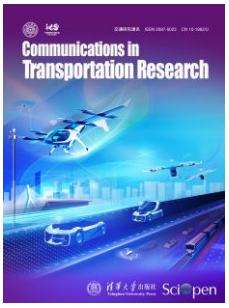
COMMTR

Communications in Transportation Research



SHORT COMMUNICATION

OVPD: A virtual-physical fusion testing dataset of OnSite Autonomous Driving Challenge

Yuhang Zhang, Jiarui Zhang, Bowen Jian, Xin Zhou, Zhichao Lv, Peng Hang, Rongjie Yu, Ye Tian, Jian Sun✉

College of Transportation, Tongji University, Shanghai 201804, China.

✉ Corresponding author.

E-mail: sunjian@tongji.edu.cn

DOI 10.26599/COMMTR.2026.9640058

Received 27 April 2026

Revised 28 May 2026

Accepted 10 October 2026

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640058

Just Accepted

COMMTR Homepage

© The Author(s) 2026.

This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Published by **Tsinghua University Press**

https://www.sciopen.com/journal/2097-5023

清華大學出版社 Tsinghua University Press | SciOpen

Commun Transp Res

Short Communication

OVPD: A virtual-physical fusion testing dataset of OnSite Autonomous Driving Challenge

Yuhang Zhang, Jiarui Zhang, Bowen Jian, Xin Zhou, Zhichao Lv, Peng Hang, Rongjie Yu, Ye Tian, Jian Sun

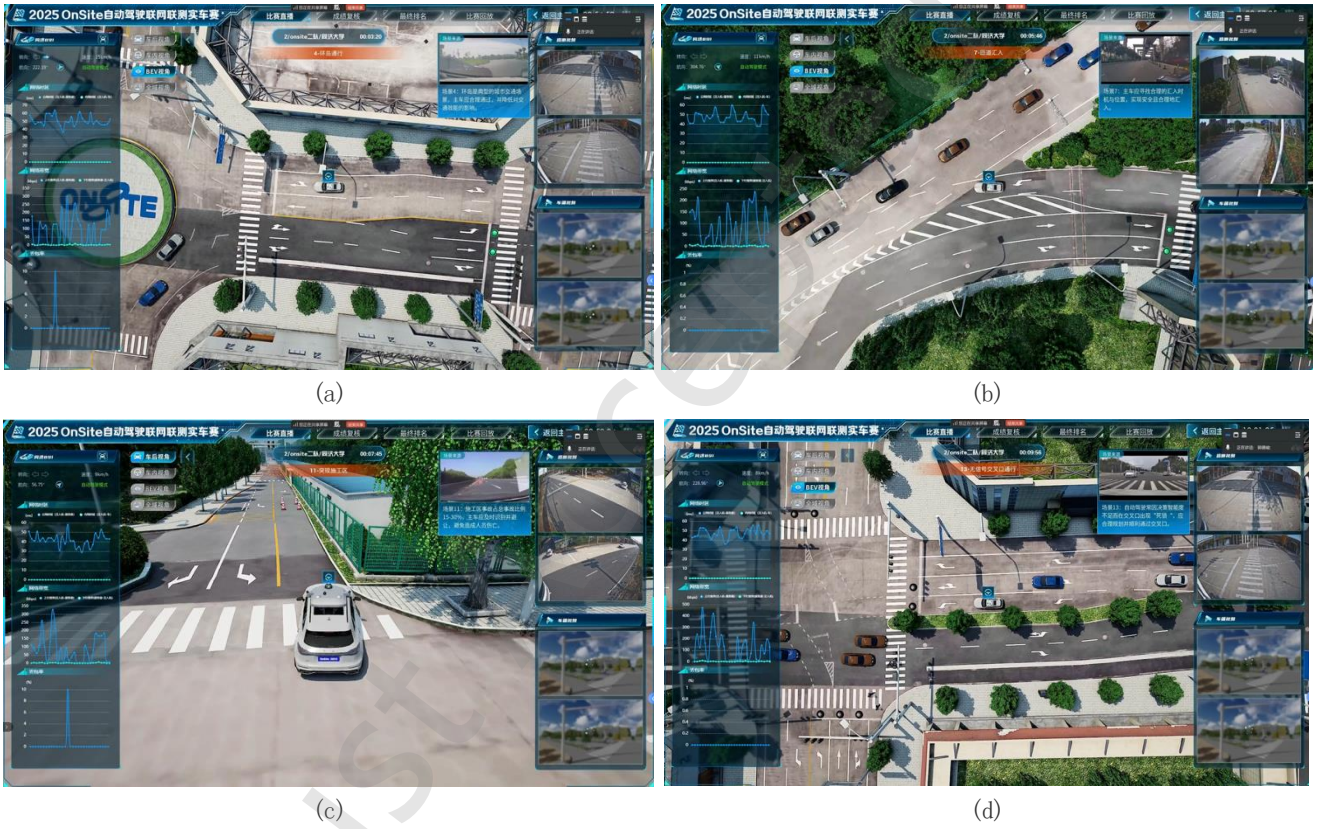


Fig. 1. A few exemplary visualizations of the OVPD dataset: (a) a roundabout scenario, highlighting yielding and interaction with circulating traffic; (b) a merge scenario, focusing on gap selection and coordination with surrounding vehicles; (c) an obstacle-avoidance scenario in a construction zone; and (d) an unsignalized intersection scenario, emphasizing right-of-way judgment and conflict negotiation.

1 Introduction

Autonomous driving is evolving toward L3+ autonomy, raising the demand for high-quality, deployment-oriented evaluation data in complex urban environments (SAE International, 2021). However, existing datasets and benchmarks share three limitations. First, long-tail, high-risk scenarios rarely occur in naturalistic driving, making systematic coverage difficult. Second, purely simulated environments leave a sim-to-real gap in vehicle dynamics, latency, and closed-loop stability. Third, closed-loop benchmarks emphasize trajectory fidelity, safety, efficiency, and comfort while underrepresenting deployment-critical rule compliance and interaction coordination, and their route-level aggregation obscures scenario-specific diagnosis (Jia et al., 2024; United Nations Economic Commission for Europe, 2025).

College of Transportation, Tongji University, Shanghai 201804, China.

Corresponding author. E-mail: sunjian@tongji.edu.cn

Received: April 27, 2026; Revised: May 28, 2026; Accepted: October 10, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

To address these challenges, the 2025 Onsite Autonomous Driving Challenge, jointly organized by Tongji University and the Ministry of Transport of China, adopts a real-vehicle-in-the-loop testing paradigm (Research Team for Traffic OPerations and Simulation, 2025). Participating vehicles running fully deployed autonomous driving algorithms interact with virtual background traffic in controllable closed-loop test environments, enabling scenario-level repeatable evaluation while preserving authentic vehicle dynamics and actuation feedback (Cui et al., 2025).

Building upon this, we release OVPD (Onsite Virtual–Physical Dataset): complete test records from 19 participating teams over a chain of 15 atomic scenarios (285 scored instances) spanning emergency response, traffic efficiency, and rule compliance, totaling about 4 hours. Each record provides vehicle trajectories and states, control commands, and digital-twin-rendered surround-view observations, together with atomic-scenario-level scores under a five-dimensional framework (safety, efficiency, comfort, rule compliance, and traffic coordination) for failure diagnosis and iterative improvement (Cui et al., 2025; Research Team for Traffic OPerations and Simulation, 2025).

OVPD differs from prior resources along five axes. In data source, open-loop, log-replay datasets and simulation-based closed-loop benchmarks (e.g., Bench2Drive; Jia et al., 2024) rely on recorded logs or synthetic engines, whereas OVPD records deployed stacks driving a real vehicle against controllable virtual traffic. In testing paradigm, it replaces open-loop regression or pure-simulation closed loops with virtual–physical fusion that preserves real vehicle dynamics. In supported tasks, it targets capability-level diagnosis rather than trajectory imitation. In evaluation metrics, it adds deployment-critical rule-compliance and traffic-coordination dimensions beyond the safety, efficiency, and comfort dimensions typically reported. In intended applications, it supports failure attribution and regression testing for deployment readiness rather than leaderboard ranking.

2 From Onsite to OVPD

2.1. Testing Organization

OVPD is collected from the on-vehicle records of the 2025 Onsite Autonomous Driving Challenge (Research Team for Traffic OPerations and Simulation, 2025). Each team's clip is a chain of 15 atomic scenarios, each with explicit trigger and termination points, connected end-to-end along the site's lane-level topology into a complete route.

Scenario design crosses road-facility types (e.g., intersections, roundabouts, ramp merges) with capability dimensions (e.g., emergency avoidance, lane-change negotiation, rule compliance), giving each atomic scenario a clear capability intent and semantic boundary. It targets three deployment bottlenecks — insufficient emergency response, rule violations, and degraded efficiency — with difficulty tuned by available driving space, reaction-time window, and participant density, so the scenarios elicit dense critical decisions and yield diagnostic, difficulty-adjustable planning and decision-making data.

2.2. Virtual-Physical Data Recording

OVPD is collected under a virtual–physical fusion framework enabled by the VP-AutoTest platform (Cui et al., 2025). The ego vehicle runs deployed autonomous driving algorithms on a real vehicle while interacting in real time with platform-generated virtual background traffic; background flow and roadside perception are synchronized to the ego domain controller, and the ego's control commands and motion states are recorded at 10 Hz. By capturing real vehicle dynamics without altering scenario semantics, this paradigm bridges simulation controllability and deployment fidelity.

This paradigm also advances conventional vehicle-in-the-loop (VIL) and hardware-in-the-loop (HIL) testing. HIL exercises only an isolated controller against simulated signals, and typical VIL either keeps the vehicle on a dynamometer or drives it against prescripted virtual objects that do not react to the ego; neither reproduces full-vehicle dynamics under genuine multi-agent interaction. OVPD instead runs the complete deployed stack on a moving real vehicle whose virtual background traffic reacts in closed loop, so it (i) captures authentic execution latency, steering response, and control stability; (ii) elicits real interactive negotiation such as yielding, merging, and cut-in response; (iii) stages high-risk, long-tail scenarios safely and repeatably that are unsafe on public roads; and (iv) yields a shared, reproducible dataset with scenario-level, five-dimensional evaluation. These risks are concretely visible in the released records, which log 9 safety-driver takeovers and 11 contacts across the 19 clips — outcomes purely simulated closed loops would not surface.

All modalities share a global timestamp for replay: ego control commands and motion states (JSON, 10 Hz), frame-by-frame surround-view and BEV renderings from the digital-twin platform, background participants' trajectories and states (JSON, 10 Hz), traffic-light states, HD maps, and the automatically computed scenario-level overall and per-dimension scores.

2.3. Comprehensive Metrics

OVPD releases evaluation outputs and computation protocols for each record, anchoring scores at the atomic-scenario level before aggregation so that each sub-score maps to a specific capability requirement or rule constraint, keeping cross-method comparison stable and interpretable (Jia et al., 2024; United Nations Economic Commission for Europe, 2025).

For a chain of N atomic scenarios, the route-level score averages the per-scenario scores S_i, each combining efficiency as a base term with safety, comfort, rule compliance, and coordination as constraint factors:

$$S_i = G_{\text{eff}}^{0.5} \times \left(0.5\, G_{\text{safety}} + 0.3\, G_{\text{comfort}} + 0.1\, G_{\text{rule}} + 0.1\, G_{\text{coord}}\right)$$

where $G_{\text{eff}} \in [0,100]$ and $G_{\text{safety}}, G_{\text{comfort}}, G_{\text{rule}}, G_{\text{coord}} \in [0,10]$. The five dimensions are defined as follows (full formulas in the Appendix). **Efficiency** combines a pass/fail completion base with a travel-time ratio against a 25 km/h reference. **Safety** is a safety-floor rule: any collision, off-road event, or takeover zeroes the score. **Comfort** penalizes the time fraction exceeding expert-set acceleration, jerk, and yaw-rate thresholds. **Rule compliance** applies severity-weighted penalties for signal, yielding, and maneuver violations logged with type, location, and time. **Traffic coordination** is the mean driving quality of background vehicles within a 30 m influence radius.

The base–constraint structure and the coefficient assignment (safety 0.5, comfort 0.3, rule 0.1, coordination 0.1) were calibrated by an expert panel using the Analytic Hierarchy Process (AHP), instantiating the safety-first, comfort-second principle common to closed-loop planning benchmarks (Jia et al., 2024), while retaining rule compliance and coordination as complementary social-driving factors. The square-root transform on the efficiency term prevents speed-chasing from dominating the product: since G_{eff} spans $[0,100]$ against the constraint terms' $[0,10]$, $G_{\text{eff}}^{0.5}$ compresses its range so the constraint dimensions retain discriminative power.

3 Dataset-level Observations

Just Accepted

We analyze OVPD as a dataset — coverage, difficulty, metric discriminability, and diagnostic granularity — not as a method ranking. All clips are 2025 Onsite Challenge records scored by the Section 2.3 metrics; full statistics and the scoring protocol are in the Appendix.

1) Category-level trends reveal a difficulty gradient: across the 285 scenarios, per-dimension mean scores decline from emergency-response (A) through traffic-efficiency (B) to rule-compliance (C) — e.g., traffic coordination averages 76.7, 75.5, and 70.4. Hard safety failures (safety score 0) occur in 31 scenarios (10.9%), most frequent in rate among rule-compliance scenarios (17%), while safety-driver takeovers concentrate in emergency-response scenarios (6 of 9).

2) ***Dimensions differ in how they separate methods*****:** safety, comfort, and rule compliance are near-saturated and bimodal (median 100 with a minority of zeros), whereas efficiency (median 70.2) and traffic coordination (median 82.5) vary continuously and separate methods more granularly. Team totals span 29.6–89.4 (median 79.4), so methods with similar overall scores can differ markedly across dimensions — heterogeneity OVPD preserves rather than collapsing into one number (Jia et al., 2024).

3) ***Scenario diversity supports a layered structure*****:** we define interaction density as the average number of distinct participants entering the ego's 30 m zone per scenario — a measure of negotiation load, not risk. It varies markedly: rule-compliance scenarios reach about 1.65× the density of traffic-efficiency scenarios (counts in the Appendix), so some scenarios are lightly interactive checks while others impose dense multi-agent negotiation. This layered structure gives a reproducible baseline for capability and regression testing and an open base for adding harder or finer-grained scenarios.

4 Conclusion

OVPD is a virtual–physical fusion testing dataset from the 2025 Onsite Autonomous Driving Challenge (Research Team for Traffic OPerations and Simulation, 2025): by running deployed algorithms on a real vehicle against controllable virtual traffic, it preserves vehicle dynamics and closed-loop interaction with scenario-level reproducibility (Cui et al., 2025). Its atomic-scenario records and five-dimensional protocol give a compact basis for diagnosing deployment-critical capabilities.

More broadly, OVPD signals a shift from offline, route-aggregated scoring toward reproducible, scenario-level, deployment-critical testing. The virtual–physical fusion paradigm can support this transition by enabling the safe, repeatable evaluation of high-risk, long-tail scenarios that are impractical to stage on public roads. It may therefore provide useful evidence for the development of future certification-oriented testing standards, particularly as shared scenario libraries incorporate new capability requirements. Two open problems remain central. First, virtual–physical consistency must become measurable: residual gaps between virtual traffic or rendered perception and reality bound the credibility of such evaluation and should govern which scenarios warrant physical testing. Second, metric legitimacy should move from expert-set choices toward standardized, contestable consensus, so that weighting, aggregation, and dimension definitions are comparable across methods and platforms. By releasing its records and full scoring protocol, OVPD offers a reproducible starting point for both.

Just Accepted

Appendix A: Summary of Test Scenarios

Table 1. Summary of Typical Autonomous Driving Test Scenarios

	Scenario Name and Description
(A) Emergency-Response Scenarios	
A1	**Pedestrian pop-out crossing:** Pedestrians suddenly enter the ego lane from roadside under limited visibility. The ego must rapidly infer intent and perform emergency braking/yielding or lateral evasion when necessary.
A2	**Suddenly exposed stalled vehicle:** A stopped broken-down vehicle is revealed after the lead vehicle changes lanes. The ego must recognize the stationary obstacle and mitigate risk within a short distance.
A3	**Red-light-running emergency vehicle:** During a left turn at a T-junction, an occluded ambulance from the left approach runs a red light. The ego must react promptly to avoid collision.
A4	**Lead vehicle hard braking:** The lead vehicle brakes sharply during car-following. The ego must choose braking/brake-and-lane-change/decelerated following conditioned on headway and surrounding traffic.
A5	**Aggressive e-scooter cut-in:** An e-scooter rapidly cuts in from the non-motorized lane. The ego must predict near-term intent within a short reaction window and avoid collision.
A6	**Fast background-vehicle cut-in:** A side vehicle accelerates and forces a merge into the ego lane. The ego must detect cut-in intent, anticipate conflict, and execute a reasonable evasive strategy.
A7	**Work-zone obstacle avoidance:** An unannotated work-zone occupation appears ahead. The ego must detect it, locally replan, and pass safely via deceleration and/or lane change.
(B) Traffic-Efficiency Scenarios	
B1	**Roundabout traversal**: The ego yields to circulating vehicles, enters at an appropriate gap, maintains a reasonable speed inside the roundabout, and plans an efficient exit while avoiding conflicts.
B2	**Ramp merging:** With high-speed mainline flow and fast-approaching ramp vehicles, the ego must perform multi-agent prediction and timely longitudinal/lateral decisions to merge safely without significantly disturbing traffic flow.
B3	**Non-motorized spillover:** A blocked non-motorized stream suddenly spills into the motorized lane. The ego must anticipate lateral motion and respond (braking/deceleration/lane change) without creating a moving bottleneck.
B4	**Unsignalized intersection negotiation:** The ego plans a left turn while vehicles from other directions and pedestrians create repeated right-of-way interactions. The ego must pass safely without being overly conservative and blocking traffic.
(C) Rule-Compliance Scenarios	
C1	**Consecutive right turns with blockage:** A failure vehicle blocks the right-turn lane. The ego must proceed under constraints, evaluating necessary rule deviation under safety-first when compliance conflicts with task reachability (criteria defined in the metric section)
C2	**Queue spillback at an intersection:** After the green phase, downstream congestion leaves limited space; crossing the stop line risks illegal stopping inside the intersection and blockage. The ego should wait until traffic clears.
C3	**Complex signal-phase composition:** The signal switches from right-turn prohibited to right-turn permitted. The ego must interpret signal constraints correctly and yield to pedestrians during the turn.
C4	**Autonomous parking under restrictions:** Roadside constraints (bus stop, solid yellow line, hatched markings) prohibit stopping. The ego must recognize markings and select a legal parking area to complete the task.

Just Accepted

Appendix B: Evaluation Metric Definitions

Every dimension is scored per atomic scenario and then uniformly averaged over the $\boldsymbol{N}$ scenarios of a chain: the overall driving score is $\text{Score}_{\text{overall}} = \frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{S_i}$, and the same scenario-wise averaging applied dimension by dimension yields the per-dimension leaderboards used for capability diagnosis.

1) Efficiency

Efficiency measures whether and how promptly the task is completed and serves as the base term of the scenario score:

$$\boldsymbol{G}_{\text{eff}} = \text{pass} \times \mathbf{30} + \mathbf{70} \times \frac{\boldsymbol{t_E}}{\boldsymbol{t}_{\text{pass}}}$$

where pass $\in \{\mathbf{0}, \mathbf{1}\}$ indicates successful completion (if the scenario is not completed, $\boldsymbol{G}_{\text{eff}} = \mathbf{0}$; $\boldsymbol{t}_{\text{pass}}$ is the actual time from the scenario trigger point to its termination point, including any stopping or waiting; and $\boldsymbol{t_E}$ is the expected time to traverse the scenario at a constant $\mathbf{25}$ km/h (scaled by a per-scenario coefficient). Because sustaining $\mathbf{25}$ km/h is rarely achievable in dense urban interaction, $\boldsymbol{t_E}/\boldsymbol{t}_{\text{pass}} \leq \mathbf{1}$ in practice, so $\boldsymbol{G}_{\text{eff}} \in [\mathbf{0}, \mathbf{100}]$.

2) Safety

Safety adopts a safety-floor principle with a hard-zero penalty for any safety-critical event:

$$\boldsymbol{G_{safety}} = \begin{cases} \mathbf{10}, & \boldsymbol{no\ collision/off-road/takeover} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases}$$

The three safety-critical events are (i) collision with any

traffic participant or road infrastructure, (ii) off-road departure of the ego trajectory, and (iii) a safety-driver takeover during real-vehicle testing, the last serving as direct evidence of on-vehicle risk. Per the challenge rule, any such event sets the entire scenario score to zero, not only the safety term.

3) Comfort

Comfort deducts, for each of five ride-dynamics indicators, the fraction of scenario time its threshold is exceeded, until the 10 points are exhausted:

$$G_{comfort} = max\left(0, 10 - 10\sum_{k=1}^{5}\frac{t_{exceed,k}}{t_{all}}\right)$$

where $t_{\text{exceed},k}$ is the time indicator k exceeds its threshold and t_{all} is the scenario duration. The five indicators and their thresholds are: longitudinal acceleration ($\pm 3\ \text{m/s}^2$), lateral acceleration ($\pm 2.5\ \text{m/s}^2$), longitudinal jerk ($\pm 6\ \text{m/s}^3$), lateral jerk ($\pm 1\ \text{m/s}^3$), and yaw rate (± 0.5 rad/s).

4) Rule Compliance

Rule compliance applies severity-weighted, event-level penalties over high-frequency urban violations, until the 10 points are exhausted:

$$G_\{rule\} = max\left(0, 10 - \sum_{e} w_e\right)$$

where the sum runs over violation events e in the scenario and w_e is the type-specific penalty. Following the challenge rule, running a red light or exceeding the speed limit by more than 50% deducts 5 points; illegal overtaking, illegal stopping, wrong-way driving, failing to yield to a vehicle or pedestrian, illegally entering an intersection, or 20–50% speeding deducts 2.5; and illegal reversing, illegal U-turns, or lane-line crossing deducts 1. Violations span signal-, yielding-, and maneuver-related types, each logged with type, location, and time for replay-based verification and error attribution.

5) Traffic Coordination

Traffic coordination quantifies the ego vehicle's externality on surrounding traffic. The set of affected background vehicles is identified by a DBSCAN-style spatial selection centered on the ego vehicle with an influence radius $r = 30$ m. For each affected vehicle, its efficiency, safety, and comfort sub-scores are computed as in B.1–B.3, and G_{coord} is their mean over the M affected vehicles:

$$G_{coord} = \frac{1}{3M}\sum_{i=1}^{M}\left(\widehat{G_{eff}^{i}} + G_{safety}^{i} + G_{comfort}^{i}\right)$$

where $\widehat{G_{\text{eff}}^{i}}$ is the efficiency score rescaled to $[0, 10]$ so that the three terms are commensurate, giving $G_{\text{coord}} \in [0, 10]$. A higher G_{coord} indicates that the ego vehicle better preserves the mobility, safety, and comfort of nearby traffic. Because traffic coordination also depends on intrinsic scenario properties, it is most informative for cross-method comparison under identical scenarios.

Appendix C: Dataset Statistics

Table 2. Dataset scale and scenario distribution

Item	Value
Clips / teams	19 / 19 (missing team IDs: team_13, team_18)
Test sites (Shanghai / Beijing clips)	12 / 7
Atomic-scenario instances	285 (= 19 × 15)
Instances by capability category (A / B / C)	133 / 76 / 76
Capability categories	A: emergency response (7 scenarios); B: traffic efficiency (4); C: rule compliance (4)
Total ego-recorded duration	4.03 h (mean clip 763.9 s)
Total ego state samples	138,047

Table 3. Trajectory counts

Item	Value
Route-level ego trajectories	19 (one per clip)
Scenario-level ego trajectory segments	282
Distinct background-participant trajectories	3,132
• motor vehicle	1,681
• traffic-flow actor	949
• pedestrian	274
• cyclist / non-motorized	183
• obstacle	45
Background state samples	2,306,791

Table 4. Interaction density (exposure / negotiation load):

Metric	Overall	A	B	C
Distinct interacting participants within 30 m (dataset total)	23,458	10,369	5,068	8,021
Mean distinct interacting participants / scenario	83.18	77.96	66.68	109.88
Highest/lowest category mean ratio	1.65×	—	—	—
Distinct participants, min gap < 10 m	5,688	2,667	1,035	1,986
Distinct participants, min gap < 5 m	1,706	887	354	465
Distinct participants, min gap < 3 m	991	518	224	249

Interaction density is the number of **distinct** background participants that enter the ego's 30 m interaction zone during a scenario — each participant counted once, regardless of how many times it enters or exits (matching the traffic-coordination radius in Section 2.3). It measures interaction/negotiation load, **not** outcome risk; the close-proximity rows below shade toward near-conflict exposure.

Participant types within 30 m: vehicle 21,279; cyclist / non-motorized 2,179; pedestrian 0.

Just Accepted

Table 5. Failure statistics

Metric	Overall	A	B	C
Safety-driver takeovers	9	6	1	2
Collision / contact events	11	2	4	5
Hard-failure scenarios (safety = 0)	31 (10.9%)	10	8	13
Incompletion scenarios (efficiency = 0)	31 (10.9%)	10	8	13
Rule-violation events	33	16	3	14

Counts from the official per-scenario diagnosis reports (n = 285/285). Counts from the official per-scenario diagnosis reports (n = 285/285).

Per-clip rates: 0.47 takeovers/clip and 0.58 contacts/clip.

Table 6. Score distributions (0–100)

Per-dimension distribution across all 285 scored scenarios:

Dimension	Min	Median	Mean	Max	SD
Safety	0	100.0	89.1	100	31.2
Efficiency	0	70.2	67.1	100	29.5
Comfort	0	97.8	85.4	100	30.5
Rule compliance	0	100.0	86.5	100	31.2
Trafficcoordination	0	82.5	74.7	100	27.7

Category mean scores:

Category	Safe.	Effi.	Comf.	Rule.	Coord.
A (emergency)	92.5	72.5	87.9	90.2	76.7
B (efficiency)	89.5	62.2	85.9	87.2	75.5
C (rule)	82.9	62.4	80.3	79.3	70.4

Team overall scores (n = 19): min 29.6, median 79.4, mean 74.2, max 89.4, SD 15.9.

Just Accepted

Author contributions

Yuhang Zhang: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Jiarui Zhang: Concept Designation, Review, Supervision, Methodology.

Bowen Jian: Concept Designation, Review, Supervision, Methodology.

Xin Zhou: Concept Designation, Review, Supervision, Methodology.

Zhichao Lv: Review & editing, Formal Analysis, Methodology.

Peng Hang: Review & editing, Formal Analysis, Methodology.

Rongjie Yu: Review & editing, Formal Analysis, Methodology.

Ye Tian: Review & editing, Formal Analysis, Methodology.

Jian Sun: Review & editing, Formal Analysis, Methodology.

Replication and data sharing

For reproducibility and verification, we have organized and open-sourced the main data. The raw data can be found at:

https://huggingface.co/datasets/Yuhang253820/Onsite_OVPD.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (52125208, 52232015) and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No.JYB2025XDXM123).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Just Accepted

References

Cui, Y., Fang, S., Zhang, J., Huang, Y., Xu, C., Zhu, B., et al., 2025. VP-AutoTest: A virtual-physical fusion autonomous driving testing platform. arXiv preprint . https://doi.org/10.48550/arXiv.2512.07507

Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J., 2024. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems, 37, 819–844.

Research Team for Traffic OPerations and Simulation, 2025. The 2025 Onsite autonomous driving connected joint testing real-vehicle competition concludes successfully, marking new breakthroughs in autonomous driving evaluation research. https://tops.tongji.edu.cn/info/1002/2513.htm (accessed May 25, 2026).

SAE International, 2021. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard J3016 202104, SAE International, Warrendale, PA.

United Nations Economic Commission for Europe, 2025. Innovation and application of testing methodology on traffic rule compliance for ADS-equipped vehicles. Informal document GRVA-23-44e, UNECE.

Author biography



Yuhang Zhang received the B.S. degree in College of Transportation from Tongji University, Shanghai, China. He is currently pursuing the M.S. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His main research interests include end-to-end autonomous driving algorithm, world model and reinforcement learning.



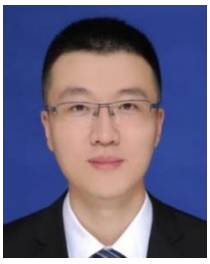
Jiarui Zhang received the B.S. degree in College of Civil Aviation from Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is currently pursuing the Ph.D. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His main research interests include autonomous driving testing and evaluation, and testing scenario generation.



Bowen Jian received the B.S. degree in College of Transportation from Tongji University, Shanghai, China, in 2023. He is currently pursuing the Ph.D. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His main research interests include law-compliance testing and enhancement for autonomous vehicles.



Xin Zhou received the B.S. degree in transportation engineering from Tongji University, Shanghai, China. He is currently pursuing the Ph.D. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His research interests include fault diagnosis and regulation for highly automated vehicles.



Zhichao Lv is a Postdoctoral Researcher with the College of Transportation Engineering, Tongji University. He received the B.Eng. degree from Taiyuan University of Technology in 2010 and the M.Sc. and Ph.D. degrees in control engineering and vehicle engineering from Jilin University and Tongji University in 2018 and 2023, respectively. His research interests include human-like decision-making and reinforcement learning for intelligent driving, and high-performance in-vehicle computing platforms.



Peng Hang is an Associate Professor at the Department of Traffic Engineering, Tongji University, Shanghai, China. He received the Ph.D. degree with the School of Automotive Studies, Tongji University, Shanghai, China, in 2019. He was a Visiting Researcher with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, in 2018. From 2020 to 2022, he served as a Research Fellow with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His research interests include vehicle dynamics and control, decision making, motion planning and motion control for autonomous vehicles. He serves as an Associate Editor of IEEE Internet of Things Journal, IEEE Transactions on Vehicular Technology, Journal of Field Robotics, IET Smart Cities, and SAE International Journal of Vehicle Dynamics, Stability, and NVH.



Rongjie Yu received the B.Sc. degree from TongjiUniversity in 2010 and the M.Sc. and Ph.D. degreesin traffic engineering from the University of Central Florida in 2012 and 2013, respectively. He iscurrently an Associate Professor with the Collegeof Transportation Engineering, Tongji University.His research interests include traffic safety, humanbehavior, and safety evaluation of connected andautonomous vehicles.



Ye Tian received the Ph.D. degreein transportation engineering from The University ofArizona, Tucson, AZ, USA, in 2015. He is currentlyan Associate Professor of transportation engineering with Tongji University, Shanghai, China. Hisresearch interests include active demand manage-ment, dynamic traffic assignment, mesoscopic traf-fic simulation, and safety assurance of automatedvehicles.



Jian Sun received the Ph.D. degree from Tongji University in 2006. Subsequently, he was at Tongji University as a Lecturer, and then promoted to the position as a Professor in 2011, where he is currently a Professor with the College of Transportation Engineering and the Dean of the Department of Traffic Engineering. His main research interests include traffic flow theory, traffic simulation, connected vehicleinfrastructure system, and intelligent transportation system.

Under Accepted