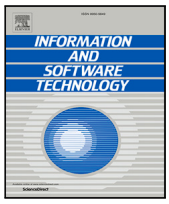
Information and Software Technology 187 (2025) 107859

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Advancing autonomous driving system testing: Demands, challenges, and future directions

Yihan Liao [a], Jingyu Zhang [a,*], Jacky Keung [a], Yan Xiao [b], Yurou Dai [c]

[a] City University of Hong Kong, Hong Kong, China
[b] Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[c] Lehigh University, Bethlehem, USA

ARTICLE INFO

Dataset link: https://github.com/ADSTestingSurvey/ADS-Testing

Keywords:
Autonomous driving systems
Software testing
Vehicle-to-Everything
Foundation models

ABSTRACT

Context: Autonomous driving systems (ADSs) promise enhanced transportation efficiency but face critical challenges in ensuring reliability across complex driving environments. Effective testing is essential to validate ADS performance and mitigate real-world risks.

Objective: This study investigates current ADS testing practices for both modular and end-to-end systems, identifies key demands (needs required by practitioners and researchers), and examines gaps between research and real-world demands. We review critical testing techniques and extend to involve Vehicle-to-Everything (V2X) communication and Foundation Models (FMs), including large language models and vision foundation models, in enhancing ADS testing performance. We provide literature reviews and outline future directions for each demand of industry practitioners and academic researchers.

Methods: A large-scale survey was conducted with 100 participants, including industry practitioners and academic researchers. We first discuss survey questions with professionals, distribute them to industry practitioners and academic researchers, and conduct follow-ups. Quantitative and qualitative analyses uncover key trends and challenges.

Results: Findings reveal that existing ADS testing techniques struggle to evaluate real-world performance comprehensively, including the diversity of corner cases, the gap between simulation and real-world testing, the lack of comprehensive testing criteria, potential attacks, practical deployment for V2X, and the computational costs for FMs. By analyzing participants' responses and 105 relevant papers, we summarize the future research directions.

Conclusion: Our study highlights critical research gaps in ADS testing and underscores the demands of industry practitioners and academic researchers. We provide future directions for ADS: comprehensive testing criteria, cross-model collaboration in V2X, enhancing cross-modality (e.g., text and image) adaptation in FM testing, and scalable ADS validation frameworks. These insights contribute to advancing software engineering practices for ADS development, ensuring safer and more reliable autonomous systems.

1. Introduction

The rapid advancements in autonomous driving systems (ADSs) have revolutionized modern transportation, but safety concerns and testing challenges remain critical barriers to widespread adoption [1–3]. Testing autonomous driving systems requires addressing complex challenges, including corner case diversity, real-world simulation fidelity, and the integration of emerging technologies like Vehicle-to-Everything (V2X) communication. For instance, leading industry players such as Tesla launched the Full Self-Driving (FSD) Beta program in 2020, allowing experienced drivers to test the latest ADS capabilities [1]. Waymo tested Level 4 [4] autonomous driving in multi-city environments and launched the autonomous driving taxi service in 2018, which is now employed more than a hundred thousand times per week [2]. Apollo Go is subject to Baidu's autonomous driving travel service, which started in 2020 and has already provided nearly nine hundred thousand rides [3]. However, despite these advancements, the safety and reliability of ADSs remain a major concern, with reports of nearly 400 crashes, particularly involving Tesla's ADS [5]. Therefore, rigorous testing methodologies are crucial to ensuring the robustness

* Corresponding author.
E-mail addresses: yihanliao4-c@my.cityu.edu.hk (Y. Liao), jzhang2297-c@my.cityu.edu.hk (J. Zhang), jacky.keung@cityu.edu.hk (J. Keung), xiaoy367@mail.sysu.edu.cn (Y. Xiao), yrd224@lehigh.edu (Y. Dai).

https://doi.org/10.1016/j.infsof.2025.107859
Received 25 March 2025; Received in revised form 21 June 2025; Accepted 28 July 2025
Available online 5 August 2025

of ADSs, a topic extensively explored in the software engineering community.

In recent years, the software engineering community has presented many testing insights and approaches, one of the prominent fields is test case generation, which aims to create diverse and challenging driving scenarios for ADS validation [6–12]. Besides, to reduce the costs of ADS real-world testing, simulation platforms have emerged as an effective alternative [3,13–15] by generating a large number of driving test scenarios. V2X communication is another promising area of research that enhances information exchange between vehicles and their surroundings, allowing wireless communication with vehicles, infrastructure, and pedestrians, improving safety and efficiency. Its integration adds complexity to testing, as cooperative ADSs require validation at both the individual vehicle and multi-agent/system-wide levels. Existing testing frameworks mainly focus on single-agent performance, neglecting V2X interactions [16]. This survey examines V2X testing practices, identifying gaps between research and real-world implementation and providing insights into future directions for ensuring V2X-enabled ADSs' robustness and reliability [17–19].

Additionally, emerging Foundation Models (FMs), such as Large Language Models (LLMs) and Vision Foundation Models (VFMs), are being integrated into ADS testing to improve methodologies [10,20]. LLMs can automate scenario generation, create adversarial corner cases, and provide natural language explanations for test results [12, 21–23]. VFMs enhance perception by synthesizing realistic environments and identifying critical failure cases [24,25]. Current research mainly explores theoretical applications, with few surveys on the practical integration of FMs into ADS testing frameworks. This gap highlights the need to explore how FMs can improve testing methodologies. Thus, we aim to investigate FMs' role in ADS testing, identify challenges, and propose future directions to enhance ADS robustness and reliability.

With the rapid development of ADS testing techniques, many survey papers have summarized and compared them [26,27]. The most related paper to our work is from Lou et al. [26], which conducted a survey and literature review on ADS testing. However, compared to their work, we have (1) covered a broader scope by incorporating over three times the number of survey questions as their study (31 questions) with a similar number of responses and analyzed the results separately for academic and industry participants, (2) expanded the discussion to include testing criteria, potential attacks, V2X collaboration, FMs, etc. Consequently, a gap exists in understanding how emerging research aligns with the current ADS testing demands of industry practitioners and academic researchers, particularly for Level-3 to Level-5 ADSs as defined by the SAE J3016 standard [28], which describe conditional to full driving automation and are the primary focus of both industrial deployment and academic investigation. To address this, we present a large-scale survey and literature review on two systems under test (SUT): modular and end-to-end (E2E) ADSs. Modular ADSs are composed of separate perception, planning, and control modules, while E2E ADSs rely on a unified deep learning model to directly map sensor inputs to driving actions. This study aims to explore three key research questions (RQs):

- **RQ1:** What are the current practices and demands in ADS testing?
- **RQ2:** To what extent does the current work contribute to the industry practitioners' and academic researchers' demands?
- **RQ3:** What are the challenges and future directions of demands?

Our survey methodology has three stages: discussion, survey, and follow-up. Specifically, we first discuss current practices and demands with professionals to acquire an initial understanding. Based on these, we design a 107-question survey and distribute it to gather industry practitioners' and academic researchers' experiences and expectations. Finally, in the follow-up stage, participants explain the reasons for their demands, helping us explore current demands in depth.

Our survey studies three aspects: the basic information about ADS and its testing techniques, testing on V2X communication, and the utilization of FMs in ADS testing. Based on the response, we summarized 7 demands, including 1 V2X-related and 2 FM-related ADS testing demands. We analyzed the perspectives of both practitioners and researchers on each demand, highlighting their respective concerns and priorities. The demands are (1) diversity in corner cases, (2) bridging gaps between simulation and real-world testing, (3) more comprehensive testing criteria, (4) effective defenses against current attacks, and (5) seamless cross-model collaboration in V2X systems. The demands of FMs for testing contain (6) Test case quality generated by LLMs and (7) reliable cross-modality integration when using FMs. To the best of our knowledge, we are the first to conduct surveys on the V2X system and utilization of FMs in ADS testing. To understand the state-of-the-art techniques in ADS testing, we conduct literature reviews on papers published in mainstream venues across various domains, including software engineering (TSE, TOSEM, ICSE, ISSTA, FSE, ASE, etc.), vehicle intelligence (TITS, TIV, IV, etc.) and artificial intelligence (AAAI, CVPR, NeurIPS, etc.) Along with the survey responses, we conclude seven emerging challenges and demands.

Contribution. The main contributions of this work are as follows:

- We conduct an advanced ADS testing survey by investigating testing practices and challenges in two SUTs: modular and E2E ADS.
- We are the first to extend the ADS testing survey by investigating V2X collaboration and the role of FMs in ADS testing, discussing their challenges and the gaps in current testing frameworks.
- We conduct a large-scale survey (107 questions) with both academic and industrial participants, identifying key testing demands and gaps between research and real-world practices.

2. Related work

Huang et al. conducted a comprehensive literature review of the safety of Deep Neural Networks (DNNs) by surveying 202 papers [29]. They discussed the testing of DNNs and identified 11 prominent challenges. Wang et al. surveyed software testing with LLMs by analyzing 102 relevant papers and summarizing several challenges, such as the case coverage and real-world application [30]. Compared to their work, our study focuses on ADS testing by not only conducting literature reviews but also a large-scale survey.

Many literature reviews center around ADS testing reviews [11, 26,27,31], including scenario-based ADS testing [11], critical scenario identification and challenges [31]. Compared to these reviews, we conducted a comprehensive survey of the ADS testing and discussed the general testing challenges. Among those overall reviews, the work from Tang et al. presented a thorough literature review of ADS testing and identified the challenges and opportunities [27]. Compared to their work, our work invited 100 ADS testing practitioners or researchers to participate in our survey and follow-up open-ended questions to analyze the current practices and challenges of ADS testing. The reviews conducted by [26] also focus on ADS testing and are most related to our study. However, they do not discuss the V2X system and the utilization of FMs in ADS testing, which we include in our paper.

3. Methodology

Following practices in [26,32], our survey design contains three stages: **Discussion**, **Survey** and **Follow-up** as displayed in Fig. 1. We first discussed with four professionals in the ADS industry to initially understand the current practices and expectations for ADS testing. Then, we designed a comprehensive and large-scale survey of 100 ADS testing practitioners and researchers. For participants who meet our follow-up criteria, we further invited them to answer open-ended questions.

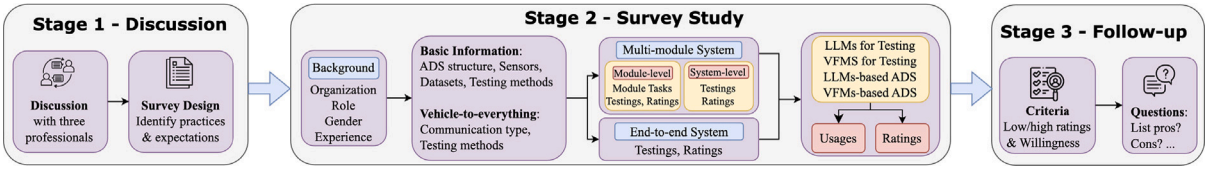


Fig. 1. The process of our survey.

Table 1
Professionals on survey design discussion stage.

Professionals	Organization	Role	Experience
P1	Company A	ADS Testing Engineer	5 years
P2	Company B	ADS Testing Engineer	3 years
P3	Research Institute A	ADS Software Engineer	4 years
P4	Research Institute A	ADS Testing Engineer	3 years

3.1. Discussion

We invited four professionals from two automobile companies and one research institute specializing in ADS to discuss current practices, challenges, and future opportunities in ADS testing. They bring extensive experience from collaborations with top-tier automobile manufacturers. Their roles and expertise are outlined in Table 1. Three are experienced ADS testing specialists, with one having five years and two with three years of involvement in ADS research and development. The fourth is a senior software engineer with four years of experience designing and optimizing ADSs.

The discussions took place across three sessions totaling six hours, where professionals provided in-depth analysis and perspectives on ADS testing. In the first session, we introduced our research motivation and then discussed current ADS testing practices, including sensors, datasets, V2X experiences, classical methods, relevant definitions, and their feedback on testing methods. In the second session, we explored tasks of modular and E2E ADSs, testing approaches, and the implementation of FMs [33] (i.e., LLMs and VFMs) in two aspects: their use in testing techniques and integration into ADSs. Afterward, we reviewed the literature to cover topics not addressed in the first two sessions, such as potential attacks and defenses. The third session focused on these missing aspects and summarized the future direction of ADS testing.

3.2. Survey

Survey design. Based on the discussion, we designed a survey with 107 questions divided into six sections: *Background*, *Basic Information*, *V2X Practices*, *Testing Practices*, *FMs Utilization*, and *Follow-up* (Fig. 1). The background section covered participants' organizations, roles, genders, and work experience in ADS testing. The second section focused on the ADS they work with, including sensors and datasets. The third section gathered V2X experience. In the fourth section, we introduced the multi-module and E2E ADS structures. Participants were asked to select the ADS they were most familiar with based on their experience. Those familiar with both structures were asked to choose one to continue the survey. After selecting the multi-module structure, participants were asked about module-level or system-level testing, with deeper questions on tasks and methods for testing.

Specifically, we asked participants to present their testing techniques, including black-box (model inputs and outputs) [34], white-box (full access) [35], and grey-box (partial access) [36]. Based on the discussion and literature [31], common methods for generating testing scenarios include knowledge-based [37], search-based [9], and data-driven [38]. Knowledge-based methods use expert knowledge, search-based methods employ algorithms (e.g., genetic algorithm [39]), and data-driven methods rely on real-world data. A more detailed literature review is provided in Section 6. Participants rated the methods on a 7-point Likert scale [40], from "Strongly disagree" (level-1) to "Strongly agree" (level-7). Those familiar with both module-level and system-level testing answered questions for both. Respondents with E2E experience were asked about sensors, output control actions, and testing ratings. The fifth section explored their experience with FMs, including usage and effectiveness compared to traditional methods. In the final section, participants could provide their email to receive study results and were invited for follow-up. "Other" options were included for each multi-choice question to capture diverse responses. Additionally, some survey questions allowed multiple selections to capture the real-world use of combined testing approaches. For example, participants could select more than one testing strategy (e.g., black-box, white-box, gray-box) or scenario generation method (e.g., knowledge-based, search-based, data-driven), reflecting hybrid practices commonly adopted in ADS testing.

Participants. We invited participants using two approaches: (1) personal networks and (2) emails crawled in GitHub. Specifically, our personal networks included individuals from both academia and industry, such as partner ADS research institutions and commercial companies. This allowed us to ensure a diverse pool of respondents in terms of organizational background and professional role. Ten pilot participants from these networks completed the questionnaire to help us identify and mitigate any sensitive or ambiguous questions.

To reach a broader audience, we further identified 48 keywords relevant to ADS testing and used GitHub's API to extract publicly available email addresses of contributors to related repositories. In total, we sent 5,200 emails and received 206 responses. However, due to the large scale of the survey, many responses were incomplete. Moreover, responses from professionals not involved in testing activities (e.g., vehicle dispatchers) were excluded.

The first two authors independently reviewed all responses and each selected 90 complete and high-quality answers. Among these, 88 responses overlapped, resulting in a 97.78% selection consistency. The final 90 responses were determined through discussion and consensus. Including the 10 pilot responses, we obtained 100 valid and complete survey responses for analysis. Regarding participant demographics, the majority (59%) are from research institutes, with research scientists (43%) being the largest professional group. The male-to-female ratio is approximately 8:2, and 73.3% have 1 to 5 years of experience in ADS testing. The respondents include both researchers and practitioners, allowing us to compare perspectives across different roles. We explicitly differentiate their responses throughout our analysis to highlight the varying concerns and testing priorities from both communities.

Additionally, participants span eight countries, with the majority from China (21 participants), the USA (17), and Japan (12). These respondents were primarily recruited through our professional network, which includes collaborators from both academic institutions and industry organizations in these countries. Together, they represent over half of our sample and cover key ADS markets with diverse regulatory and deployment paradigms: China and the USA emphasize rapid field application and commercialization, while Japan and several European countries prioritize safety and cautious deployment strategies [41]. Although our survey may not fully reflect every global regulatory environment, the inclusion of participants from these major regions ensures

that our findings capture broadly relevant practices and challenges in ADS testing.

Data Analysis. In this study, we counted and summarized all the data collected in the survey in the form of text or graphs. In the survey, we placed "Other" options for each multiple-choice question to prevent the provided options from not covering the participants' experience. We also analyzed this part of the alternative answer. If the answer belonged to a given option, it was correctly categorized. Otherwise, it was briefly introduced when summarizing the relevant questions. We analyzed participants' demands in linear-scale ratings by comparing the rating distributions.

3.3. Follow-up

We conducted follow-up surveys of participants with the following criteria: (1) rating low/high on the 7-point Likert scale, (2) willing to attend the follow-up stage. There were 27 participants in our follow-up phase, and we developed multiple targeted open-ended questions for each of them according to their responses. For instance, *Please specify why the current testing approaches cannot meet the testing requirements*. Regarding the data analysis in this stage, the first two authors categorized and summarized all open-ended questions collected separately to identify current demands further. The results were then discussed with each other to refine the results.

3.4. Literature review

We follow the literature review guidelines proposed by [42] to ensure a comprehensive and structured exploration of relevant research. Our literature search strategy consists of three main methods: keyword search, citation tracking, and journal browsing.

For the keyword search, we carefully designed 52 keywords based on a preliminary review of ADS testing research and relevant survey papers (e.g., autonomous driving testing, E2E testing, online testing, CARLA, etc.). The keyword list was iteratively refined to ensure broad coverage of relevant studies. The final set of keywords was combined using Boolean operators such as "OR" and "AND" to expand coverage, and adapted for use in major databases including IEEE Xplore, ACM Digital Library, SpringerLink, and ScienceDirect. Search results were first automatically filtered by removing duplicates and unrelated domains, followed by manual screening based on titles and abstracts. Where ambiguity remained, the full-text review was conducted. To reduce bias, two authors independently screened each candidate paper, resolving disagreements through discussion. This ensured consistency and minimized selection bias in the final corpus of reviewed literature.

We then employed citation tracking, which includes both forward and backward tracking. Forward tracking identifies papers that have cited a given study, helping us trace its impact and subsequent advancements, while backward tracking involves reviewing the reference lists of key papers to identify foundational and related works. To ensure the robustness of our selection, we prioritize highly cited papers and those published in premier SE, ADS, and AI venues. Forward tracking is conducted using Google Scholar, while backward tracking is performed manually. To further refine our selection, we conduct journal and conference browsing. In particular, we focused on software engineering journals and conferences (TSE, TOSEM, IST, ICSE, ESEC/FSE, ASE, ISSTA, ICST, etc.), and ADS and AI-related venues (TITS, TIV, IV, AAAI, NeurIPS, CVPR, etc.). We prioritized recent papers from the last five years while including seminal works significantly shaping ADS testing research. A detailed discussion of these contributions, existing challenges, and research gaps is presented in Section 6.

4. Common practices of ADS testing

This section provides an overview of ADS testing practices used by industry and academia. Based on our survey results, academic researchers tend to focus on innovation and novel techniques, while industry practitioners are more concerned with system dependability, robustness, and practical deployment. Reflecting these differences, we examine the testing practices of two types of SUT: multi-module and E2E ADSs, followed by a review of the sensors involved. We then examine testing practices for single-agent ADSs, including sensor testing, and analyze multi-module systems with module-level and system-level testing. We also cover E2E system testing approaches and focus on V2X communication testing for multi-agent and E2E systems. Finally, we highlight the use of FMs in ADS testing and their impact on system performance.

4.1. Autonomous driving systems under test

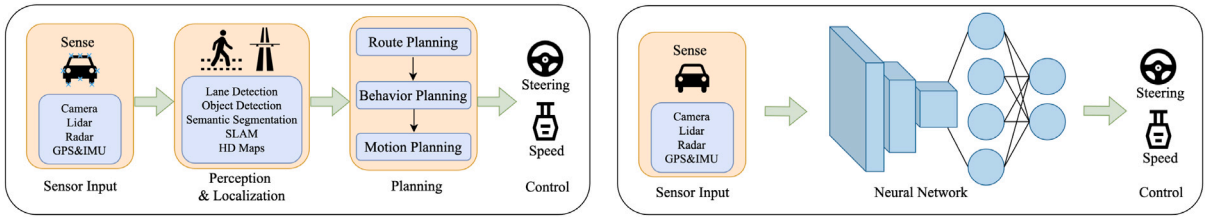
Participants in our survey work on two types of ADSs: multi-module and E2E. In this section, we first introduce the definition and structure of these two ADSs and then analyze the survey results, followed by the utilization rate of sensors.

Multi-module ADSs: Multi-module ADSs use a modular architecture, decomposing the system into interconnected components, each dedicated to a specific task. This structure is shown in Fig. 2(a), enables flexibility, scalability, and easier debugging compared to monolithic designs. Typically, a multi-module ADS includes sensing, perception and localization, planning, and control modules. The Sensing module collects raw data from sensors like cameras, LiDAR, Radar, and GPS to understand the vehicle's surroundings. The Perception and Localization module processes this data to detect road objects (e.g., vehicles, pedestrians) and localize the ego vehicle. The Planning module then generates an optimal driving trajectory considering road constraints, traffic rules, and dynamic obstacles. Finally, the Control module converts the planned trajectory into vehicle control commands (e.g., steering, throttle, braking) to execute the driving behavior.

Regarding the distribution of models used in ADSs, participants were allowed to select multiple techniques used in their systems, as real-world ADS modules often integrate several approaches simultaneously. Although DL, reinforcement learning (RL), and traditional ML are all categorized as subfields of ML, we analyze them separately in our survey. Because each has distinct application patterns in ADS testing, and they can also be used in combination (e.g., deep reinforcement learning). To capture this flexible and often overlapping usage, we designed the question as a multiple-choice item. This allows respondents to report all applicable methods and enables to analysis of the adoption trends and co-occurrence patterns of these techniques.

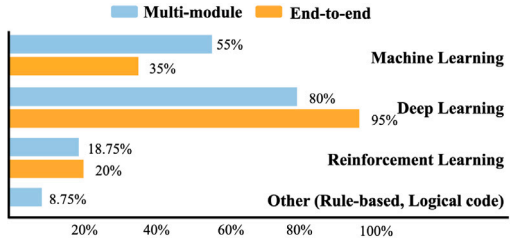
Our survey results show that 80% participants work on multi-module ADSs, with 45% from research institutes and 35% from companies, indicating strong adoption in both academia and industry. As shown in Fig. 3(a), 64 respondents integrate DL into their modules (45% industry practitioners), while 55% use ML techniques, of which approximately 58% are academic researchers, including supervised and unsupervised learning models. RL is used by 18.75% of respondents, half of whom are practitioners. Only 8.75% use rule-based or logic-based systems, mostly from industry and confined to Planning and Control modules. The rationale behind this trend is discussed in Section 4.2.1, where we analyze the role of hybrid approaches in enhancing ADS performance.

End-to-end ADSs: The E2E system integrates all modules into a single DNN, mapping sensor data directly to control actions, as shown in Fig. 2(b). Compared to multi-module ADSs, E2E systems have a simpler structure and rely more on data-driven learning, reducing the need for manual rules or feature engineering. Most E2E systems use imitation learning, where DNNs are trained on expert datasets to mimic human

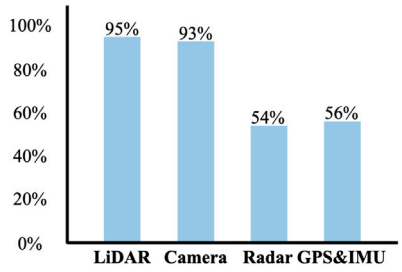


(a) Multi-module system. (b) End-to-end system.

Fig. 2. The structure of the multi-module system and end-to-end system.



(a) Reported usage of learning techniques in ADS testing (ML, DL, and RL are shown separately to reflect their distinct applications and overlapping combinations in practice).



(b) Distributions of sensors in ADSs.

Fig. 3. The survey results related to multi-module and end-to-end systems.

driving behavior [8]. Despite their simplicity, E2E ADSs face challenges in generalization and interpretability, as decision-making is embedded within the network's representations, complicating debugging and failure analysis. Our survey shows that 20% of participants work on E2E ADSs, much lower than the 80% adoption rate of multi-module systems. Among E2E users, 75% are from research institutes, and 95% use DNN-based architectures, while ML and RL are only used by academic researchers, with 35% and 20% adoption rates, respectively.

Given the heavy reliance on DL models in both multi-module and E2E ADSs, safety concerns, particularly adversarial attacks, have gained attention in autonomous driving [43]. Among survey participants, 54% have conducted adversarial testing, with 87.04% using defenses like adversarial training and model regularization. In addition to adversarial robustness, privacy vulnerabilities in ML-based ADSs are emerging as a concern. One expert noted that ML-based ADSs are susceptible to inference attacks, such as membership and attribute inference [44], which could allow adversaries to infer sensitive user data like locations, restaurants, or driving habits from sensor data.

> ***Practice 1***: The majority of participants use DL/ML-based ADSs as system components under test, which are 87.50% and 45% on average, respectively, where the adversarial attack is a common threat.

All of the professionals expressed their preference for LiDAR and cameras, which can provide precise 3D point clouds and abundant visual information. While radar and GPS&IMU contribute to specific scenarios, such as collision warning systems, they have limitations in perception and semantic information.

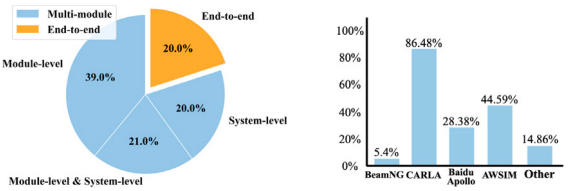
Fig. 3(b) shows the distribution of sensor types used in ADSs, based on participants' multiple-choice responses. LiDAR and cameras are the most widely adopted, with utilization rates of 95% and 93%, while Radar and GPS & IMU are used in approximately 50% of systems. The sensor utilization is similar among industry practitioners and academic researchers. All industry professionals preferred LiDAR and cameras for their high-resolution depth perception and rich visual information, crucial for object detection, localization, and scene understanding. Radar and GPS & IMU serve complementary roles, where Radar is valued for its robustness in adverse weather and its ability to measure object velocity, while GPS & IMU are essential for global localization and vehicle odometry in GPS-denied environments. However, Radar and GPS & IMU have limitations in perception, resulting in lower utilization rates. Sensor fusion is a key trend, and integrating these sensors effectively remains a research challenge to improve system robustness across diverse driving conditions.

4.2. Testing for single-agent ADSs

Next, we focus on single-agent ADS testing. We first summarize the testing distribution of single-agent ADS, followed by presenting the sensor-related findings, and then analyze the testing methods in further detail by dividing multi-module into module-level and system-level testing. Finally, we discuss testing issues related to the E2E system.

Fig. 4(a) shows the distribution of testing practices in single-agent ADSs, where multi-module ADSs account for 80% of participants, with 36% from industry. 39% focus on module-level testing (over half from academia), and 21% conduct both module and system-level testing. We found 52.73% of industry practitioners focus on real-world testing, while 62.30% of academic researchers use simulation platforms, likely due to cost and safety constraints in real-world testing. Fig. 4(b) shows CARLA [13] is the most popular simulation platform, praised for its flexibility, high-fidelity urban scenarios, and open-source nature. Since the question allowed multiple selections, 74% of participants worked on online testing, while 55% conducted offline testing, with more industry practitioners preferring offline testing and more researchers using online testing for cost reasons. Professional discussions highlighted that system-level testing emphasizes cooperative behavior, requiring online testing for real-time evaluation, while offline testing is used for controlled experiments and reproducible evaluations.

Testing on Sensors. 27% of participants reported conducting sensor testing, with physical attacks like jamming and spoofing posing significant risks to ADS perception, leading to sensor failures or misdetections. To mitigate these risks, 85.19% of respondents, both from industry and academia, utilize multi-sensor fusion and anomaly detection. Multi-sensor fusion enhances robustness by cross-validating data from multiple sensors, while anomaly detection flags or discards corrupted sensor data. Additionally, 30% of participants use adversarial training, primarily among researchers (59.52%), to improve model robustness against adversarial sensor data. Beyond algorithms, 18.5%



(a) Distribution of single-agents. (b) Distribution of simulations.

Fig. 4. The survey result of single-agent testing in ADSs.

have implemented physical protection measures, such as shields, to safeguard sensors from interference and tampering. These findings highlight the importance of combining software-based defenses with physical protection to ensure sensor reliability and ADS perception integrity.

> ***Practice 2***: The attacks (jamming, spoofing) of sensors have common defenses, including multi-sensor fusion (85.19%), anomaly detection (85.19%), adversarial training (30%), and physical protection (18.5%).

4.2.1. Multi-module system.

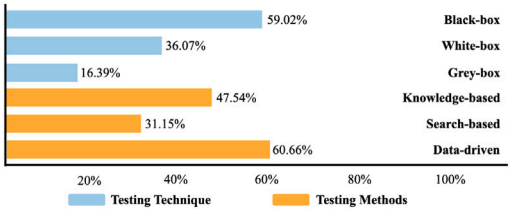
In this section, we present the multi-module ADS, where functional components are separately designed and tested. We discuss the corresponding testing methodologies, including module-level and system-level testing. Module-level testing ensures individual components function correctly in isolation, while system-level testing evaluates the entire ADS's performance, ensuring seamless integration of all modules in real-world scenarios.

Module-level Testing. There are 75% participants who conduct module-level testing in the multi-module system, of which 25% are industry practitioners. Among these participants, 70% are testing the perception module, 65% are focusing on the planning module, and 26.67% working on the control module.

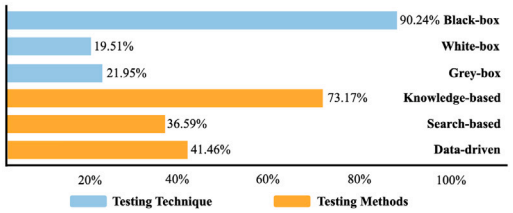
In our survey, 95.24% of perception modules utilize DL models for downstream tasks (e.g., object detection), with 9.52% (mainly industry practitioners) using rule-based, scan-matching, or classical algorithms. Among the respondents, 76.19% focus on object detection and tracking, while semantic segmentation accounts for around 20%, highlighting object detection as the foundation of ADS perception. In the planning module, ML-based and DL-based approaches each make up about 40%, with RL-based approaches at 12.82%. The "Other" option includes rule-based, and the state machine also occupied 12.82%, indicating that hybrid approaches combining classical and learning-based methods are still prevalent. The most common tasks are path planning (84.62%) and behavior planning (58.94%). In the control module, the majority of participants (62.5%) shared the classical algorithms they used from both academia and industry, such as proportional integral derivative (PID), model predictive control (MPC), or private algorithms. The specific practices we investigated found that all of the control module respondents engaged in steering control, in which 87.5% and 56.25% of them also worked on speed and stability control.

The distribution of testing approaches in module-level testing and system-level testing is shown in Fig. 5, both of which are multiple-choice questions for reflecting real-world testing practices, where different techniques are often used in combination. Not only reveals the popularity of individual methods but also captures how they are jointly applied in practice, offering a more delicate view of ADS testing strategies.

According to Fig. 5(a), black-box testing is more commonly applied at the module-level testing (60%), likely due to its ability to evaluate functional correctness and system behavior without requiring internal



(a) Distribution of testing approaches in module-level testing.



(b) Distribution of testing approaches in system-level testing.

Fig. 5. The survey result of testing approaches in single-agent systems.

model access, making it suitable for testing individual components in isolation. Data-driven testing is also prevalent, with 60.66% of participants adopting this approach, indicating a preference for testing in real-world scenarios or realistic simulation environments to evaluate module performance under diverse and dynamic conditions. Among the three testing approaches, knowledge-based and data-driven methods receive equal attention from both industry and academia, and search-based testing is more prevalent in research institutes.

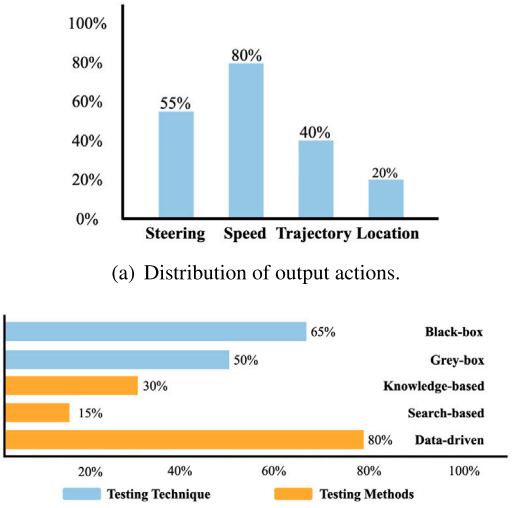
System-level Testing. In system-level testing, black-box testing is the most commonly used method, with a usage rate of 90.24%, as shown in Fig. 5(b). This method evaluates the overall performance of ADSs without focusing on the internal details of individual modules. The knowledge-based approach is also widely used, with a rate of 71.79%, of which 60% are from industry. Industry professionals emphasize that system-level testing involves evaluating the interaction and integration of multiple modules, which introduces complexity. The knowledge-based method is particularly useful for designing critical scenarios that reflect real-world challenges, such as unexpected road conditions, multi-agent interactions, and rare corner cases. It leverages domain knowledge and predefined rules to explore inter-module dependencies, ensuring safe and reliable operation in diverse environments.

> ***Practice 3***: The module-level testing is more flexible in multi-module systems, and both module-level and system-level testing tends to black-box testing, which is 74.63% on average.

4.2.2. End-to-End system.

In this section, we present the findings of the E2E system. We introduce the E2E system first, summarizing the output actions utilized by respondents and the testing methodology. The questions in this section also allow participants to select multiple answers.

The E2E system is an architecture that maps sensor data directly to outputs (e.g., steering, speed, trajectory, etc.). One professional specializing in E2E systems noted that although E2E is less widespread in ADSs, it reduces system complexity compared to multi-module architectures due to fewer interface and integration points. However, we observed diverging views based on participants' roles. Researchers and algorithm developers often view E2E as simpler, given its streamlined pipeline and lack of inter-module dependencies. In contrast, deployment-focused practitioners emphasized that E2E systems may be more complex in practice. Specifically, the opacity and non-interpretability of DNNs make validation, debugging, and compliance



(a) Distribution of output actions.

(b) Distribution of testing approaches.

Fig. 6. The survey result of end-to-end system in ADSs.

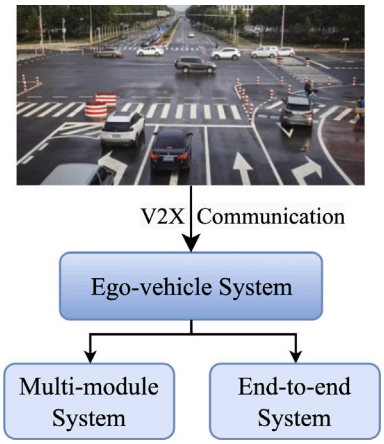


Fig. 7. V2X structure.

verification more challenging. This contrast highlights how professional backgrounds shape perceptions of architectural trade-offs in ADSs.

Our survey results show that E2E testing mainly focuses on key sensors and output actions, as seen in Fig. 6(a). LiDAR (70%) and cameras (70%) are the most commonly tested sensors, while Radar (30%) is used less due to its lower spatial resolution. On the output side, speed control (80%) is the most frequently tested function, followed by steering (55%), trajectory generation (40%), and location tracking (20%), which reflect core vehicle motion control tasks.

Regarding testing methodologies, our findings (Fig. 6(b)) show that E2E testers employ both black-box (65%) and gray-box (50%) testing approaches. Black-box testing is useful for evaluating overall system behavior without internal access. While some participants reported applying gray-box testing, this may reflect efforts to improve transparency through techniques such as model interpretability tools, intermediate signal analysis, or limited observability into DNN structures. For example, some practitioners may use feature attribution methods like SHAP to understand model sensitivity to specific input features [45,46]. Others might insert diagnostic checkpoints at intermediate layers of the DNN (e.g., activations from the perception module or intermediate driving intention vectors) to analyze whether internal representations behave as expected during simulation. However, the scope and depth of such testing are still limited compared to conventional gray-box testing in traditional software systems.

Notably, no participant reported using white-box testing, due to E2E systems involving integrated components where the system's internal workings (such as DNNs and fused sensor data) are not directly accessible. We acknowledge that directly applying white-box or even gray-box testing techniques to neural networks remains an open challenge in the field, and interpretability research may play a crucial role in enabling such testing in the future.

Additionally, the focus is often on testing the system's external behaviors and interactions rather than its internal workings. Among testing techniques, data-driven testing is the most widely used, as illustrated in Fig. 6. Respondents emphasize the importance of simulating real-world environments and leveraging historical driving data to generate diverse testing scenarios that closely resemble actual deployment conditions.

4.3. Testing for V2X communication

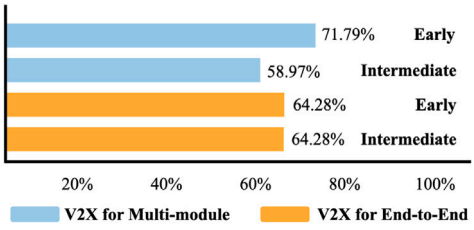
Following the discussion on single-agent ADS testing, we now turn to the testing of V2X communication systems. V2X communication is essential for enabling effective interactions between ADSs and other vehicles or infrastructure. In this section, we explore how V2X testing varies between multi-agent systems and E2E systems. We begin by defining V2X and then present the survey results regarding the use and testing of V2X approaches across different ADS types.

V2X is a wireless communication technology that enables data exchange between vehicles and external entities, such as other vehicles (V2V), infrastructure (V2I), pedestrians (V2P), and networks (V2N), with the primary goal of enhancing traffic safety and optimizing transportation efficiency [16]. By sharing real-time information, V2X improves situational awareness, cooperative decision-making, and collision avoidance, making it vital for the development of ADSs. These benefits drive the need to explore its current use and testing in real-world applications.

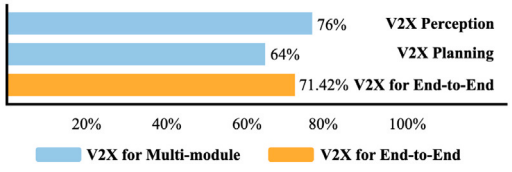
The structure of V2X is shown in Fig. 7. A key aspect of V2X-enabled ADSs is data fusion, which integrates information from multiple vehicles and roadside sensors. Data fusion in V2X is categorized into three levels: *early*, *intermediate*, and *late* fusion [47]. *Early* fusion combines raw sensor data (e.g., LiDAR, camera images) from multiple sources before feature extraction, requiring high bandwidth and computation. *Intermediate* fusion uses processed feature representations, while *late* fusion integrates final predictions (e.g., detected objects, planned trajectories), minimizing communication overhead but limiting cross-vehicle feature learning.

The questions in the V2X section are multiple-choice questions. Our survey shows that common V2X datasets include V2X-Sim [48] and V2X-Seq [49], which feature communication scenarios between vehicles and infrastructure. Additionally, 36.36% of V2X users reported conducting cyber security evaluations to assess communication safety and data integrity, mainly in simulated environments. These evaluations focus on threats such as message spoofing, jamming, and data injection. Defense mechanisms include encrypted communication, security protocols, identity authentication, and intrusion detection systems.

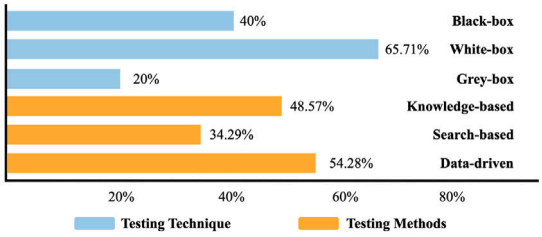
V2X for Multi-module Agents. According to our survey, 48.75% of participants have used V2X in multi-module ADSs, with 60% from academia, highlighting its growing role in collaborative perception and decision-making. Regarding data fusion, 71.79% transmitted raw data, 58.97% used intermediate features, and none used *late* fusion (Fig. 8(a)). Specifically, 60% of those using *early* fusion are from research institutes, while *intermediate* fusion is equally favored by both industry and academia. The preference for early and intermediate fusion reflects the need for rich perception and better scene understanding, while the absence of late fusion suggests a focus on lower-latency, high-fidelity fusion to enhance real-time performance.



(a) Distribution of data fusion.



(b) Distribution of V2X testing.



(c) Distribution of V2X testing approaches.

Fig. 8. The survey result related to V2X communication.

From our professional discussion, we identified perception and planning as the primary ADS modules utilizing V2X communication. Among V2X users in multi-module systems, 64.10% reported conducting V2X-specific testing, with 36% from industry (Fig. 8(b)). Specifically, 44% of academic researchers and 32% of industry practitioners tested V2X-enhanced perception, while 44% of researchers and 20% of practitioners tested V2X-based planning. No participants tested V2X integration in other modules, highlighting its key role in environmental understanding and decision making.

V2X for E2E agents. In the E2E system, half of the participants stated they used V2X communication, where participants used *early* fusion and *intermediate* fusion are both 64.28%, which is exhibited in Fig. 8(a). The preference for these fusion strategies aligns with the nature of E2E systems, which rely on rich sensory data and learned feature representations for autonomous decision-making. Regarding testing-related findings, according to Fig. 8(b), 71.42% of V2X users in E2E systems have conducted testing, of which 64.29% are academic researchers. Our survey shows that 100% of participants tested LiDAR point cloud transmission, making it the most widely adopted V2X-perceived modality. 80% utilized RGB images, highlighting the importance of visual perception, while 40% incorporated radar point clouds, with half from industry, likely due to their robustness in adverse weather conditions.

The testing techniques result is shown in Fig. 8(c). Black-box testing is the most prevalent (40%) used by all E2E testers for overall system evaluation without internal access. 65.71% of multi-module V2X testers use white-box testing to analyze module interactions and verify logic compliance. No participant conducts white-box testing in E2E ADSs because it requires deep knowledge of the internal structure of the system, which may not always be feasible or practical in V2X systems due to the complexity of interactions between various vehicle models, sensors, and communication networks. Additionally, 20% of V2X testers have conducted gray-box testing, the majority of whom are affiliated with research institutes. This type of testing is primarily employed for assessing model interpretability and robustness. In certain cases, gray-box testing also encompasses partial inspections of communication modules or sensor fusion outputs.

In terms of test generation strategies, data-driven (54.28%) and knowledge-based (48.57%) are the most frequently employed. Besides, the use of knowledge-based and data-driven approaches is evenly distributed between industry and academia, while search-based methods are predominantly adopted by academic researchers (75%). On the basis of the professionals' statements, the data-driven method can meet the requirements of V2X systems that need to process large volumes of real-time data, and the knowledge-based method contributes to infrequent corner cases, such as edge cases in cooperative perception, or communication failures.

> ***Practice 4***: Most participants leverage early data fusion in V2X communication and conduct testing on perception and planning modules by leveraging the black-box testing technique (65.71%) and the data-driven method (54.28%) most frequently.

4.4. Utilization of FMs

This section introduces LLMs and VFMs, discusses FMs' utilization for testing, and also slightly introduces survey results on FMs-based ADS, which use VFMs and LLMs for tasks like scenario generation, test case evaluation, and system validation. Survey questions in the FMs sections also allow participants to select multiple choices to share their experiences.

LLMs are rapidly growing DL-based NLP models with powerful generative capabilities in autonomous driving [50]. Traditional ADS testing approaches struggle with the diversity of real-world driving scenarios, but FMs, especially LLMs, can generate diverse test scenarios to improve test coverage and adaptability [12,51]. Besides, VFMs are models for describing and modeling vehicle characteristics, which are utilized broadly in ADSs [52]. For instance, ADS perception models are highly susceptible to environmental variations (e.g., lighting changes, occlusions, adversarial perturbations), and VFMs can provide a robust way to simulate complex sensor inputs, allowing for stress testing of ADS perception modules under diverse and adversarial conditions. Additionally, they enable data augmentation for synthetic dataset generation, helping improve model generalization and robustness [24].

FMs for Testing ADSs. According to the discussion results, all four professionals have relevant experience in testing with FMs, and one has both experience using LLMs-based and VFMs-based ADSs. Professionals stated that LLMs and VFMs both contribute to generating critical scenarios in testing, while VFMs can help retrain by providing real-time feedback from the vehicle.

Survey results show that 70% of respondents are from research institutes. 26.25% use LLMs in multi-module ADS testing, while 12.5% use VFMs, as shown in Fig. 9(a). Among module-level testers, 63.64% focus on perception, and 45.45% on planning. The lower adoption of VFMs suggests their strength lies in system optimization rather than module-level testing. VFMs are primarily used for camera-based perception systems, where they model scene variations and generate perturbations. However, radar and LiDAR-based systems, which rely on different principles, may need alternative approaches for testing.

In the E2E system, we found that using VFMs is more frequent (25%) than LLMs. Unlike multi-module ADSs, where decision-making is modularized, E2E architectures require rapid, continuous feedback for real-time adjustments, making VFM-driven optimization particularly effective in enhancing response time and control stability.

Additionally, from Fig. 9(b), participants are more likely to choose LLMs to generate testing scenarios (67.86%) due to their ability to synthesize complex environments, while VFMs are preferred for retraining (60.71%). This aligns with the observation that LLMs are

well-suited for generating diverse, scenario-based tests, whereas VFMs are more effective for refining real-time feedback loops. The differences in sensor modality and testing target suggest the need for a more tailored approach when selecting FMs for ADS evaluation.

FMs-based ADSs. LLMs-based ADSs utilize LLMs for tasks like decision-making, scenario understanding, and human-machine interaction, while VFMs-based ADSs improve perception, scene recognition, and environmental understanding. LLMs focus on reasoning and language tasks, while VFMs enhance visual data interpretation. Emerging models like CLIP [53], when combined with Vision Transformers (ViT) and CNNs, enhance tasks like semantic segmentation and object recognition, crucial for ADS perception. Our survey showed 10% of participants have integrated FMs into ADS workflows, with only 2% using them as a core component, likely due to computational and adaptation challenges. Respondents also reported using SAM [54] to improve semantic segmentation in ADS perception.

> ***Practice 5***: Participants tend to employ LLMs to generate diverse and complex testing scenarios in multi-module and utilize VFMs for retraining in E2E systems.

4.5. Hybrid testing practices

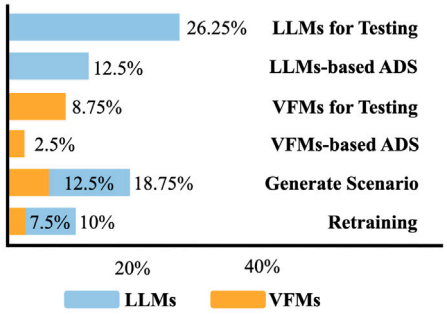
Across the ADS testing survey results, participants often employ a combination of testing techniques in ADS testing, which is reflected in the multi-selection design of our survey question. This design choice was intentional, as it mirrors real-world practices where different methods are combined to meet varied testing goals. For instance, combining black-box and white-box testing allows testers to verify both system behavior and internal logic. Similarly, integrating knowledge-based and data-driven methods helps cover domain-specific edge cases and statistically representative scenarios. These hybrid strategies highlight the practical demands of ADS testing, where achieving comprehensive test coverage often requires the complementary strengths of multiple approaches.

Additionally, we observe that even among researchers, testing practices may vary based on their roles. Those focused on developing novel techniques for future ADSs tend to experiment with exploratory or AI-driven testing approaches, while those responsible for evaluating deployed systems prioritize robustness and regulatory compliance. This variation further supports the adoption of hybrid testing practices tailored to different development stages and responsibilities.

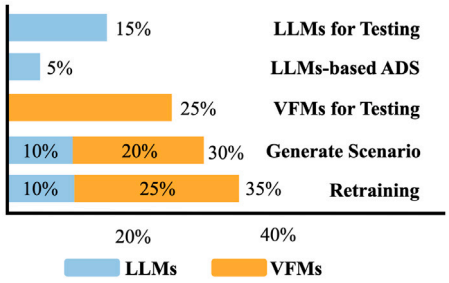
5. Demands

In this section, we summarize 7 demands of ADS testing, including 1 V2X-related and 2 FM-related demands. We analyze demands according to the ratings and the follow-up open-ended questions. The rating result of current testing approaches of single-agent ADSs is illustrated in Fig. 10(a), which contains a 7-level rate from strongly disagree to strongly agree. According to the result, at the module-level testing, the perception module and the control module received the most and the least positive ratings, 79% and 63%, respectively, while 2% and 12% of the participants expressed disagreed that the current testing meets the requirements (e.g., safety, robustness, real-time response, user experience), respectively. Besides, no one fully agreed that the current testing of the control module fulfilled the requirements. In the system-level testing, only 5% of the participants indicated test methods sometimes failed to meet the testing requirements. The E2E system received the most negative voices (20% sometimes disagree, at level-3).

In summary, 13% of the participants strongly disagreed or disagreed that their current testing approaches meet their testing requirements. We designed follow-up questions to investigate further (e.g., *Please specify why the current testing approaches cannot meet the testing requirements*),



(a) Distribution of using FMs for testing in multi-module system.



(b) Distribution of using FMs for testing in end-to-end system.

Fig. 9. The survey result of FMs in ADSs.

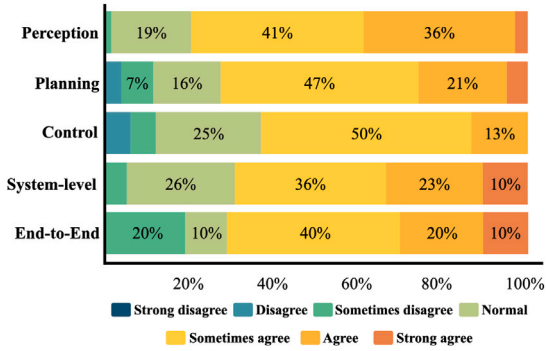
and the response rate reached 84.61%. 30% of the participants take a neutral stance on the testing approaches they leverage, the follow-up questions including *Please briefly list the advantages and disadvantages of the current testing approaches*, and received an 80% response rate. 64% of the participants sometimes agree on the effectiveness of current testing approaches. The open-ended questions contain *Please describe the shortcomings of the current testing approaches*, with a 60.94% response rate.

For those respondents with experience with FMs, we asked two types of questions: "*Compared with other testing methods, the introduction of LLMs into ADS testing improves the testing capability. (Strong disagree to strong agree)*" and "*Compared with other ADS, the introduction of LLMs-based ADS improves the performance. (Strong disagree to strong agree)*". The rating result is displayed in Fig. 10(b). Overall, both LLMs for testing and LLMs-based ADSs were rated better than ADSs involving VFMs. The average percentage of those who indicated agreement in LLMs-related usage amounted to 65.50%.

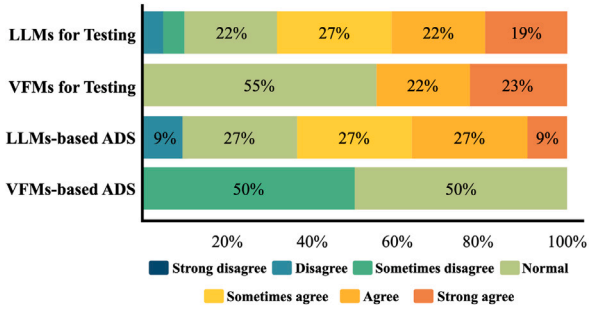
In general, we analyze the demands for both industry and academic researchers separately, and we totally summarized 7 demands: (1) diversity in corner cases, (2) bridging gaps between simulation and real-world testing, (3) more comprehensive testing criteria, (4) effective defenses against current attacks, (5) seamless cross-model collaboration in V2X systems, (6) Test case quality generated by LLMs and (7) reliable cross-modality integration when using FMs.

5.1. Demand 1: Diversity in corner cases

Corner cases are extreme or rare driving scenarios under which the ADSs may make wrong decisions due to improper responses, leading to severe traffic accidents [55]. From the follow-up result, 58.90% responses have mentioned the lack of diversity of corner cases. Among them, 53.49% of participants are from industry, and 46.51% are from research institutes. This demand is primarily put forward by perception module testers and E2E testers, who rely on diverse and realistic edge-case scenarios to evaluate model reliability. Practitioners focus



(a) Ratings for current testing approaches of single-agent ADSs.



(b) Ratings for LLMs/VFMs testing approaches and LLMs/VFMs-based ADSs.

Fig. 10. The rating result of current testing approaches.

on constructing challenging real-world scenarios, while academia may explore methods for automatically generating or identifying new corner cases. Some responses said that *"Lack of complex corner cases, like a vehicle driving at high speed in heavy fog at night, suddenly encounter a black animal crossing the road."* and *"The current corner cases are limited, especially when combined with different extreme situations, such as pedestrians suddenly appearing to cross the road in blizzard weather."* Both of these statements are sophisticated corner cases that are currently lacking, particularly those involving adverse weather conditions, poor visibility, dynamic obstacles, and multi-agent interactions. Therefore, the diversity of corner cases is one of the crucial bottlenecks in ADS testing.

5.2. Demand 2: Bridging gaps between simulation and real-world testing

50.68% of the follow-up responses (including 21.92% from industry) indicated that simulation environments may fail to fully capture real-world conditions (e.g., lighting variations, sensor noise), potentially causing performance discrepancies when transitioning to real-world deployment. Practitioners are primarily concerned with the discrepancy between simulation and real-world testing, and researchers focus on building higher-fidelity simulation models to bridge the gap between simulation and reality. A key issue raised by many practitioners is the difference between open-loop and closed-loop testing, which affects how ADSs interact with dynamic environments. Closed-loop testing enables real-time feedback, where system outputs influence subsequent inputs, allowing the ADS to dynamically adjust its decisions based on evolving perception data. In contrast, open-loop testing lacks feedback, meaning that perception errors cannot propagate to downstream modules, making it difficult to evaluate the full impact of sensor inaccuracies or decision-making flaws [56]. One response said, "*The open-loop testing lacks feedback, so it is hard to measure the perception error on downstream modules.*"

The professionals argue that both open-loop and closed-loop testing are conducted in simulation environments, but their purposes differ, where open-loop testing is primarily used for evaluating individual components (e.g., perception accuracy), whereas closed-loop testing is crucial for assessing system interactions and decision-making under dynamic conditions. While closed-loop testing is often considered more representative of real-world behavior, most practitioners still rely on simulation-based closed-loop testing before real-world deployment due to the high cost and complexity of physical testing [57]. In this demand, 29.73% of participants emphasized "*the lack of dynamic simulation*", reflecting concerns that existing simulation environments may not fully support closed-loop evaluations. Additionally, some practitioners who worked on control module testing claimed that the dynamics modeling needs to be more precise.

5.3. Demand 3: More comprehensive testing criteria

Our follow-up survey results reveal that 30.14% (with 19.18% of researchers) of participants identified the lack of comprehensive testing criteria as one of the demands, especially in system-level testing and E2E system testing, which both evaluate the performance of the entire ADS. Researchers are more focused on developing comprehensive evaluation systems, such as new metrics, to improve test coverage and interoperability. Unlike module-level testing, in which specific components can be independently validated, system-level and E2E testing suffer from poor interpretability, making it difficult to pinpoint the root cause of failures. A follow-up response has mentioned that "*We can acquire outcomes easily when conducting system-level testing, but it is hard to find what part went wrong.*" This highlights a critical limitation that while system-level testing can indicate ADS performance, it often fails to provide actionable insights into which specific components contributed to errors.

To address this challenge, researchers have introduced quantitative metrics to evaluate ADSs, such as Vulnerable Road User (VRU), Operational Design Domain (ODD), Conflict Area (CA), etc. [58]. However, these metrics do not always correlate with real-world driving quality. One practitioner pointed out this limitation, stating that "*A good metric does not equal a good driving performance. For example, an ADS that shows a low collision rate in testing as it chooses too conservative a driving strategy, such as braking hard frequently, may lead to rear-end collisions or impact on user experiences.*" Hence, the comprehensive testing criteria is one of the demands for ADS testing.

5.4. Demand 4: Effective defenses against current attacks

26.03% of follow-up responses came up with concerns about safety by referring to various current attacks, including adversarial, sensor, cyber, and inference attacks. Out of these participants, 84.21% are conducting academic work, indicating that the industry has fewer concerns about ADS attacks. Practitioners are concerned with deploying defenses in real-world systems, such as adversarial training and encrypted communication. In this demand, 15.79% of participants claimed the lack of polygonal defenses, and one of them said, "*Autonomous vehicle security not receive enough attention (maybe only until the accident happens).*" This reflects a reactive rather than proactive approach to ADS security, suggesting that current defenses are often implemented only after vulnerabilities have been taken advantage of in real incidents, rather than being integrated into the system from the outset. Regarding specific attack vectors, 73.68% of practitioners emphasized adversarial attacks, which exploit small perturbations in sensor inputs (e.g., images, LiDAR point clouds) to mislead ADS perception models. While adversarial training is a common defense, many practitioners noted that it requires extensive computational resources and can introduce latency, which is often unacceptable in real-time, safety-critical ADS operations. Besides, sensor attacks obtained 52.63% attention, in which the main focus is jamming (e.g., sending interference signal), spoofing (e.g., sending forging signal), and physical damage attack.

Moreover, 26.32% of responses reported that most of the cyber attacks involved the V2X system, "*In ADSs, some communication paths are unique, and a single point of failure for communications makes the vehicle unable to obtain the real-time data.*" Additionally, 10.53% of participants mentioned inference attacks in ADSs, and one of the statements said that "*In collaborative perception scenarios, such as V2X system, real-time data sharing between vehicles may disclose sensitive information.*" This reflects the growing challenge of privacy-preserving cooperative ADS perception, where ensuring effective data sharing while mitigating privacy risks is an open research problem. The prevalence of these security concerns highlights the urgent demand for stronger, multi-layered defense mechanisms in ADSs.

5.5. Demand 5: Seamless cross-model collaboration in V2X systems

V2X ADSs enhance autonomous driving by enabling vehicles to share real-time data and contribute to traffic efficiency and safety. However, current V2X communication among different vehicles has an inevitable challenge in seamless cross-model collaboration, where autonomous vehicles equipped with different DL models and sensors cannot effectively communicate and make joint decisions due to compatibility issues.

Almost all V2X users in our survey identified model compatibility as a major barrier to the widespread deployment of V2X-enabled ADSs, 83.33% of them are industry practitioners. The primary concern stems from interoperability challenges between different automobile manufacturers and their proprietary DL models. Specifically, a response said, "*If an autonomous vehicle is in the V2X DNN of a Tesla company, it is difficult to transmit data with vehicles that joined other automobile companies, such as BYD.*" Specifically, the core challenge includes adopting distinct DL architectures and frameworks, using different sensor configurations (e.g., variations in LiDAR, camera, and radar setups) that lead to information inconsistencies, and storing and transmitting data in non-standardized formats, all of which prevent seamless cross-company data exchange and real-time decision-making.

5.6. Demand 6: Test case quality generated by LLMs

The advantages of FMs in ADSs include scalability, adaptability, and automation, enabling more diverse scenario construction. VFMs can process and generate realistic driving scenes, while LLMs offer significant advantages in generating test cases in ADSs by reducing manual effort and increasing diversity in testing scenarios. However, LLMs often struggle with real-world consistency, which refers to their ability to align with the physical laws of the world, adhere to traffic regulations, and maintain logical coherence across generated scenarios. This inconsistency leads to test cases that may defy the laws of physics (e.g., vehicles accelerating instantly from 0 to 100 mph), violate established traffic rules (e.g., traffic lights changing unpredictably), or contain contradictions in the sequence of events (e.g., a vehicle stopping at an intersection but still being described as moving forward). 76.19% of responses put forward the testing scenarios generated by LLMs are not accurate (violation of real-world physical or inconsistent logic), most of them are researchers (81.25%), and one of the responses declared that "*Repeated modifications and confirmations with the LLMs are necessary to generate a satisfactory critical scenario,*" highlighting the inefficiency of the current process. It implies that the LLMs lack a deep understanding of real-world physical interactions, affecting testing scenario generation. Besides, professionals also claimed that the limitation of LLMs in context and multi-modal understanding exacerbates the situation of non-accurate testing scenarios.

5.7. Demand 7: Reliable cross-modality integration when using FMs

Apart from test case quality, cross-modality adaptation poses a significant challenge. LLMs, trained on textual data, struggle to process raw sensor inputs (e.g., camera, LiDAR, radar), requiring manual conversion into text, which can lead to errors or missing details. Additionally, translating real-world scenarios into language-based prompts often results in vague test cases, as natural language lacks the precision to capture sensor-driven data. Even with multi-modal FMs, inconsistencies arise in unifying low-level sensor fusion (e.g., camera and LiDAR for object detection) with high-level decision-making (e.g., safe driving maneuvers), creating incoherent tests across modalities.

42.85% of FMs users mentioned the cross-modality adaptation between NLP and vision. This demand involves frontier topics that are primarily driven by academic research, which extract more attention from researchers. Specifically, FMs in ADS testing may struggle with integrating full-scale awareness (an understanding of all relevant objects and events) in complex driving environments. As one practitioner noted, "*LLMs are hard to get the input of the scenario, we have to translate the scenario into language,*" underscoring the fundamental challenge of mapping low-level sensor data into high-level textual descriptions. The inherent ambiguity of natural language further complicates the precise generation of required test scenarios, increasing the risk of misinterpretation in testing. These limitations collectively highlight the effective integration of FMs into ADS testing, especially when bridging perception and reasoning across different modalities.

6. Literature review, current challenges and future direction

In this section, we summarize the literature related to the demands we identified in Section 5, followed by analyzing the challenges and future directions. This work contains 105 references in total, including software engineering journals (TSE, TOSEM, JSS, IST, ESE, SOSP, and SPE) and conferences (ICSE, ISSTA, ASE, FSE, QRS, ICST, and COMPSAC), intelligence vehicle journals (TITS, TII, RAL, IV, and TIV) and conferences (ICRA, CoRL, and DSC), and artificial intelligence conferences (NeurIPS, ICML, AAAI, CVPR, ICCV, and ECCV).

6.1. Challenges and directions for demand 1

Literature: We briefly discussed the testing critical scenario approaches in Section 4.3, including knowledge-based [37,59,60], search-based [6,9,61–65], and data-driven [38,66,67] methods. In the knowledge-based approach, Deng et al. introduced a rule-based metamorphic testing framework for flexible testing scenarios [37]. Li et al. focused on generating test cases for road topology, especially junctions [59]. Guo et al. proposed FuzzScene, which generates perturbations to capture rare behaviors [60]. In the search-based approach, Zhou et al. combined specification-based methods with search techniques to automatically find test cases that violate specifications [62]. The MOSAT uses genetic algorithms to generate test scenarios [61], and scenoRITA identifies and eliminates duplicate scenarios [9]. Tian et al. leveraged real-world traffic data to generate road topology test scenarios [6]. In the data-driven approach, Amini et al. introduced VISTA, a data-driven simulation to generate test cases for transformations and perturbations [66]. Zohdinasab et al. developed DeepAtash-LR, combining data-driven methods with search-based strategies for resource-efficient ADS testing [38]. Additionally, corner case coverage has drawn significant attention. Zhang et al. categorized corner cases and evaluated the suitability of existing metrics and datasets [68]. Li et al. created CODA, a real-world road corner case dataset [55], and Biagiola et al. proposed GenBo, which generates challenging scenarios at the behavior boundary [69].

Challenges: Many papers discuss the limitation of coverage in corner cases [31,68]. Whether in simulation or real-world testing, generating corner cases is difficult. Although the existing testing datasets cover

common driving scenarios, the coverage for extreme scenarios is still an issue. Moreover, the evolving nature of real-world traffic conditions necessitates continuous dataset updates and scenario adaptation, further complicating comprehensive coverage.

Future Directions: (1) LLMs can generate diverse and rare scenarios by integrating knowledge and complex combinations of objects and environments. Recent studies have demonstrated the potential of LLMs in scenario-based reasoning for autonomous driving [10,20]. However, challenges remain in ensuring scenario validity and physical plausibility. (2) Appeal companies and institute open-source datasets, boost the ADSs community to accumulate more data on corner cases. Efforts such as Waymo Open Dataset [70] have enhanced the diversity of test scenarios. One participant pointed out, "*To improve ADS robustness, we need more datasets with real-world adversarial scenarios, not just synthetic data.*" Future work should focus on curating datasets with more adversarial and long-tail driving situations, ensuring better generalization and safety evaluation for ADSs.

6.2. Challenges and directions for demand 2

Literature: Simulation platforms like CARLA [13], Apollo [3], and BeamNG.tech [15] bridge the gap between real-world and simulation testing by offering realistic environments, diverse scenarios, and automated feedback. Studies have provided insights into their effectiveness [57,71–73]. Osinski et al. evaluated simulation-based RL in real-world performance, showing high noise levels with limited randomness [71]. Stocco et al. examined the transferability between simulation and real-world testing, identifying gaps due to random noise and mechanical stochastic effects [57]. Dai et al. introduced SCTrans to generate simulation scenarios from real-world traffic datasets, but it lacked other objects, which may cause discrepancies [74]. Wang et al. tested an E2E ADS in the real world, achieving good performance but with slow inference speed [73].

Challenges: Although simulators can provide a large number of testing data, the uncertainty of the real-world cannot be fully covered, which is mainly attributed to the discrepancies in the received data from sensors between the simulation platforms and the real-world [7,16,68,75,76]. This results in ADSs that perform well in simulators but degrade in the real-world. Additionally, the imprecise dynamic modeling of real-world physics and traffic behavior further exacerbates the problem [77–80]. Consequently, achieving reliable transferability between simulated and real-world testing remains an ongoing challenge.

Future Directions: (1) Testing on hybrid simulation and real-world by screening valuable testing scenarios in simulators and then validating them in real-world experiments, this hybrid approach improves testing efficiency while ensuring scenarios reflect real-world conditions. Simulators can identify valuable scenarios, which are then tested in controlled real-world settings, bridging the gap between simulation limitations and real-world traffic complexities. (2) Explore the fusion technology of multi-modal sensors, especially testing when sensor data fails in extreme environments to enhance the ADSs' robustness. This would improve ADS reliability by ensuring it can adapt and function even when typical sensor inputs fail or are inaccurate. One professional noted, "*Heavy rain or fog often degrades LiDAR and camera performance, making it critical to rely on radar or sensor fusion techniques.*" Recent advancements in multi-sensor fusion, particularly in real-world testing [16], can significantly improve ADS robustness in dynamic and unpredictable environments.

6.3. Challenges and directions for demand 3

Literature: Evaluation metrics for simulation and real-world testing are designed to measure the safety, reliability, and ability of ADSs, typically including collision rates, trajectory deviations, passenger comfort, etc. In system-level and E2E simulation testing of ADSs, many studies utilized prevalent simulation benchmarks [81–83] like CoRL2017 [13], NoCrash [84] and CARLA Leaderboard [85] to assess performance. The benchmarks focus on criteria such as task success rate, infraction, navigation, etc. CoRL2017 pays more attention to decision-making ability in complex navigation tasks [86], NoCrash mainly evaluates the robustness and adaptability of automatic driving systems in complex traffic scenarios [87], and CARLA Leaderboard provides a comprehensive assessment framework [88,89]. Laurent et al. presented an alternative method that combines different parameters for testing to evaluate different decision outputs, thus ensuring each decision path of the ADS is tested [90]. Besides, in the real-world testing, the criteria also mainly focused on lane-following and obstacle avoidance [73].

Challenges: Despite the widespread use of simulation benchmarks such as CoRL2017, NoCrash, and CARLA Leaderboard for evaluating ADS performance, the existing evaluation metrics often fail to address the full spectrum of challenges faced by autonomous systems in real-world environments. One of the primary limitations is that these metrics focus on specific, isolated aspects of ADS performance. However, they do not sufficiently assess the long-term stability or robustness of ADSs during continuous operation. This oversight becomes crucial when considering the need for an ADS to operate safely and reliably over extended periods, often under varying conditions [91]. Additionally, these metrics lack the ability to provide granular feedback on system failures, making it difficult to diagnose performance issues or bottlenecks. This limitation also hinders the ability to compare ADS performance directly with human driving behavior [72,92].

Future Directions: (1) Consider the long-term performance of ADSs and assess adaptability in multi-tasking scenarios. This includes assessing the system's ability to handle diverse and dynamic environments, particularly in multi-tasking scenarios where the system must handle multiple tasks simultaneously or transition between tasks smoothly. Therefore, one participant emphasized that "*We need better ways to test how ADSs adapt to long-term driving scenarios instead of just short tests.*" (2) Introduce comprehensive feedback systems that measure reaction time in complex scenarios and assess responses to unexpected driving scenarios.

6.4. Challenges and directions for demand 4

Literature: We investigate attacks frequently mentioned in the responses: physical, cyber, inference, and adversarial attacks. Physical attacks disrupt sensor data or fabricate signals using external hardware. Two common examples are jamming [93] and spoofing attacks [27]. Jamming adds noise to degrade sensor data, while spoofing injects signals during data collection. Beyond model-agnostic defenses, Sun et al. proposed an ADS defense against sensor attacks, though it leads to false alarms [34]. Cao et al. compared existing defenses, noting that the best defense only reduces the attack success rate to 66% [94].

Cyber-attacks on vehicle systems and transport infrastructure pose significant threats to V2X communication [95]. For instance, attackers can control the upload process of HD Maps to the cloud [96]. Denial of Service attacks can also disrupt V2X network availability [97]. Defenses against cyber-attacks span areas like data transmission, authentication, and network routing, though they require considerable computational resources [95]. Additionally, inference attacks, evaluated in IoT and cloud techniques, pose threats [98,99]. He et al. showed that membership inference attacks leak information about the perception training dataset, with defenses like Gaussian noise and differential privacy potentially reducing model accuracy [100].

Adversarial attacks, which involve pixel-level perturbations to input data, cause models to make incorrect decisions [101–104]. ADEPT generates adversarial attacks based on ADS feedback from various test scenarios [103]. Von et al. applied adversarial testing to vehicle

trajectories to refine perturbations [104], while Wu et al. developed white-box adversarial attacks against the E2E system [101], and Xiang et al. integrated adversarial attacks into the V2X system [102]. The most common defense is adversarial training, though it requires significant resources [102,104].

Challenges: These attacks can affect multiple components of an ADS simultaneously, requiring defense mechanisms that address various dimensions of the system, including data integrity, communication security, sensor fusion, and more [27,68,96,105]. A major challenge lies in ensuring that defenses can work across the physical, communication, and model layers to provide robust protection without significantly hindering system performance.

Future Directions: (1) Construct a unified, comprehensive security assessment standard or framework that is flexible to adapt to emerging attack strategies and ensure a holistic defense across different ADS components [68]. The professionals agree that "*Different companies use different security testing criteria, making it hard to compare ADS security levels objectively.*" Therefore, a well-defined framework should integrate attack simulation, real-time threat detection, and system-wide robustness evaluation. (2) Develop ADSs that can sense threats in real-time and respond dynamically. This would require the implementation of adaptive security mechanisms that continuously monitor the vehicle's environment, system state, and communication network for potential threats [27].

6.5. Challenges and directions for demand 5

Literature: Communication and cross-model collaboration are inevitable issues in the V2X system [17–19]. Sedar et al. presented an onboard unit that contributes to communication among various equipment manufacturers, who leveraged different V2X protocols [17]. Besides, the standard communication protocols and Ethernet-based networks have leverage widespread [18]. However, different organization develops their own DNN models for ADSs, which impacts model compatibility. The common DNN model in ADSs, including Tesla's Full Self-Driving (FSD) DNN [1], Apollo DNN [3], Waymo DNN [70], etc. Among them, Tesla's FSD and Waymo DNNs rely on their closed-loop ADS, which makes it difficult to exchange data with other ADSs seamlessly.

Challenges: The research on model compatibility among multi-vendors is limited [106]. The DNN models developed by different automobile manufacturers utilized different data processing methods, feature extraction methods, or differences in decision modules, making distinct ADS difficult to collaborate seamlessly. For example, different V2X interaction models choose different data fusion strategies (early, intermediate, late) according to their application requirements.

Future Directions: (1) Introduce knowledge distillation to extract useful knowledge from different DNN models and transfer it into a common lightweight model that can be applied across platforms [107]. One of the professionals states that "*If we can distill a smaller, cross-model collaboration ADS, it would greatly improve real-time decision-making.*" It proposes that knowledge distillation may serve as a bridge between proprietary ADS models, ensuring cross-platform compatibility and computational efficiency. (2) Establish standardized interfaces for models in V2X systems, calling for models from various vendors to collaborate through the unified interface. This would reduce the barriers to cross-platform collaboration, enabling better communication and cooperation in V2X systems [19]. One participant indicates the potential benefits of a standardized framework, "*If every company uses its own protocol, cross-vehicle communication will always be a challenge. A standardized model interface would significantly improve vehicle coordination in V2X.*"

6.6. Challenges and directions for demand 6 and demand 7

Because Demand 6 and Demand 7 are both related to FMs, we present literature relevant to LLMs and VFMs in the same section and discuss the challenges, respectively.

Literature: Several studies have applied LLMs to generate testing scenarios for ADSs [10,12,20–23]. Zhang et al. [20] used GPT-3.5 to generate metamorphic relations for metamorphic testing. Guo et al. proposed SoVAR, utilizing GPT-4's ability to extract textual information and generate driving trajectories in the Apollo simulator [10]. Zhang et al. also used GPT-4 to create complex testing scenarios in the CARLA simulator [12]. Lu et al. leveraged GPT-4 to diagnose safety violations in simulation testing [23]. In VFMs, SAM [54] segments objects from images without specific training datasets, while DINO [24] uses self-supervised learning to generate visual representations, enabling the generation of diverse testing scenarios in unseen driving environments.

Challenge of Demand 6: Improving the generated scenario quality is a demand for introducing LLMs for testing ADSs. The current LLMs do not provide satisfactory results in understanding traffic jargon and non-standard phrase structures in accident reports [10,21]. Specifically, Song et al. leveraged GPT-4 to identify benign scenarios, but the scenarios lacking surrounding objects were recognized as inconsistent [108].

Challenge of Demand 7: LLMs are based on natural languages. [22,23,108] translates domain-specific language into natural language or leverages LLMs setup viewpoint's position and angle by input natural language. However, the tasks for the generation may not match the expectations due to natural language ambiguity. Errors in LLMs' comprehension of commands also affect task accuracy. Besides, testing standardization and consistency are hard to guarantee, and the different test instructions may generate different scenarios.

Future Directions: (1) Divide test cases into multiple levels (e.g., basic driving behavior, scene complexity, interaction dynamics), allowing LLMs to generate high-quality test cases by controlling different levels. One of the professionals argues that "*LLMs generate test cases but often lack structure. If we categorize test cases into different levels, we can systematically control scenario difficulty and complexity.*" Another professional also agrees, "*Some test cases are too trivial, while others are unrealistic. If we define structured levels, LLMs can better generate relevant and meaningful cases.*" (2) Building the translation module, transforming the natural language description into the parameterized scene configuration, which helps to generate more controllable scenes. One practitioner claims the limitations of current test case creation, "*We often describe a scenario verbally. If we had an automated way to translate test descriptions into simulation-ready parameters, it would significantly streamline testing.*" It shows a direction for setting up a translation layer that bridges high-level test case descriptions with executable simulation parameters, ensuring that generated scenarios are both realistic and reproducible.

7. Threats to validity

We summarize threats to validity following the criteria from [109], which are divided into four parts: internal, external, construct, and conclusion validity.

Internal Validity. To ensure internal validity and avoid sensitive questions, we avoided questions that might make respondents worried about disclosing company information to encourage truthful responses from participants. During the pilot survey, we adjusted the questions based on feedback until all 10 pilots claimed that the questions were reasonable and did not involve sensitive information. Besides, the anonymity and confidentiality of the data were explained to respondents before

our survey to ensure their answers did not reveal any personal or company information.

External Validity. Two main external threats may affect the generalizability of our findings: participant sampling bias and discussion bias.

Regarding participant sampling bias: ***(1)*** our recruitment strategy relied on personal networks and GitHub email crawling. While these methods may introduce sampling bias, potentially favoring academic researchers and open-source ADS developers, we ensured diversity within our personal network by including both industry practitioners and academic contacts. Additionally, we applied strict filtering to remove incomplete, low-quality, or irrelevant responses (e.g., from vehicle scheduling). As a result, our final sample comprises 100 validated ADS testers, with 59% from academia and 41% from industry, supporting a balanced representation of perspectives. Nonetheless, we acknowledge that practitioners working on confidential, closed-source ADS projects may still be under-represented. To mitigate this further, we explicitly differentiate academic and industry perspectives throughout our analysis. ***(2)*** Geographical imbalance may exist. While our participants span eight countries, the majority were recruited through our professional network and are from China (24 participants), the USA (16), Japan (12), and the UK, which together represent major markets and deployment paradigms in ADS development. Although countries like Germany and other regions may be underrepresented, the inclusion of these leading regions, covering both aggressive field deployment (e.g., China, USA) and cautious safety-focused strategies (e.g., Japan, UK), helps ensure that our findings reflect diverse and globally relevant testing practices and challenges.

Regarding the discussion bias, the results may be affected by the initial discussion during the study. Discussions with four professionals involved in ADS testing may have been subjective due to their different experiences and backgrounds, resulting in the lack of representation of the general population. To prevent the survey from focusing only on areas more familiar to the professionals, we conducted literature reviews to summarize the large-scale survey with 107 questions. Besides, we invited 10 experienced candidates as pilot respondents to double-check the comprehensiveness of the survey. The survey was distributed to academic researchers and industry practitioners, and 100 responses were received.

Construct Validity. If the survey cannot adequately cover the different test dimensions, may lead to misunderstandings about the overall performance of the ADS testing. To improve construct validity, we comprehensively analyzed ADS testing by covering multiple subareas involved, including V2X, module-level testing, system-level testing, etc. Although most surveys are multiple-choice, we have provided alternative answers to receive more voices. In the follow-up section, we designed open-ended questions for distinct respondents to further investigate their demands.

Our study did not explicitly restrict the scope to a particular Society of Automotive Engineers (SAE) automation level, which defines six levels of driving automation from Level 0 (no automation) to Level 5 (full automation) [28]. Although responses suggest alignment with Level-3 to Level-5 systems based on system descriptions and affiliations, this lack of explicit separation may introduce interpretation variance across participants. Nevertheless, this was intended to ensure broad coverage of ADS testing practices across all automation levels, and many companies evaluate features spanning multiple levels (e.g., Level-3 and Level-4) under unified testing frameworks. We inferred the approximate SAE levels based on participants' affiliations and response content, where practitioners primarily referenced Level-3 and Level-4 systems, while researchers discussed challenges related to Level-4 and Level-5 autonomy. Therefore, the collected responses still reasonably reflect a wide range of ADS testing practices and challenges.

Conclusion Validity. Regarding the conclusion validity of the survey, the first two authors selected 90 complete and high-quality responses separately, with a consistency rate of 97.78%, and finally obtained a total of 100 responses as the thematic analysis foundation. Additionally, we further surveyed participants who met the follow-up criteria with a response rate of 68.22%. To capture respondents' thoughts, we designed different open-ended questions based on the participants' ratings to determine the responses' consistency.

8. Conclusion

This paper presents a comprehensive survey of testing practices for ADSs, including both modular and E2E systems. It highlights the key demands and challenges faced by both industry practitioners and academic researchers. The survey methodology involved discussions with professionals, a detailed survey of ADS testers and researchers, and follow-up open-ended questions. Seven critical demands were identified, with a particular focus on the diversity of corner cases, testing criteria, potential attacks, and V2X interoperability. Additionally, the increasing use of LLMs for generating test scenarios is noted, alongside demands for improving the quality of test cases through FMs. The paper also provides an in-depth literature review of software engineering research to evaluate progress in addressing these challenges. This work offers actionable insights and future research directions to enhance ADS testing methodologies, ultimately contributing to safer and more reliable autonomous driving systems.

CRediT authorship contribution statement

Yihan Liao: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jingyu Zhang:** Writing – review & editing, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jacky Keung:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Yan Xiao:** Writing – review & editing, Supervision, Conceptualization. **Yurou Dai:** Methodology, Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (Grant No. 24qnpy153), the Shenzhen Science and Technology Program (Grant No. KJZD20240903095700001), and the National Natural Science Foundation of China (Grant No. 62402499).

Data availability

Data is available at https://github.com/ADSTestingSurvey/ADS-Testing.

References

[1] Tesla, Tesla, 2022, URL: https://www.tesla.com/autopilot.
[2] Waymo, Waymo safety report, 2021, URL: https://waymo.com/safety/.
[3] ApolloAuto, Apollo, 2021, URL: https://github.com/ApolloAuto/apollo.
[4] On-Road Automated Driving (ORAD) Committee, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE International, 2021.
[5] Associated Press, US report: Nearly 400 crashes of automated tech vehicles, 2022, URL: https://www.voanews.com/a/us-report-nearly-400-crashes-of-automated-tech-vehicles-/6618927.html.

[6] H. Tian, et al., Generating critical test scenarios for autonomous driving systems via influential behavior patterns, in: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, 2022, pp. 1–12.

[7] W. Ding, et al., A survey on safety-critical driving scenario generation—A methodological perspective, IEEE Trans. Intell. Transp. Syst. 24 (7) (2023) 6971–6988.

[8] C. Lu, Test scenario generation for autonomous driving systems with reinforcement learning, in: 2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings, ICSE-Companion, IEEE, 2023, pp. 317–319.

[9] Y. Huai, et al., sceno RITA: Generating diverse, fully-mutable, test scenarios for autonomous vehicle planning, IEEE Trans. Softw. Eng. (2023).

[10] A. Guo, et al., SoVAR: Building generalizable scenarios from accident reports for autonomous driving testing, 2024, arXiv preprint arXiv:2409.08081.

[11] Q. Song, et al., An empirically grounded path forward for scenario-based testing of autonomous driving systems, in: Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, 2024, pp. 232–243.

[12] J. Zhang, et al., ChatScene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15459–15469.

[13] A. Dosovitskiy, et al., CARLA: An open urban driving simulator, in: Conference on Robot Learning, PMLR, 2017, pp. 1–16.

[14] Tier I.V., AWSIM, 2022, URL: https://github.com/tier4/AWSIM/tree/main.

[15] BeamN.G. GmbH, BeamNG.tech, 2022, URL: https://github.com/BeamNG/BeamNGpy.

[16] R. Xu, et al., V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13712–13722.

[17] R. Sedar, et al., Standards-compliant multi-protocol on-board unit for the evaluation of connected and automated mobility services in multi-vendor environments, Sensors 21 (6) (2021) 2090.

[18] A. Martínez-Cruz, et al., Security on in-vehicle communication protocols: Issues, challenges, and future research directions, Comput. Commun. 180 (2021) 1–20.

[19] B. Yang, B. Wu, Y. You, C. Guo, L. Qiao, Z. Lv, Edge intelligence based digital twins for internet of autonomous unmanned vehicles, Softw.: Pr. Exp. (2022).

[20] Y. Zhang, et al., Automated metamorphic-relation generation with ChatGPT: An experience report, in: 2023 IEEE 47th Annual Computers, Software, and Applications Conference, COMPSAC, IEEE, 2023, pp. 1780–1785.

[21] Q. Lu, et al., Multimodal large language model driven scenario testing for autonomous vehicles, 2024, arXiv preprint arXiv:2409.06450.

[22] Y. Wei, et al., Editable scene simulation for autonomous driving via collaborative llm-agents, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15077–15087.

[23] Y. Lu, et al., DiaVio: LLM-empowered diagnosis of safety violations in ADS simulation testing, in: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis, 2024, pp. 376–388.

[24] M. Caron, et al., Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.

[25] M. Käppeler, et al., Few-shot panoptic segmentation with foundation models, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 7718–7724.

[26] G. Lou, et al., Testing of autonomous driving systems: where are we and where should we go? in: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 31–43.

[27] S. Tang, et al., A survey on automated driving system testing: Landscapes and trends, ACM Trans. Softw. Eng. Methodol. 32 (5) (2023) 1–62.

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

[29] X. Huang, et al., A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Comput. Sci. Rev. 37 (2020) 100270.

[30] J. Wang, et al., Software testing with large language models: Survey, landscape, and vision, IEEE Trans. Softw. Eng. (2024).

[31] Q. Song, et al., Industry practices for challenging autonomous driving systems with critical scenarios, ACM Trans. Softw. Eng. Methodol. 33 (4) (2024) 1–35.

[32] F. Shull, et al., Guide to Advanced Empirical Software Engineering, vol. 93, Springer, 2008.

[33] R. Bommasani, et al., On the opportunities and risks of foundation models, 2021, arXiv preprint arXiv:2108.07258.

[34] J. Sun, Y. Cao, Q.A. Chen, Z.M. Mao, Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures, in: 29th USENIX Security Symposium, USENIX Security 20, 2020, pp. 877–894.

[35] K. Pei, et al., Deepxplore: Automated whitebox testing of deep learning systems, in: Proceedings of the 26th Symposium on Operating Systems Principles, 2017, pp. 1–18.

[36] L.J. Moukahal, et al., Boosting grey-box fuzzing for connected autonomous vehicle systems, in: 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion, QRS-C, IEEE, 2021, pp. 516–527.

[37] Y. Deng, et al., A declarative metamorphic testing framework for autonomous driving, IEEE Trans. Softw. Eng. 49 (4) (2022) 1964–1982.

[38] T. Zohdinasab, et al., Focused test generation for autonomous driving systems, ACM Trans. Softw. Eng. Methodol. (2024).

[39] F. Klück, et al., An empirical comparison of combinatorial testing and search-based testing in the context of automated and autonomous driving systems, Inf. Softw. Technol. 160 (2023) 107225.

[40] A. Joshi, et al., Likert scale: Explored and explained, Br. J. Appl. Sci. Technol. 7 (4) (2015) 396–403.

[41] T. Alqahtani, Recent trends in the public acceptance of autonomous vehicles: A review, Vehicles 7 (2) (2025) 45.

[42] S. Keele, et al., Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report, ver. 2.3 ebse technical report. ebse, 2007.

[43] J. Zhang, et al., UniAda: Universal adaptive multiobjective adversarial attack for end-to-end autonomous driving systems, IEEE Trans. Reliab. (2024).

[44] Y. Liu, et al., {ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models, in: 31st USENIX Security Symposium, USENIX Security 22, 2022, pp. 4525–4542.

[45] A. Kuznietsov, B. Gyevnar, C. Wang, et al., Explainable AI for safe and trustworthy autonomous driving: a systematic review, IEEE Trans. Intell. Transp. Syst. (2024).

[46] S. Nazat, O. Arreche, M. Abdallah, On evaluating black-box explainable AI methods for enhancing anomaly detection in autonomous driving systems, Sensors 24 (11) (2024) 3515.

[47] R. Xu, et al., V2x-vit: Vehicle-to-everything cooperative perception with vision transformer, in: European Conference on Computer Vision, Springer, 2022, pp. 107–124.

[48] Y. Li, et al., V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving, IEEE Robot. Autom. Lett. 7 (4) (2022) 10914–10921.

[49] H. Yu, et al., V2X-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[50] C. Cui, et al., A survey on multimodal large language models for autonomous driving, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 958–979.

[51] L. Chen, et al., Driving with llms: Fusing object-level vector modality for explainable autonomous driving, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 14093–14100.

[52] Z. Wei, et al., Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 28619–28630.

[53] A. Radford, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[54] A. Kirillov, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[55] K. Li, et al., Coda: A real-world road corner case dataset for object detection in autonomous driving, in: European Conference on Computer Vision, Springer, 2022, pp. 406–423.

[56] R. Donà, et al., Virtual testing of automated driving systems. A survey on validation methods, IEEE Access 10 (2022) 24349–24367.

[57] A. Stocco, et al., Mind the gap! A study on the transferability of virtual versus physical-world testing of autonomous driving systems, IEEE Trans. Softw. Eng. 49 (4) (2022) 1928–1940.

[58] L. Westhofen, C. Neurohr, T. Koopmann, et al., Criticality metrics for automated driving: A review and suitability analysis of the state of the art, Arch. Comput. Methods Eng. 30 (1) (2023) 1–35.

[59] C. Li, et al., Simulation-based validation for autonomous driving systems, in: Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis, 2023, pp. 842–853.

[60] A. Guo, et al., Semantic-guided fuzzing for virtual testing of autonomous driving systems, J. Syst. Softw. 212 (2024) 112017.

[61] H. Tian, et al., MOSAT: finding safety violations of autonomous driving systems using multi-objective genetic algorithm, in: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 94–106.

[62] Y. Zhou, et al., Specification-based autonomous driving system testing, IEEE Trans. Softw. Eng. 49 (6) (2023) 3391–3410.

[63] A.A. Babikian, et al., Concretization of abstract traffic scene specifications using metaheuristic search, IEEE Trans. Softw. Eng. (2023).

[64] L. Giamattei, et al., Causality-driven testing of autonomous driving systems, ACM Trans. Softw. Eng. Methodol. 33 (3) (2024) 1–35.

[65] Z. Li, et al., VioHawk: Detecting traffic violations of autonomous driving systems through criticality-guided simulation testing, in: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis, 2024, pp. 844–855.
[66] A. Amini, et al., Learning robust control policies for end-to-end autonomous driving from data-driven simulation, IEEE Robot. Autom. Lett. 5 (2) (2020) 1143–1150.
[67] N. Neelofar, et al., Identifying and explaining safety-critical scenarios for autonomous vehicles via key features, ACM Trans. Softw. Eng. Methodol. 33 (4) (2024) 1–32.
[68] X. Zhang, et al., Finding critical scenarios for automated driving systems: A systematic mapping study, IEEE Trans. Softw. Eng. 49 (3) (2022) 991–1026.
[69] M. Biagiola, et al., Boundary state generation for testing and improvement of autonomous driving systems, IEEE Trans. Softw. Eng. (2024).
[70] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, et al., Scalability in perception for autonomous driving: Waymo open dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2446–2454.
[71] B. Osiński, et al., Simulation-based reinforcement learning for real-world autonomous driving, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2020, pp. 6411–6418.
[72] M. Biagiola, et al., Two is better than one: digital siblings to improve autonomous driving testing, Empir. Softw. Eng. 29 (4) (2024) 1–33.
[73] T.H. Wang, et al., Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 6687–6694.
[74] J. Dai, et al., SCTrans: Constructing a large public scenario dataset for simulation testing of autonomous driving systems, in: Proceedings of the 46th IEEE/ACM International Conference on Software Engineering, 2024, pp. 1–13.
[75] A. Prakash, et al., Multi-modal fusion transformer for end-to-end autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7077–7087.
[76] S. Suo, et al., Trafficsim: Learning to simulate realistic multi-agent behaviors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10400–10409.
[77] F.U. Haq, et al., Comparing offline and online testing of deep neural networks: An autonomous car case study, in: 2020 IEEE 13th International Conference on Software Testing, Validation and Verification, ICST, IEEE, 2020, pp. 85–95.
[78] Y. Xiao, et al., Deep neural networks with koopman operators for modeling and control of autonomous vehicles, IEEE Trans. Intell. Veh. 8 (1) (2022) 135–146.
[79] A. Stocco, et al., Model vs system level testing of autonomous driving systems: a replication and extension study, Empir. Softw. Eng. 28 (3) (2023) 73.
[80] J. Cheng, et al., Rethinking imitation-based planners for autonomous driving, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 14123–14130.
[81] Z. Huang, et al., Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding, IEEE Sensors J. 21 (10) (2020) 11781–11790.
[82] P. Wu, et al., Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline, Adv. Neural Inf. Process. Syst. 35 (2022) 6119–6132.
[83] D. Coelho, et al., RLfOLD: Reinforcement learning from online demonstrations in urban autonomous driving, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, (10) 2024, pp. 11660–11668.
[84] F. Codevilla, et al., Exploring the limitations of behavior cloning for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9329–9338.
[85] CARLA, Autonomous driving leaderboard, 2022, URL: https://leaderboard.carla.org/leaderboard.
[86] K. Ishihara, et al., Multi-task learning with attention for end-to-end autonomous driving, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2902–2911.
[87] M. Ahmed, et al., Policy-based reinforcement learning for training autonomous driving agents in urban areas with affordance learning, IEEE Trans. Intell. Transp. Syst. 23 (8) (2021) 12562–12571.
[88] F.U. Haq, et al., Many-objective reinforcement learning for online testing of dnn-enabled systems, in: 2023 IEEE/ACM 45th International Conference on Software Engineering, ICSE, IEEE, 2023, pp. 1814–1826.
[89] C. Lu, et al., Epitester: Testing autonomous vehicles with epigenetic algorithm and attention mechanism, IEEE Trans. Softw. Eng. (2024).
[90] T. Laurent, et al., Parameter coverage for testing of autonomous driving systems under uncertainty, ACM Trans. Softw. Eng. Methodol. 32 (3) (2023) 1–31.
[91] S. Mozaffari, et al., Deep learning-based vehicle behavior prediction for autonomous driving applications: A review, IEEE Trans. Intell. Transp. Syst. 23 (1) (2020) 33–47.
[92] Y. Zhao, et al., Cadre: A cascade deep reinforcement learning framework for vision-based autonomous urban driving, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (3) 2022, pp. 3481–3489.
[93] M. Pham, et al., A survey on security attacks and defense techniques for connected and autonomous vehicles, Comput. Secur. 109 (2021) 102269.
[94] Y. Cao, et al., Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks, in: 2021 IEEE Symposium on Security and Privacy, SP, IEEE, 2021, pp. 176–194.
[95] K. Kim, et al., Cybersecurity for autonomous vehicles: Review of attacks and defense, Comput. Secur. 103 (2021) 102150.
[96] Y. Deng, et al., Deep learning-based autonomous driving systems: A survey of attacks and defenses, IEEE Trans. Ind. Informatics 17 (12) (2021) 7897–7912.
[97] Z. Pethő, et al., Quantifying cyber risks: The impact of DoS attacks on vehicle safety in V2X networks, IEEE Trans. Intell. Transp. Syst. (2024).
[98] X. Gong, et al., Private data inference attacks against cloud: Model, technologies, and research directions, IEEE Commun. Mag. 60 (9) (2022) 46–52.
[99] R. Yang, et al., Practical feature inference attack in vertical federated learning during prediction in artificial internet of things, IEEE Internet Things J. 11 (1) (2023) 5–16.
[100] Y. He, et al., Segmentations-leak: Membership inference attacks and defenses in semantic image segmentation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16, Springer, 2020, pp. 519–535.
[101] H. Wu, et al., Adversarial driving: Attacking end-to-end autonomous driving, in: 2023 IEEE Intelligent Vehicles Symposium, IV, IEEE, 2023, pp. 1–7.
[102] H. Xiang, et al., V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2023, pp. 3584–3591.
[103] S. Wang, et al., ADEPT: A testing platform for simulated autonomous driving, in: Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, 2022, pp. 1–4.
[104] M. von Stein, et al., DeepManeuver: Adversarial test generation for trajectory manipulation of autonomous vehicles, IEEE Trans. Softw. Eng. (2023).
[105] Z. Zhong, et al., Neural network guided evolutionary fuzzing for finding traffic violations of autonomous vehicles, IEEE Trans. Softw. Eng. 49 (4) (2022) 1860–1875.
[106] S. Liu, et al., Towards vehicle-to-everything autonomous driving: A survey on collaborative perception, 2023, arXiv preprint arXiv:2308.16714.
[107] J. Gou, et al., Knowledge distillation: A survey, Int. J. Comput. Vis. 129 (6) (2021) 1789–1819.
[108] R. Song, et al., Enhancing LLM-based autonomous driving agents to mitigate perception attacks, 2024, arXiv preprint arXiv:2409.14488.
[109] C. Wohlin, et al., Experimentation in Software Engineering, vol. 236, Springer, 2012.