Chen, "The importance of the normality assumption in large public health data sets," *Annu. Rev. Public Health*, vol. 23, no. 1, pp. 69–151, 2002.
- [42] M. W. Fagerland, "T-tests, non-parametric tests, and large studies—A paradox of statistical practice?" *BMC Med. Res. Methodol.*, vol. 12, p. 78, Jun. 2012.
- [43] E. Skovlund and G. U. Fenstad, "Should we always choose a nonparametric test when comparing two apparently nonnormal distributions?" *J. Clin. Epidemiol.*, vol. 54, no. 1, pp. 86–92, Jan. 2001.
- [44] J. S. Brunner, M. A. Makridis, and A. Kouvelas, "Comparing the observable response times of ACC and CACC systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19299–19308, Oct. 2022.
- [45] D. Zhang, X. Chen, J. Wang, Y. Wang, and J. Sun, "A comprehensive comparison study of four classical car-following models based on the large-scale naturalistic driving experiment," *Simul. Model. Pract. Theory*, vol. 113, Dec. 2021, Art. no. 102383.

- [46] A. Alhariqi, Z. Gu, and M. Saberi, "Calibration of the intelligent driver model (IDM) with adaptive parameters for mixed autonomy traffic using experimental trajectory data," *Transportmetrica B, Transp. Dyn.*, vol. 10, no. 1, pp. 421–440, 2021.
- [47] A. Sharma, Z. Zheng, and A. Bhaskar, "Is more always better? The impact of vehicular trajectory completeness on car-following model calibration and validation," *Transp. Res. B, Methodol.*, vol. 120, pp. 49–75, Feb. 2019.
- [48] V. Punzo, Z. Zheng, and M. Montanino, "About calibration of car-following dynamics of automated and human-driven vehicles: Methodology, guidelines and codes," *Transp. Res. C, Emerg. Technol.*, vol. 128, Jul. 2021, Art. no. 103165.
- [49] V. Punzo, M. Montanino, and B. Ciuffo, "Do we really need to calibrate all the parameters? Variance-based sensitivity analysis to simplify microscopic traffic flow models," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 184–193, Feb. 2015.
- [50] J. Sun, Z. Zheng, and J. Sun, "The relationship between car following string instability and traffic oscillations in finite-sized platoons and its use in easing congestion via connected and automated vehicles with IDM based controller," *Transp. Res. B, Methodol.*, vol. 142, pp. 58–83, Dec. 2020.
- [51] B. Ciuffo et al., "Requiem on the positive effects of commercial adaptive cruise control on motorway traffic and recommendations for future automated driving systems," *Transp. Res. C, Emerg. Technol.*, vol. 130, Sep. 2021, Art. no. 103305.
- [52] J. Sun, Z. Zheng, and J. Sun, "Stability analysis methods and their applicability to car-following models in conventional and connected environments," *Transp. Res. B, Methodol.*, vol. 109, pp. 212–237, Mar. 2018.
- [53] R. Nishi, A. Tomoeda, K. Shimura, and K. Nishinari, "Theory of jam-absorption driving," *Transp. Res. B, Methodol.*, vol. 50, pp. 116–129, Apr. 2013.
- [54] Z. He, L. Zheng, L. Song, and N. Zhu, "A jam-absorption driving strategy for mitigating traffic oscillations," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 802–813, Apr. 2017.
- [55] R. E. Stern et al., "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 205–221, Apr. 2018.



Zuduo Zheng is currently TMR TAP Chair (Deputy) and an Associate Professor with the School of Civil Engineering, The University of Queensland, Australia. His research interests include traffic flow modeling, travel behavior and decision making, and advanced data analysis techniques, including mathematical modeling, econometrics, numerical optimization in transport engineering, and metaresearch. He was a DECRA Research Fellow supported by the Australian Research Council. He is/was on the Editorial Advisory Board of several transport journals, including *Transportation Research—B: Methodological*, *Transportation Research—C: Emerging Technologies*, and *Journal of Advanced Transportation*. He is currently an Associate Editor of IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.



Danjue Chen received the Ph.D. degree from the Georgia Institute of Technology, in 2012. She is currently an Assistant Professor with the Department of Civil and Environmental Engineering, University of Massachusetts Lowell. Her research interests include the experimental study of automated vehicles, traffic flow theory, control, and the simulation of connected automated vehicles (CAVs), human-cyber-physical-system of CAVs, smart cities. She is a member of the Transportation Research Board Traffic Flow Theory Committee (ACP50). She received the NSF Career Award in 2020. She is/was on the Editorial Advisory Board of *Transportation Research—C: Emerging Technologies*.



Xiangwang Hu received the B.E. degree in transport engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently pursuing the Ph.D. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His research interests include traffic flow modeling, the optimal operation of traffic flow with connected, and autonomous vehicles.



Jian Sun received the Ph.D. degree in transportation engineering from Tongji University, Shanghai, China. He is currently a Professor of transportation engineering with Tongji University. His research interests include intelligent transportation systems, traffic flow theory, AI in transportation, and traffic simulation.